diagnostics

Special Issue Reprint

Advances in Medical Image Processing, Segmentation and Classification

Edited by
Wan Azani Mustafa and Hiam Alquran

mdpi.com/journal/diagnostics

MDPI

Advances in Medical Image Processing, Segmentation and Classification

Advances in Medical Image Processing, Segmentation and Classification

Guest Editors

Wan Azani Mustafa
Hiam Alquran

MDPI

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

Guest Editors

Wan Azani Mustafa
Faculty of Electrical
Engineering & Technology
Universiti Malaysia Perlis
Perlis
Malaysia

Hiam Alquran
Department of Biomedical
Systems and Informatics
Engineering
Yarmouk University
Irbid
Jordan

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Diagnostics* (ISSN 2075-4418), freely accessible at: https://www.mdpi.com/journal/diagnostics/special_issues/RD7FWD9ZE7.

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

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

ISBN 978-3-7258-4123-3 (Hbk)
ISBN 978-3-7258-4124-0 (PDF)
https://doi.org/10.3390/books978-3-7258-4124-0

Contents

About the Editors . . . . . **vii**

Wan Azani Mustafa and Hiam Alquran
Editorial for the Special Issue "Advances in Medical Image Processing, Segmentation, and Classification"
Reprinted from: *Diagnostics* **2025**, *15*, 1114, https://doi.org/10.3390/diagnostics15091114 . . . . **1**

İrem Çetinkaya, Ekin Deniz Çatmabacak and Emir Öztürk
Detection of Fractured Endodontic Instruments in Periapical Radiographs: A Comparative Study of YOLOv8 and Mask R-CNN
Reprinted from: *Diagnostics* **2025**, *15*, 653, https://doi.org/10.3390/diagnostics15060653 . . . . . **8**

Se-Yeol Rhyou, Minyung Yu and Jae-Chern Yoo
Mixture of Expert-Based SoftMax-Weighted Box Fusion for Robust Lesion Detection in Ultrasound Imaging
Reprinted from: *Diagnostics* **2025**, *15*, 588, https://doi.org/10.3390/diagnostics15050588 . . . . . **24**

Ahmet Bozdag, Muhammed Yildirim, Mucahit Karaduman, Hursit Burak Mutlu, Gulsah Karaduman and Aziz Aksoy
Detection of Gallbladder Disease Types Using a Feature Engineering-Based Developed CBIR System
Reprinted from: *Diagnostics* **2025**, *15*, 552, https://doi.org/10.3390/diagnostics15050552 . . . . . **43**

Madallah Alruwaili and Mahmood Mohamed
An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification
Reprinted from: *Diagnostics* **2025**, *15*, 551, https://doi.org/10.3390/diagnostics15050551 . . . . . **58**

Yi-Ching Cheng, Yi-Chieh Hung, Guan-Hua Huang, Tai-Been Chen, Nan-Han Lu, Kuo-Ying Liu and Kuo-Hsuan Lin
Deep Learning-Based Object Detection Strategies for Disease Detection and Localization in Chest X-Ray Images
Reprinted from: *Diagnostics* **2024**, *14*, 2636, https://doi.org/10.3390/diagnostics14232636 . . . . **76**

Mofleh Hannuf AlRowaily, Hamzah Arof, Imanurfatiehah Ibrahim, Haniza Yazid and Wan Amirul Mahyiddin
Enhancing Retina Images by Lowpass Filtering Using Binomial Filter
Reprinted from: *Diagnostics* **2024**, *14*, 1688, https://doi.org/10.3390/diagnostics14151688 . . . . **95**

Aneesha Baral, Simone Lee, Farah Hussaini, Brianna Matthew, Alfredo Lebron, Muyang Wang, et al.
Clinical Trial Validation of Automated Segmentation and Scoring of Pulmonary Cysts in Thoracic CT Scans
Reprinted from: *Diagnostics* **2024**, *14*, 1529, https://doi.org/10.3390/diagnostics14141529 . . . . **107**

Irene Ligato, Giorgio De Magistris, Emanuele Dilaghi, Giulio Cozza, Andrea Ciardiello, Francesco Panzuto, et al.
Convolutional Neural Network Model for Intestinal Metaplasia Recognition in Gastric Corpus Using Endoscopic Image Patches
Reprinted from: *Diagnostics* **2024**, *14*, 1376, https://doi.org/10.3390/diagnostics14131376 . . . . **117**

Semin Kim, Huisu Yoon and Jongha Lee
Semi-Supervised Facial Acne Segmentation Using Bidirectional Copy–Paste
Reprinted from: *Diagnostics* **2024**, *14*, 1040, https://doi.org/10.3390/diagnostics14101040 . . . . **128**

Yuting Xie, Fulvio Zaccagna, Leonardo Rundo, Claudia Testa, Ruifeng Zhu, Caterina Tonon, et al.
IMPA-Net: Interpretable Multi-Part Attention Network for Trustworthy Brain Tumor Classification from MRI
Reprinted from: *Diagnostics* **2024**, *14*, 997, https://doi.org/10.3390/diagnostics14100997 . . . . . **141**

Daphné Mulliez, Edouard Poncelet, Laurie Ferret, Christine Hoeffel, Blandine Hamet, Lan Anh Dang, et al.
Three-Dimensional Measurement of the Uterus on Magnetic Resonance Images: Development and Performance Analysis of an Automated Deep-Learning Tool
Reprinted from: *Diagnostics* **2023**, *13*, 2662, https://doi.org/10.3390/diagnostics13162662 . . . . **159**

Reham Kaifi
A Review of Recent Advances in Brain Tumor Diagnosis Based on AI-Based Classification
Reprinted from: *Diagnostics* **2023**, *13*, 3007, https://doi.org/10.3390/diagnostics13183007 . . . . **172**

Wan Azani Mustafa, Shahrina Ismail, Fahirah Syaliza Mokhtar, Hiam Alquran and Yazan Al-Issa
Cervical Cancer Detection Techniques: A Chronological Review
Reprinted from: *Diagnostics* **2023**, *13*, 1763, https://doi.org/10.3390/diagnostics13101763 . . . . **204**

Taye Girma Debelee
Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review
Reprinted from: *Diagnostics* **2023**, *13*, 3147, https://doi.org/10.3390/diagnostics13193147 . . . . **229**

About the Editors

Wan Azani Mustafa

Wan Azani Mustafa is an accomplished academic and researcher specializing in biomedical, electronic, and mechatronic engineering. He holds a bachelor's degree in Biomedical Electronic Engineering (2013) and a Ph.D. in Mechatronic Engineering (2017) from Universiti Malaysia Perlis (UniMAP). Dr. Wan Azani has been a registered member of the Board of Engineers Malaysia (BEM) since 2014 and the Malaysia Board of Technologists (MBOT) since 2017. He is also recognized as a Senior Member of the IEEE, reflecting his active engagement in the international engineering community. Currently serving as a Senior Lecturer at UniMAP, Dr. Wan Azani has demonstrated exceptional research productivity, having published over 400 academic articles and achieving a Scopus H-Index of 22. His research interests are interdisciplinary, spanning image processing, biomechanics, intelligent systems, and computer science. His contributions significantly impact the fields of engineering and applied technology, positioning him as a prominent figure in his areas of expertise.

Hiam Alquran

Hiam Alquran is a Professor at the Department of Biomedical Systems and Informatics Engineering, Yarmouk University, Jordan. Alquran received her Ph.D. (2014) in Biomedical and Biotechnology Engineering from Massachusetts Lowell University, USA, her M.Sc. (2008) in Automation Engineering from Yarmouk University, and her B.Sc. in Biomedical Engineering from JUST, Jordan (2005). Her research interests are in medical image processing, digital signal processing, pattern recognition, and deep learning.

diagnostics

MDPI

Editorial

Editorial for the Special Issue "Advances in Medical Image Processing, Segmentation, and Classification"

Wan Azani Mustafa [1,2,*] and Hiam Alquran [3]

[1] Faculty of Electrical Engineering Technology, Campus Pauh Putra, Universiti Malaysia Perlis, Arau 02600, Malaysia
[2] Advanced Computing (AdvComp), Centre of Excellence (CoE), Universiti Malaysia Perlis, Arau 02600, Malaysia
[3] Department of Biomedical Systems and Informatics Engineering, Yarmouk University, Irbid 21163, Jordan; heyam.q@yu.edu.jo
* Correspondence: wanazani@unimap.edu.my

Medical data include various health indicators, such as physiological signals, images, and treatment histories, providing crucial insights into a patient's condition and disease progression. Computer-aided diagnosis (CAD) systems, encompassing detection, segmentation, and classification, have become integral in modern clinical practice, offering healthcare professionals accurate and efficient diagnostic support. These systems utilize advanced image-processing techniques to ensure reliable and consistent analysis across different medical imaging modalities, including CT, MRI, X-ray, and ultrasound. The incorporation of artificial intelligence (AI), particularly machine learning and deep learning, has revolutionized medical imaging by enabling automated disease detection and classification. However, the development of highly accurate AI models demands large datasets, requiring specialized expertise in medical data processing and analysis.

Many researchers are considering the impacts of employing AI in enhancing the outcomes of medical imaging systems. İrem Çetinkaya et al. [1] assessed the performance of YOLOv8 and the Mask Region Convolutional Neural Network (R-CNN) in detecting fractured endodontic instruments and root canal treatments (RCTs), comparing their effectiveness with experienced endodontists. They used 1050 periapical radiograph images to train and evaluate the performance of both models. Their findings were assessed using various matrices: the accuracy intersection over union (IoU), mean average precision (mAP50), and inference time. YOLO8 achieved 97.4% accuracy, 98.4% mean precision, and a rapid interference time of 14.6 ms, whereas Mask R-CNN showed 98.21% accuracy, 95% mean precision, and 88.4 ms for the interference time. They conclude that both models are appropriate for endodontics; however, YOLO is suitable for real-time applications, while Mask R-CNN is more accurate for pixel-wise segmentation. On the other hand, Se-Yeol Rhyou, Minyung Yu, and Jae-Chern Yoo [2] propos a novel framework, CSM-FusionNet, to enhance the detection of hepatocellular carcinoma from ultrasound images, with their model achieving a promising result in terms of accuracy (97.25%), as well as 100% sensitivity. Taking a different approach, Ahmet Bozdag et al. [3] propose a content-based image retrieval (CBIR) model that combines descriptors from three pre-trained CNN structures, namely, GoogleNet, InceptionV3, and NasNetLarge. They compared performance with two texture-based methods and six CNN models, utilizing cosine similarity to evaluate the similarity between them. The proposed CBIR model outperformed six existing models, achieving an AP (average precision) value of 0.94, demonstrating its effectiveness in detecting gallbladder diseases and making ultrasound-based detection more accessible and efficient.

Received: 5 April 2025
Revised: 20 April 2025
Accepted: 27 April 2025
Published: 28 April 2025

Citation: Mustafa, W.A.; Alquran, H. Editorial for the Special Issue "Advances in Medical Image Processing, Segmentation, and Classification". *Diagnostics* **2025**, *15*, 1114. https://doi.org/10.3390/diagnostics15091114

For skin disease detection, Madallah Alruwaili and Mahmood Mohamed [4] explored the benefits of fused features from three powerful deep learning models, EfficientNet-B0, EfficientNet-B2, and ResNet50. A fusion mechanism operates by passing the extracted features through dense and dropout layers, ensuring better generalization and reducing dimensionality for improved model performance. They obtained a significant result of 99.14% using the 27,153-image Kaggle Skin Diseases Image Dataset. For chest X-ray disease detection, Yi-Ching Cheng et al. [5] employed results obtained using convolutional neural networks (CNNs) and Transformer-based models to specify the most effective architecture for X-ray image analysis. Their outcomes show the effectiveness of CNNs in the detection of abnormal regions in chest X-ray images.

Several researchers conducted experiments on brain tumor detection and classification. Alquran et al. [6] employed a modified version of U-Net structures to segment brain tumors using a huge dataset. The central modifications included the kernel size, the number of channels, the dropout ratio, and changing the activation function from ReLU to Leaky ReLU. The obtained global accuracy was 99.4%, with the most important parameter for accuracy, the Dice similarity, being 90.2%. Deshan Liu [7] focused in their study on brain tumor segmentation utilizing an enhanced super-pixel technique to precisely cluster pathology topological blocks, preventing the common issue of the misgrouping of pixels with low similarity near tumor boundaries. Then, they segmented the entire tumor based on the topological relationships and weights among these blocks. The validation process was conducted using the BraTS 2015 dataset and clinical images from 123 patients. The proposed method achieved high performance, according to the Dice (0.91), Jaccard (0.92), precision (0.90), and recall (0.91) values, in distinguishing tumors from their surrounding regions. On the other hand, in their research, Saravanan Srinivasan et al. [8] focused on brain tumor misdiagnosis, employing three CNN model designs. Their approach achieved high accuracy in tumor detection (99.53%), classification into five types (93.81%), and grading (98.56). Their approach was trained and tested on a huge public database to show the reliability of its use in clinical cases. J. Jebastine [9] investigated brain tumor detection based on a novel convolution extreme gradient boosting model enhanced through Salp Swarm Optimization (CEXGB-ESSO). They assessed the rapid growth of specified tumors and their variability in shape, size, and location, considering issues affecting accurate classification. Their experiment assessed preprocessed MRI images using bilateral filtration, followed by deep feature extraction using a CNN where the last traditional fully connected layer was replaced by the Extreme Gradient Boosting (EXGB) classifier. Enhanced Salp Swarm Optimization (ESSO) was then employed to optimize the model's hyperparameters. The proposed model achieved high performance, with 99% accuracy, 97.52% sensitivity, 98.2% precision, and 97.7% specificity, revealing its efficiency and reliability in the identification and classification of brain tumors.

Moreover, kidney stones represent one of the most familiar diseases of the urinary tract. Traditionally, ultrasound and computed tomography (CT) are the most popular imaging techniques utilized for people who have chronic kidney pain. Alquran et al. [10] designed a fully automated system of segmenting 3D kidney structures and evaluating kidney stones in CT abdominal images; the designed system obtained almost 99% accuracy for a clinical dataset. Erdal Özbay [11] addresses the increasing need for accurate and efficient kidney tumor diagnosis, utilizing a deep learning-based approach with a Masked Autoencoder (MAE) integrated with self-supervised learning and self-distillation (SSLSD-KTD) to successfully categorize kidney tumors, even with limited data availability. The model employs local and global attention mechanisms within its encoder–decoder structure to enhance feature extraction and classification accuracy. It showed remarkable results of 98.04% and 82.14% for accuracy on the KAUH-kidney and CT-kidney datasets, respectively,

with performance further improving to 99.82% and 95.24% with transfer learning. The proposed approach could replace traditional diagnostic methods due to its reliability and robustness. Jorge Gonzalez-Zapata [12] focused on the Guided Deep Metric Learning (DML) approach to enhancing automated kidney stone detection during ureteroscopy, specifically for rare stone types with constrained labeled data. Conventional deep learning methods struggle with such low-data circumstances; therefore, the proposed approach uses a teacher–student framework inspired by Few-Shot Learning. The teacher model (GEMINI) constrains the hypothesis space, exploiting a ResNet50 student model to extract more representative features. The designed experiments conducted on two types of image datasets—stone surface and section—reveal the method's remarkable results, showing 10–12% accuracy gains over existing deep learning and deep machine learning.

Several contributions focus on the classification of cervical cancer using medical images. Alquran et al. [13] enhanced classification accuracy by combining hand craft and deep features with traditional machine learning classifiers, whereas, in a second work, they propose Cervical Net [14], which is based on a feature fusion model that surpasses individual CNNs in recognizing cervical cancer images. Several papers aimed to improve the utilization of artificial intelligence in medical imaging for the detection and classification of COVID-19. Amel Imene Hadj Bouzid [15] investigated the effectiveness of widely used public datasets in training deep learning models to diagnose COVID-19 based on CT scans, using datasets from 13 countries. Several CNNs, including ResNet, DenseNet, and EfficientNet, were trained and evaluated through internal cross-validation and external testing with clinical data. The results emphasize the dilemma of generalization issues due to variations in acquisition conditions and devices. Transfer learning techniques were employed, and the most effective models were customized, yielding an enhanced diagnostic performance in COVID-19 detection. Aboshosha [16] developed an artificial intelligence framework to diagnose, and develop treatment strategies for, COVID-19 through medical imagery, highlighting its automation and clinical applicability. Aggarwal et al. [17] review the most recent developments in COVID-19 image classification employing deep learning, delineating the key advances, current challenges such as data shortages and variability, and prospects for enhancing the generalization and robustness of the proposed models based on previous research perspectives.

In the most recent study, Çağatay Berke Erdaş [18] proposes the use of UNet3+ to localize colorectal abnormalities, followed by utilizing a Cross-Attention Multi-Scale Vision Transformer to distinguish between five types of abnormalities. The proposed UNet3+ achieved a Dice coefficient 0.9872, while the classification model outperformed others, with high accuracy (93.4%) and precision (94.46%). This approach had a great impact in improving colorectal cancer diagnosis. Moreover, Zhihe Zhao et al. [19] propose a deep learning-based for classifying colon diseases using endoscopic images. The data were subjected to preprocessing techniques, and two networks were employed, namely, A_Vit and MobileNet. Both were trained under the same conditions using the Adam optimizer. The A_Vit model incorporated MobileNet, achieving 95.76% accuracy and 97.21% recall.

Focusing on prostate cancer, Yao Zheng et al. [20] introduce a weakly supervised UNet (WSUNet) model that is designed to detect MRI-invisible prostate cancers (MIPCas), which pose a major challenge due to their similarity to normal tissue on MRI. The research was conducted with 777 patients: 600 for training and the remainder for evaluating the performance of the model. Ground truth-labeled prostate biopsies were performed using an MRI–ultrasound fusion system. The validation process was based on biopsy results, achieving an AUC of 0.764 and significantly improving the precision by 91.3% ($p < 0.01$) compared to traditional biopsy methods. Meanwhile, Zhenzhen Dai et al. [21] analyzed biparametric MRI (bp-MRI) scans from 262 prostate cancer patients, grouped into three

cohorts for model development and evaluation. In Cohort 1 (64 patients), histopathology images were used for precise lesion annotation, split into training, validation, and testing sets. Cohort 2 (158 patients) underwent bp-MRI-based lesion delineation and was similarly divided. Cohort 3 included 40 unannotated patients for semi-supervised learning. A non-local Mask (R-CNN) was utilized and enhanced using various training scenarios. Its performance was benchmarked against a baseline Mask (R-CNN), 3D U-Net, and expert radiologist annotations using metrics such as the detection rate, Dice similarity coefficient (DSC), sensitivity, and Hausdorff Distance (HD), demonstrating its effectiveness in prostate lesion segmentation. Pablo Cesar Quihui-Rubio et al. [22] introduced FAU-Net, which is a deep learning model for segmenting prostate zones in MRI images. The proposed model incorporates additive and feature pyramid attention modules, achieving a mean Dice similarity index of 84.15% and an intersection of union of 76.9%, outperforming several other U-Net-based architectures.

Basel Elsayed et al. [23] reviewed the potential of deep learning in the detection of leukemia in pediatric cases from 2013 to 2023 across various countries, reaching the conclusion that artificial intelligence techniques showed effectiveness in achieving a promise result over conventional methods. Meanwhile, A. Khuzaim Alzahrani et al. [24] present a deep learning-based framework for the early detection and classification of leukemia. The proposed model incorporates a novel UNET architecture for segmentation, feature extraction, and classification. The model was evaluated using four datasets, achieving promising results, with 97.82% accuracy and a 98.64% F-score. This approach provides a cost-effective, accurate, and efficient solution that surpasses traditional methods. Morteza MoradiAmin [25] designed an automated system for accurately diagnosing acute lymphoblastic leukemia (ALL), enhancing the images using histogram equalization, segmenting nuclei using fuzzy C-means clustering, then classifying six cell types using a custom convolutional neural network (CNN). The model achieved around 97% classification accuracy—outperforming VGG-16, DenseNet, and Xception—highlighting its effectiveness in ALL detection. Moreover, Syed Ijaz Ur Rahman et al. [26] reviewed the most up-to-date Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines on using AI in ALL detection and classification. Their review focuses on the impact of early WBC analysis based on blood or bone marrow images and categorizes detection approaches into image processing, traditional machine learning, and advanced deep learning models. The review thoroughly evaluates current methodologies and recommends future research directions to support the advancement of effective, AI-driven leukemia diagnostic systems. Md Manowarul Islam [27] proposes an AI-based Internet of Medical Things (IoMT) framework for the automatic detection of acute lymphoblastic leukemia (ALL) based on peripheral blood smear (PBS) images. Through the integration of deep learning into cloud-connected microscopic devices, the system transmits PBS images to a server with a novel fusion model based on a combination of automated features from VGG16 and DenseNet-121. The training phase used 6512 images from 89 individuals, and the model achieved outstanding accuracy (99.89%), precision (99.80%), and recall (99.72%), outperforming existing CNN models. A beta web application simulated this process, emphasizing its potential in achieving precise early leukemia diagnosis.

In gastrointestinal disease diagnosis, Ejaz Ul Haq et al. [28] addressed the high mortality rate of gastric cancer by proposing a deep learning-based classification and segmentation method for endoscopic images. They propose a new model that classifies images into three categories: normal, early gastric cancer, and advanced gastric cancer. Combining the modified GoogLeNet, vision transformer (ViT), and Faster R-CNN models, the proposed system precisely segments and classifies the affected region. The model achieved outstanding results, with 97.4% accuracy, 97.5% sensitivity, and a 95.9% F1-score for classification and

96.7% accuracy, 96.6% sensitivity, and an 95.5% F1-score for segmentation. These findings highlight the impact of the proposed approach in improving gastric cancer diagnosis compared with existing methods. Yiheng Shi et al. [29] conducted a meta-analysis to evaluate the performance of machine models and clinicians in the early diagnosis of gastric cancer. Their analysis integrated 21 articles and assessed the sensitivity, specificity, and receiver operating characteristic (ROC). Machine learning models showed high performance, with a sensitivity of 0.91, specificity of 0.85, and ROC of 0.94 in the training set, and a sensitivity of 0.90, specificity of 0.90, and ROC of 0.96 in the validation set. Specialist clinicians showed a better diagnostic performance than non-specialists; however, with the assistance of machine learning models, non-specialists' sensitivity significantly improved (0.76 vs. 0.64). This study concludes that machine learning models can improve diagnostic accuracy, especially among non-specialist clinicians, providing significant support in the clinical setting in EGC diagnosis during endoscopy.

Fan Li et al. [30] conducted a large-scale analysis in over 8000 inflammatory bowel disease (IBD) patients to study how long-term comorbid conditions affected clinical results. Using hidden class analysis, they recognized patterns of various morbidity in both Crohn's disease and ulcerative colitis. They conclude that patients with hypertension and chronic pain had higher risks of mortality, cardiovascular events, and IBD-related surgeries. Their results underscore the importance of accounting for comorbidity patterns in IBD treatment planning. Finally, Chiraag Kulkarni [31] reviews over 80 recent studies employing artificial intelligence to ulcerative colitis (UC). These studies investigated a range of clinical tasks, including diagnosis, prognosis, biomarker identification, and complication prediction using structured and imaging data, employing various methods, such as random forests, support vector machines, and convolutional neural networks, and assessed their cost-effectiveness to consider whether these tools could be widely adopted in clinical practice.

In conclusion, the integration of artificial intelligence (AI) and deep learning (DL) into medical image processing, segmentation, and classification has revolutionized diagnostic accuracy and efficiency across diverse clinical applications. This Special Issue highlights groundbreaking advancements, such as YOLOv8 and Mask R-CNN for the real-time detection of endodontic fractures, U-Net variants achieving over 99% accuracy in brain tumor segmentation, and hybrid models including CSM-FusionNet for hepatocellular carcinoma detection in ultrasound images. These innovations underscore the potential of AI to augment clinical workflows, reduce human error, and enable rapid decision-making, particularly in resource-constrained settings. However, challenges persist, including the dependency on large, annotated datasets and the limited generalizability of models across diverse populations and imaging modalities. Studies in COVID-19 diagnosis revealed significant performance drops when models trained on public data were tested on external clinical datasets, emphasizing the need for robust data harmonization and transfer learning strategies. Additionally, rare disease detection, such as in MRI-invisible prostate cancers or kidney stones, requires specialized architectures such as WSUNet or SSLSD-KTD to address data scarcity and anatomical complexity.

Future research should prioritize multi-center collaborations to curate diverse, representative datasets and develop lightweight, interpretable models for real-world deployment. Techniques such as federated learning, few-shot learning, and explainable AI (XAI) could enhance model transparency and adaptability. Furthermore, the integration of AI into emerging technologies such as the Internet of Medical Things (IoMT) promises scalable, real-time diagnostic solutions. By addressing these challenges and fostering interdisciplinary innovation, AI-driven systems can transition from experimental tools to indispensable clinical assets, ultimately improving patient outcomes and healthcare accessibility worldwide. The papers in this Special Issue collectively demonstrate the transformative role of

AI in medical imaging while charting a roadmap for overcoming existing limitations and maximizing clinical impacts.

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

Funding: This research received no external funding.

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

References

1. Çetinkaya, İ.; Çatmabacak, E.D.; Öztürk, E. Detection of Fractured Endodontic Instruments in Periapical Radiographs: A Comparative Study of YOLOv8 and Mask R-CNN. *Diagnostics* **2025**, *15*, 653. [CrossRef] [PubMed]
2. Rhyou, S.-Y.; Yu, M.; Yoo, J.-C. Mixture of Expert-Based SoftMax-Weighted Box Fusion for Robust Lesion Detection in Ultrasound Imaging. *Diagnostics* **2025**, *15*, 588. [CrossRef] [PubMed]
3. Bozdag, A.; Yildirim, M.; Karaduman, M.; Mutlu, H.B.; Karaduman, G.; Aksoy, A. Detection of Gallbladder Disease Types Using a Feature Engineering-Based Developed CBIR System. *Diagnostics* **2025**, *15*, 552. [CrossRef]
4. Alruwaili, M.; Mohamed, M. An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification. *Diagnostics* **2025**, *15*, 551. [CrossRef]
5. Cheng, Y.-C.; Hung, Y.-C.; Huang, G.-H.; Chen, T.-B.; Lu, N.-H.; Liu, K.-Y.; Lin, K.-H. Deep Learning-Based Object Detection Strategies for Disease Detection and Localization in Chest X-Ray Images. *Diagnostics* **2024**, *14*, 2636. [CrossRef]
6. Alquran, H.; Alslatie, M.; Rababah, A.; Mustafa, W.A. Improved Brain Tumor Segmentation in MR Images with a Modified U-Net. *Appl. Sci.* **2024**, *14*, 6504. [CrossRef]
7. Liu, D.; Zhang, Y.; Wang, X.; Jiang, Y.; Wang, H.; Fang, L. Brain tumor segmentation algorithm based on pathology topological merging. *Multimed. Tools Appl.* **2024**, *83*, 88019–88037. [CrossRef]
8. Srinivasan, S.; Francis, D.; Mathivanan, S.K.; Rajadurai, H.; Shivahare, B.D.; Shah, M.A. A hybrid deep CNN model for brain tumor image multi-classification. *BMC Med. Imaging* **2024**, *24*, 21. [CrossRef]
9. Jebastine, J. Detection and classification of brain tumor using convolution extreme gradient boosting model and an enhanced salp swarm optimization. *Neural Process. Lett.* **2024**, *56*, 135. [CrossRef]
10. Alquran, H.; Alslity, M.; Qasmieh, I.A.; Alawneh, K.Z.; Alqudah, A.M.; Al-Rasheed, A.; Al-Hawari, M. Three-dimensional kidney's stones segmentation and chemical composition detection. *Int. J. Electr. Comput. Eng.* **2021**, *11*, 3988. [CrossRef]
11. Yang, H.; Wu, X.; Liu, W.; Yang, Z.; Wang, T.; You, W.; Liu, H. CT-based AI model for predicting therapeutic outcomes in ureteral stones after single extracorporeal shock wave lithotripsy through a cohort study. *Int. J. Surg.* **2024**, *110*, 6601–6609. [CrossRef] [PubMed]
12. Gonzalez-Zapata, J.; Lopez-Tiro, F.; Villalvazo-Avila, E.; Flores-Araiza, D.; Hubert, J.; Ochoa-Ruiz, G.; Mendez-Vazquez, A. A metric learning approach for endoscopic kidney stone identification. *Expert Syst. Appl.* **2024**, *255*, 124711. [CrossRef]
13. Alquran, H.; Mustafa, W.A.; Qasmieh, I.A.; Yacob, Y.M.; Alsalatie, M.; Al-Issa, Y.; Alqudah, A.M. Cervical cancer classification using combined machine learning and deep learning approach. *Comput. Mater. Contin.* **2022**, *72*, 5117–5134. [CrossRef]
14. Alquran, H.; Alsalatie, M.; Mustafa, W.A.; Abdi, R.A.; Ismail, A.R. Cervical Net: A Novel Cervical Cancer Classification Using Feature Fusion. *Bioengineering* **2022**, *9*, 578. [CrossRef]
15. Hadj Bouzid, A.I.; Berrani, S.A.; Yahiaoui, S.; Belaid, A.; Belazzougui, D.; Djouad, M.; Tliba, S. Deep learning-based Covid-19 diagnosis: A thorough assessment with a focus on generalization capabilities. *EURASIP J. Image Video Process.* **2024**, *2024*, 40. [CrossRef]
16. Aboshosha, A. AI Based Medical Imagery Diagnosis for COVID-19 Disease Examination and Remedy. *Sci. Rep.* **2025**, *15*, 1607. [CrossRef]
17. Aggarwal, P.; Mishra, N.K.; Fatimah, B.; Singh, P.; Gupta, A.; Joshi, S.D. COVID-19 Image Classification Using Deep Learning: Advances, Challenges and Opportunities. *Comput. Biol. Med.* **2022**, *144*, 105350. [CrossRef]
18. Erdaş, Ç.B. Computer-Aided Colorectal Cancer Diagnosis: AI-Driven Image Segmentation and Classification. *PeerJ Comput. Sci.* **2024**, *10*, e2071. [CrossRef]
19. Zhao, Z.; Gao, Z.; Zhang, K.; Lun, L.; Xu, W.; Wu, H.; Liu, B. Colon Disease Classification Method Based on Deep Learning. In *Advances in Biomedical and Bioinformatics Engineering*; IOS Press: Amsterdam, The Netherlands, 2023; pp. 689–695.

20. Zheng, Y.; Zhang, J.; Huang, D.; Hao, X.; Qin, W.; Liu, Y. Detecting MRI-Invisible Prostate Cancers Using a Weakly Supervised Deep Learning Model. *Int. J. Biomed. Imaging* **2024**, *2024*, 2741986. [CrossRef]
21. Dai, Z.; Jambor, I.; Taimen, P.; Pantelic, M.; Elshaikh, M.; Dabaja, A.; Wen, N. Prostate Cancer Detection and Segmentation on MRI Using Non-Local Mask R-CNN with Histopathological Ground Truth. *Med. Phys.* **2023**, *50*, 7748–7763. [CrossRef]
22. Quihui-Rubio, P.C.; Flores-Araiza, D.; Gonzalez-Mendoza, M.; Mata, C.; Ochoa-Ruiz, G. FAU-Net: An Attention U-Net Extension with Feature Pyramid Attention for Prostate Cancer Segmentation. In Proceedings of the Mexican International Conference on Artificial Intelligence, Tonantzintla, Mexico, 13–18 November 2023; Springer Nature: Cham, Switzerland, 2023; pp. 165–176.
23. Elsayed, B.; Elhadary, M.; Elshoeibi, R.M.; Elshoeibi, A.M.; Badr, A.; Metwally, O.; Yassin, M. Deep Learning Enhances Acute Lymphoblastic Leukemia Diagnosis and Classification Using Bone Marrow Images. *Front. Oncol.* **2023**, *13*, 1330977. [CrossRef] [PubMed]
24. Alzahrani, A.K.; Alsheikhy, A.A.; Shawly, T.; Azzahrani, A.; Said, Y. A Novel Deep Learning Segmentation and Classification Framework for Leukemia Diagnosis. *Algorithms* **2023**, *16*, 556. [CrossRef]
25. MoradiAmin, M.; Yousefpour, M.; Samadzadehaghdam, N.; Ghahari, L.; Ghorbani, M.; Mafi, M. Automatic Classification of Acute Lymphoblastic Leukemia Cells and Lymphocyte Subtypes Based on a Novel Convolutional Neural Network. *Microsc. Res. Tech.* **2024**, *87*, 1615–1626. [CrossRef] [PubMed]
26. Rahman, S.I.U.; Abbas, N.; Ali, S.; Salman, M.; Alkhayat, A.; Khan, J.; Gu, Y.H. Deep Learning and Artificial Intelligence-Driven Advanced Methods for Acute Lymphoblastic Leukemia Identification and Classification: A Systematic Review. *Comput. Model. Eng. Sci.* **2025**, *142*, 1199–1231. [CrossRef]
27. Islam, M.M.; Rifat, H.R.; Shahid, M.S.B.; Akhter, A.; Uddin, M.A. Utilizing Deep Feature Fusion for Automatic Leukemia Classification: An Internet of Medical Things-Enabled Deep Learning Framework. *Sensors* **2024**, *24*, 4420. [CrossRef]
28. Haq, E.U.; Yong, Q.; Yuan, Z.; Jianjun, H.; Haq, R.U.; Qin, X. Accurate Multiclassification and Segmentation of Gastric Cancer Based on a Hybrid Cascaded Deep Learning Model with a Vision Transformer from Endoscopic Images. *Inf. Sci.* **2024**, *670*, 120568. [CrossRef]
29. Shi, Y.; Fan, H.; Li, L.; Hou, Y.; Qian, F.; Zhuang, M.; Miao, B.; Fei, S. The value of machine learning approaches in the diagnosis of early gastric cancer: A systematic review and meta-analysis. *World J. Surg. Oncol.* **2024**, *22*, 40. [CrossRef]
30. Li, F.; Chang, Y.; Wang, Z.; Wang, Z.; Zhao, Q.; Han, X.; Tang, T. Classification of Long-Term Disease Patterns in Inflammatory Bowel Disease and Analysis of Their Associations with Adverse Health Events. *BMC Public Health* **2024**, *24*, 3102. [CrossRef]
31. Kulkarni, C.; Liu, D.; Fardeen, T.; Dickson, E.R.; Jang, H.; Sinha, S.R.; Gubatan, J. Artificial Intelligence and Machine Learning Technologies in Ulcerative Colitis. *Ther. Adv. Gastroenterol.* **2024**, *17*, 17562848241272001. [CrossRef]

diagnostics

MDPI

Article

Detection of Fractured Endodontic Instruments in Periapical Radiographs: A Comparative Study of YOLOv8 and Mask R-CNN

İrem Çetinkaya [1,*], Ekin Deniz Çatmabacak [1] and Emir Öztürk [2]

[1] Department of Endodontics, Faculty of Dentistry, Trakya University, Edirne 22030, Turkey; edenizcatmabacak@trakya.edu.tr
[2] Department of Computer Engineering, Faculty of Engineering, Trakya University Edirne 22030, Turkey; emirozturk@trakya.edu.tr
* Correspondence: irem.cetinkaya@trakya.edu.tr

Abstract: Background/Objectives: Accurate localization of fractured endodontic instruments (FEIs) in periapical radiographs (PAs) remains a significant challenge. This study aimed to evaluate the performance of YOLOv8 and Mask R-CNN in detecting FEIs and root canal treatments (RCTs) and compare their diagnostic capabilities with those of experienced endodontists. **Methods:** A data set of 1050 annotated PAs was used. Mask R-CNN and YOLOv8 models were trained and evaluated for FEI and RCT detection. Metrics including accuracy, intersection over union (IoU), mean average precision at 0.5 IoU (mAP50), and inference time were analyzed. Observer agreement was assessed using inter-class correlation (ICC), and comparisons were made between AI predictions and human annotations. **Results**: YOLOv8 achieved an accuracy of 97.40%, a mAP50 of 98.9%, and an inference time of 14.6 ms, outperforming Mask R-CNN in speed and mAP50. Mask R-CNN demonstrated an accuracy of 98.21%, a mAP50 of 95%, and an inference time of 88.7 ms, excelling in detailed segmentation tasks. Comparative analysis revealed no statistically significant differences in diagnostic performance between the models and experienced endodontists. **Conclusions:** Both YOLOv8 and Mask R-CNN demonstrated high diagnostic accuracy and reliability, comparable to experienced endodontists. YOLOv8's rapid detection capabilities make it particularly suitable for real-time clinical applications, while Mask R-CNN excels in precise segmentation. This study establishes a strong foundation for integrating AI into dental diagnostics, offering innovative solutions to improve clinical outcomes. Future research should address data diversity and explore multimodal imaging for enhanced diagnostic capabilities.

Keywords: fractured endodontic instruments; root canal therapy; YOLOv8; Mask R-CNN; Artificial Intelligence; periapical radiography; diagnostic radiology

Academic Editor: Dechang Chen

Received: 4 February 2025
Revised: 25 February 2025
Accepted: 5 March 2025
Published: 7 March 2025

Citation: Çetinkaya, İ.; Çatmabacak, E.D.; Öztürk, E. Detection of Fractured Endodontic Instruments in Periapical Radiographs: A Comparative Study of YOLOv8 and Mask R-CNN. *Diagnostics* **2025**, *15*, 653. https://doi.org/10.3390/diagnostics15060653

1. Introduction

The identification and accurate localization of fractured endodontic instruments (FEIs) in root canal treatment (RCT) represent a significant diagnostic challenge that directly impacts clinical outcomes. The presence of FEIs obstructs the effective cleaning and disinfection of root canals, thereby compromising the success of endodontic therapy. These challenges point to the critical need for advanced diagnostic techniques to improve precision and reliability in clinical practice [1].

Radiopaque materials aid in detecting FEIs in empty canals but pose challenges in complete RCT, where FEIs and filling materials share similar radiopacity [2]. Comprehensive

studies addressing the radiopacity differences between FEIs and root canal fillings, as well as the influence of factors such as anatomical superimposition, geometric distortion, and material-induced artifacts, remain limited in the literature [3]. Advances in digital imaging, including high-resolution sensors and specialized filters, have improved the visibility of small instrument fragments. However, detection remains subjective, time-intensive, and reliant on clinician expertise and image quality [4]. Consequently, alongside advanced diagnostic techniques, the integration of artificial intelligence (AI) models presents a promising approach to improve the accuracy of radiographic examinations and support clinicians in addressing FEI-related challenges [3].

Radiology provides a direct entry point for AI algorithms in dentistry as its digitally encoded images facilitate their application. These algorithms effectively identify variations in tissue density and dental structures, while also detecting and localizing complex features in medical imaging, improving diagnostic accuracy [5].

Recent advancements in convolutional neural networks (CNNs) have shown significant potential in processing large imaging data sets. These algorithms effectively identify variations in tissue density and dental structures, while also detecting and localizing complex features in medical imaging, improving diagnostic accuracy [5].

Among these models, Mask R-CNN and YOLO are notable for their unique capabilities. While previous studies have demonstrated the success of both models in object detection and segmentation tasks [6], these structural differences position Mask R-CNN as more suitable for tasks requiring detailed object-background differentiation, whereas YOLOv8's rapid detection capability offers a distinct advantage in scenarios demanding quick responses.

Beyond these models, CNNs and long short-term memory networks have recently been employed for the detection of FEIs, achieving notable success. For instance, CNN-based models have shown sensitivity of approximately 81% and specificity up to 87% in identifying FEIs on panoramic images [7]. Similarly, a study employing Mask R-CNN on PAs achieved a mAP exceeding 98%, highlighting its potential for accurate segmentation and localisation of fractured instruments [5]. These methods theoretically enhance diagnostic efficiency by segmenting and localizing fractured instruments. While the diagnostic capabilities of AI models are well-documented, the inclusion of observer comparisons introduces a crucial dimension for assessing their practical applicability in clinical contexts.

The aim of this study is to evaluate the performance of Mask R-CNN and YOLOv8 models in the diagnosis of FEIs and RCTs and to compare their diagnostic capabilities with those of experienced endodontists, thereby assessing their suitability for clinical application.

2. Materials and Methods

The research protocol was approved by the Non-interventional Clinical Research Ethical Committee of Trakya University (TÜTF-GOBAEK 2024/384) and was conducted in accordance with the principles of the Helsinki Declaration. Patient names were excluded, and all data were anonymized to ensure confidentiality and ethical compliance.

2.1. Data Collection

PAs were selected as the study material due to their accessibility during and after RCT and their common use in cases with complications. Patients from the Department of Endodontics at Trakya University who underwent PA examinations were categorized into the following groups:

(1) FEIs with RCT
(2) FEIs without RCT
(3) Teeth with complete RCT

(4) Teeth without RCT or FEIs

The radiographic images were captured using a Planmeca Pro X™ 2D intraoral X-ray unit (Planmeca®, Helsinki, Finland) equipped with a size 2 photostimulable phosphor plate (PSP) detector. Exposure parameters were standardised at 65 kVp, 7 mA, and 0.2 s. Images were digitised using the VistaScan Mini Easy system (Dürr, Biertigheim-Bissingen, Germany), ensuring uniform and reliable imaging quality across all samples.

A total of 396 teeth with FEIs were initially identified, but 16 were excluded for failing to meet quality standards for PAs. Criteria for quality included adequate resolution, absence of distortions, and clear visibility of key anatomical landmarks such as the root apex and canal structure. Radiographs exhibiting artefacts such as cone cuts or motion blurring were excluded. Additionally, 360 teeth with complete RCT and no instrument fractures, as well as 310 teeth without RCT or FEIs, were included. All data were standardised to a resolution of 788 × 612 pixels and stored in PNG format.

2.2. Ground Truth Determination and Observer Agreement

Annotations were performed using an open-source tool (CVAT, version 2.2.0). Two endodontic specialists, with 3 and 10 years of experience, respectively, annotated each radiograph for the presence of FEIs and/or RCT using detailed polygonal labelling. Radiographs without these features were left unannotated (Figure 1). To ensure compatibility with analytical models, polygonal annotations were converted to bounding boxes, as Mask R-CNN and YOLO algorithms primarily operate using this format. While Mask R-CNN can generate pixel-level segmentation masks, it first identifies objects using bounding boxes before applying segmentation. YOLO focuses exclusively on bounding boxes for rapid detection [8].

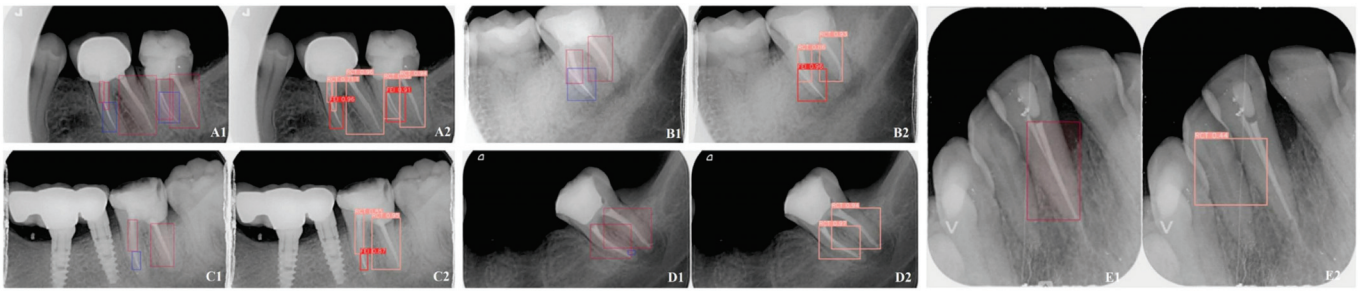


Figure 1. Representative examples of Mask R-CNN's performance on periapical radiographs (PAs) for detecting fractured endodontic instruments (FEI) and root canal treatments (RCT). The bounding boxes and associated confidence scores highlight the model's ability to accurately identify and localize objects. Panels (**A1**–**E1**) represent the ground truth annotations marked with blue boxes for FEI and red boxes for RCT, while panels (**A2**–**E2**) depict the segmentations generated by the Mask R-CNN model, where FEI is marked with red boxes and RCT with pink boxes.

To ensure high inter-observer reliability and accurate bounding box annotations for model training, the Inter-Class Correlation (ICC) analysis was applied with a threshold of 0.9. Radiographs meeting this threshold were considered sufficiently reliable for inclusion in the study. For annotations with ICC values below 0.9, indicating discrepancies between observers, a re-evaluation process was implemented to address these inconsistencies. This comprehensive approach is particularly important in medical imaging studies, where precise localization and accurate annotations are essential for model performance and validity.

In instances where re-evaluation did not resolve disagreements, a third observer was consulted to finalize the annotations. In cases requiring third observer input, the decision-making process was guided by predefined criteria. The third observer assessed 11 radio-

graphs that required further review due to unresolved discrepancies. These disagreements predominantly arose in images where both RCT and FEIs were present, presenting added interpretative complexity.

The final data set included 1050 teeth with corresponding annotations, demonstrating excellent inter-observer agreement and serving as the ground truth for model training. Among these, FEIs were identified in 37.7% (n = 396), providing a sample size for robust model evaluation.

2.3. Model Selection and Training

Mask R-CNN was chosen for its ability to delineate precise boundaries, making it particularly suited for detecting FEIs within complex anatomical structures. Its advanced segmentation capabilities were advantageous in identifying FEIs within radiopaque areas. YOLOv8 was selected for its rapid detection capabilities, offering practical benefits for clinical decision-making [9–11].

Mask R-CNN employs a two-stage segmentation approach for object detection. In the first stage, the Region Proposal Network (RPN) identifies potential object regions. In the second stage, these regions undergo further analysis: one layer classifies the object, another refines the bounding box, and a third generates the binary mask. To enhance segmentation accuracy, the Region of Interest (RoI) Align layer is utilized [12]. The Mask R-CNN architecture is shown in Figure 2.

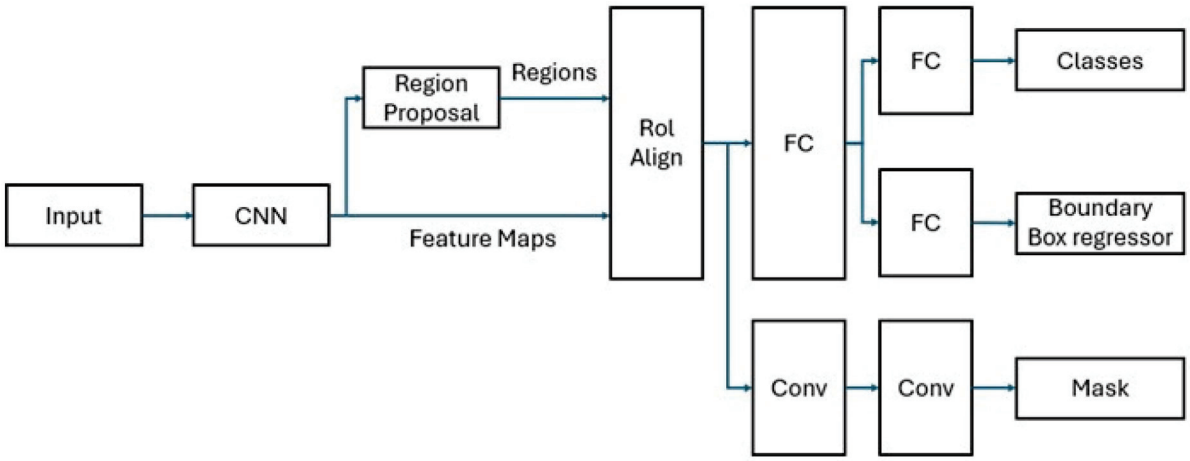


Figure 2. Flowchart of Mask R-CNN architecture. CNN extracts feature maps from the input image. The Region Proposal Network generates candidate regions, which are processed through RoI (Region of Interest) Align to ensure accurate spatial alignment. The extracted features are passed through FC (Fully Connected) layers for classification and bounding box regression. Additionally, Conv (Convolutional) layers are used for mask prediction.

YOLO processes the entire image in a single pass, simultaneously computing bounding boxes and class probabilities using a CNN-based architecture. The CNN output is fed into a prediction stage that determines bounding box coordinates and class probabilities. Due to its single-pass design, YOLO is optimized for real-time applications, prioritizing speed [13]. The YOLO architecture is shown in Figure 3.

The evaluation was performed on a system featuring a Ryzen 7 5700 x processor, 32 GB RAM, and an RTX 4070 Ti GPU, operating on the Windows platform. The data set was divided into training (60%), validation (20%), and test (20%) sets, ensuring class balance across all subsets. Both models were implemented within a structured framework and trained on a curated set of annotated radiographs to ensure accurate detection and

localization of FEIs and RCTs. The training process incorporated hyperparameter tuning to optimize performance.

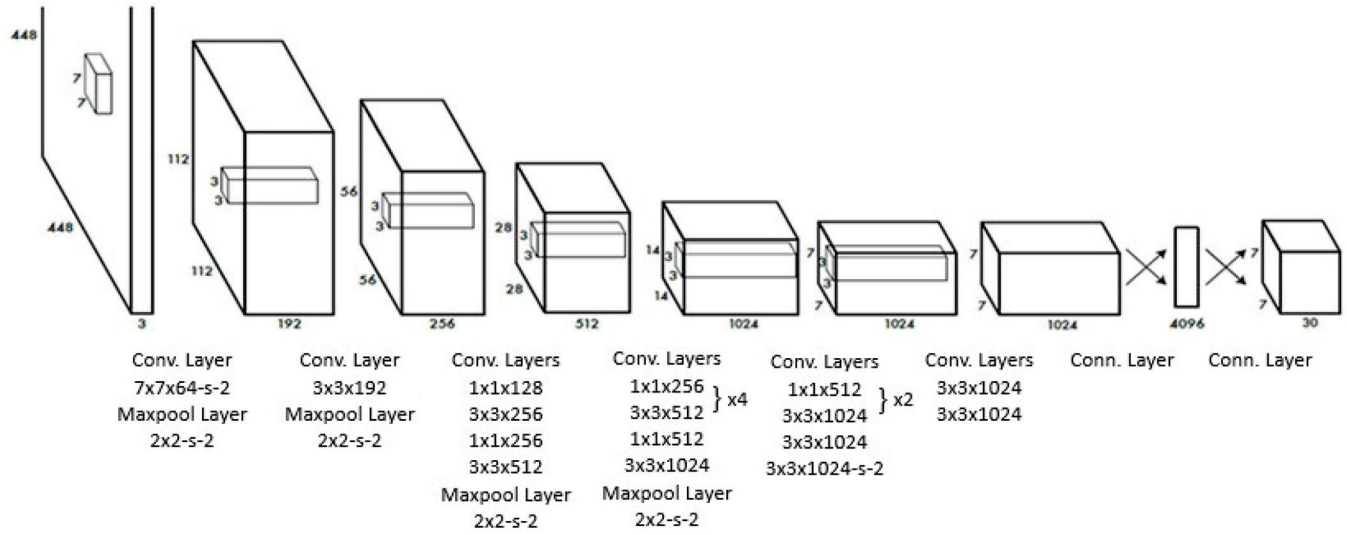


Figure 3. Flowchart of YOLO architecture.

For YOLOv8, key hyperparameters such as the learning rate (initially set at 0.001), batch size (set to 32 samples per batch), and intersection over Union (IoU) threshold (fixed at 0.5) were optimized using a grid search strategy to achieve optimal detection performance. Similarly, for Mask R-CNN, the backbone architecture (ResNet-50), the learning rate schedule (step decay starting at 0.002), and the number of epochs (set to 30) were fine-tuned to enhance segmentation accuracy. Data augmentation techniques were applied to enhance model robustness against variations in radiographic quality and noise. These augmentations help the model generalize better by introducing real-world variations commonly observed in clinical settings, such as horizontal/vertical flipping (50% probability), rotation ($\pm$ 15° in 10° increments), brightness/contrast adjustments ($\pm$20%), Gaussian noise, translation (10% of image size, 10–20% probability), and shear (2°). Rotation ensures diversity in image capture angles without excessive distortion, while translation and shear introduce minor positional shifts reflective of patient or device variability. Flipping leverages dental symmetry to improve feature representation and data set diversity.

2.4. Box Loss Calculation

During training, box loss was used to evaluate the distance between predicted bounding boxes and ground truth boxes. L1 and L2 losses were used to minimize localization errors and improve bounding box precision. This metric played a crucial role in refining model accuracy throughout the training process.

2.5. Follow-Up Analysis

Six months after the initial annotation, the two observers re-evaluated the data set, and their updated annotations were compared with the predictions of YOLOv8 and Mask R-CNN using confusion matrices. These matrices facilitated a detailed analysis of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) across the classes (FEIs, RCTs, BG), offering insights into observer consistency, model performance, and areas where the models might face challenges. This comparison provided valuable metrics for assessing the effectiveness of the models in detecting and classifying these categories, as well as identifying areas requiring further refinement.

2.6. Methods for Evaluating Model Effectiveness

These matrices offered insights into the models' ability to detect FEIs, RCTs, and background (BG). Key metrics—accuracy, IoU, mAP50, and interference time—are crucial for assessing model performance in detecting and localizing RCTs and FEIs in PAs. Each is detailed below.

Accuracy (Equation (1)) measures the proportion of correct predictions, including both TPs and TNs, out of all predictions made by the model. IoU (Equation (2)), ranging from 0 to 1, was used to measure overlap and assess localization accuracy between predicted and ground truth bounding boxes. mAP50 (Equation (3)) computes average precision at an IoU threshold of 0.50, requiring at least 50% overlap for correct detection. Inference time refers to the time taken by the model to generate a prediction after receiving input. By selecting mAP50 as evaluation metric, we aim to balance accuracy with practical applicability, ensuring our models are both effective and deployable in diverse clinical scenarios.

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

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{2}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{3}$$

2.7. Statistical Analysis

A Z-test was conducted to compare accuracy, IoU, and mAP50 between YOLOv8 and Mask R-CNN. Given the sufficiently large sample size (n = 1050), the normality assumption was considered valid. The test was based on the null hypothesis (H_0) that no significant difference exists between the models, while the alternative hypothesis (H_1) proposed a statistically significant difference. A two-tailed Z-test was applied, with statistical significance set at $p < 0.05$. All statistical analyses were performed using Microsoft Excel (Microsoft 365, Version 2501, 64-bit).

3. Results

3.1. Model Performance

The evaluation of YOLO v8 and Mask-R-CNN models revealed consistently high performance across key metrics, showcasing their suitability for detection and segmentation tasks (Table 1). Accuracy rates were observed at 97.40% for YOLO v8 and 98.21% for Mask-R-CNN. Despite the slight numerical advantage of Mask-R-CNN, the difference in accuracy between the two models was not statistically significant (p = 0.571). Similarly, IoU scores of both models were remarkably high, with YOLO v8 achieving a perfect score of 100% and Mask-R-CNN scoring 99%. This difference also lacked statistical significance (p = 0.146), indicating comparable segmentation precision between the two methods.

In contrast, mAP50 metric highlighted a statistically significant difference (p = 0.020). YOLO v8 outperformed Mask-R-CNN with a mAP50 of 98.9%, compared to 95% for Mask-R-CNN, suggesting a superior ability to correctly detect and classify objects across varying confidence thresholds.

Another notable distinction was observed in computational efficiency, as measured by interference time. The YOLO v8 model demonstrated a significant advantage in speed, with an interference time of only 14.6 ms, far surpassing the Mask-R-CNN model's 88.7 ms. This substantial reduction in processing time underscores YOLO v8's potential for real-time applications, where rapid decision-making is critical.

Table 1. Comparison of Performance Metrics between YOLO v8 and Mask-R-CNN Models for Detection and Segmentation.

	Accuracy	IoU (Intersection Over Union)	mAP50	Interference Time
YOLO v8	0.973975	1.00	0.989	14.6 ms
Mask-R-CNN	0.982075	0.99	0.950	88.7 ms
p	0.571	0.146	0.020	

3.2. Algorithmic Overview of YOLOv8 and Mask R-CNN

Explainable AI (XAI) techniques were employed to enhance the interpretability and reliability of object-detection results. Saliency mapping was utilized to identify the most influential regions contributing to the detection of FEIs, providing valuable insights into model transparency and robustness. Based on the documentation [14], the Detector-RISE (D-RISE) saliency algorithm was selected as it aligns with the object-detection approach used in this study. The results were generated accordingly.

In the saliency map, red areas indicate the regions that contribute most to object recognition, while blue areas correspond to regions with minimal relevance. Changes and variations in the red areas are considered to have the highest impact on the decision-making process. In cases where multiple classes are present in the examples, saliency values are provided for a single class. Apart from this, saliency graphs can be calculated separately for each value [15].

For YOLO and Mask-R-CNN models, saliency map outputs for FEI and RCT examples are presented in Figure 4.

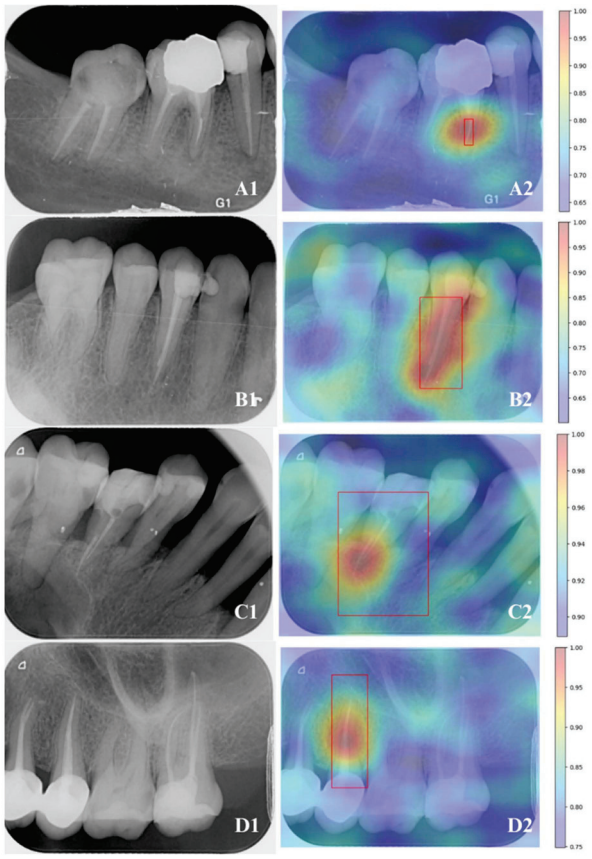


Figure 4. Saliency map outputs for FEI and RCT detection using YOLO and Mask R-CNN. (**A1**–**D1**) Raw periapical radiographs, (**A2**–**D2**) corresponding saliency maps. (**A**) YOLO-based saliency map for FEI detection, (**B**) YOLO-based saliency map for RCT detection, (**C**) Mask R-CNN-based saliency

map for FEI detection, and (**D**) Mask R-CNN-based saliency map for RCT detection. The red boxes indicate the regions identified by the models as containing FEI or RCT, highlighting the areas of interest detected by the respective deep learning approaches.

3.3. Training and Validation Stability

YOLO v8—Box Loss: Both training and validation box loss steadily decreased, approaching 0.5, indicative of progressive improvements in bounding box predictions. The close alignment between training and validation losses suggests consistent performance across data sets and minimal overfitting (Figure 5A).

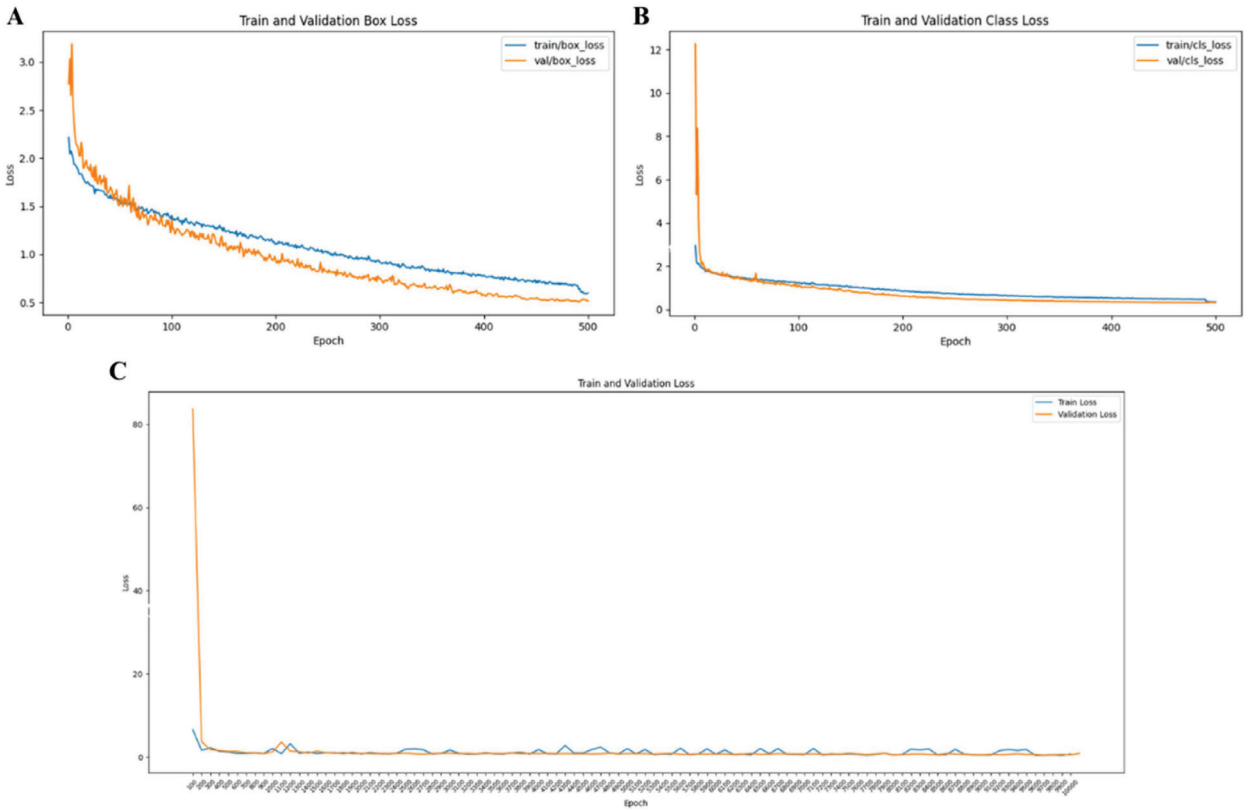


Figure 5. Comparison of training and validation losses for YOLOv8 (top) and Mask R-CNN (bottom) models. The YOLOv8 graphs depict box loss (**A**) and class loss (**B**), illustrating a steady decrease in both training and validation losses with minimal divergence, indicating strong generalization and effective performance in object localization and classification. In contrast, the Mask R-CNN graph (**C**) shows the total loss across training and validation, with training loss decreasing rapidly and validation loss stabilizing with slight fluctuations, reflecting its ability to perform detailed segmentation tasks. Overall, YOLOv8 demonstrates faster convergence and smoother loss reduction, while Mask R-CNN exhibits robustness in tasks requiring precise segmentation.

YOLO v8—Class Loss: Class loss decreased rapidly during training, reflecting improved accuracy in classifying objects. The alignment of training and validation losses demonstrated the model's robustness and reduced likelihood of overfitting in classification tasks (Figure 5B).

Mask-R-CNN—Overall Loss: Validation loss peaked early but stabilised rapidly, indicating improved object localisation and classification accuracy. The close alignment of training and validation losses highlighted the model's strong generalisation capabilities (Figure 5C).

3.4. Comparative Performance Between Models and Endodontists

The comparative analysis between the models and endodontists revealed no statistically significant differences, reflecting the consistency and reliability of both human and machine performance (Table 2). For the FEI F1 score, Endodontist A and Endodontist B achieved identical values of 0.9947 ($p = 1.000$). This uniformity extended to comparisons between Endodontist A and the machine learning models, as no significant differences were found for either Mask-R-CNN ($p = 0.640$) or YOLO v8 ($p = 0.447$). Similarly, Endodontist B's performance did not significantly differ from Mask-R-CNN ($p = 0.640$) or YOLO v8 ($p = 0.447$).

Specifically, the diagnostic accuracy values for Endodontist A and Endodontist B exhibited slight variations over the 6-month period (Table 2). Accuracy values were as follows: 0.9947 at the initial assessment and 0.9947 after 6 months for both endodontists, followed by 0.9944 at both time points. A decrease in diagnostic accuracy was observed in a particular instance, with values dropping to 0.967 for both endodontists. These findings suggest a degree of temporal inconsistency in human diagnostic decisions, focusing on the influence of cognitive and interpretative factors over time.

The RCT F1 score further supported these findings, as no significant differences were detected among the endodontists or between the models ($p > 0.05$). These results suggest that both models perform at a level comparable to experienced practitioners, further validating their applicability in clinical or research settings.

Within-group comparisons of F1 scores also showed no significant variations across metrics. For example, the BG metric did not significantly differ from FEI ($p = 0.971$) or RCT ($p = 0.064$). This lack of statistically significant differences across multiple comparisons indicates a high level of consistency and robustness in both human and model performance.

YOLO v8 and Mask-R-CNN demonstrated performance metrics comparable to skilled endodontists, with added benefits of automation and scalability. YOLO v8 outperformed Mask-R-CNN in computational efficiency and mAP50, making it more suitable for high-speed, accurate applications.

Table 2. Inter- and Intra-group Comparisons of F1 Scores for Detection of FEI, RCT, and BG by Endodontist A, Endodontist B, Mask R-CNN, and YOLOv8.

	Endodontist A	**Endodontist B**	**Mask-R-CNN**	**YOLOv8**	**A-B (*p*)**	**A-Mask-R-CNN (*p*)**	**A-YOLOv8 (*p*)**	**B-Mask-R-CNN (*p*)**	**B-YOLOv8 (*p*)**	**Mask-R-CNN-YOLOv8 (*p*)**
BG F1 Score	N/A	N/A	0.9893	N/A	---	---	---	---	---	---
FEI F1 Score	0.9947	0.9947	0.9890	1.0000	1.000	0.640	0.447	0.640	0.447	0.272
RCT F1 Score	0.9944	0.9944	1.0000	0.9964	1.000	0.447	0.832	0.447	0.832	0.542
BG- FEI (*p*)	---	---	0.971	---						
BG- RCT (*p*)	---	---	0.064	---						
FEI- RCT (*p*)	0.967	0.967	0.126	0.382						

4. Discussion

This study demonstrates the accuracy and reliability of YOLOv8 and Mask R-CNN models in detecting and localizing FEIs and RCTs. Both models demonstrated robust performance, with YOLOv8 and Mask R-CNN effectively identifying all FEI cases while maintaining minimal error rates in RCT classification. The reported mAP50 scores for YOLOv8 and Mask R-CNN align closely with results from similar studies [5,16,17], further validating their potential as clinical diagnostic tools.

Accurate identification of FEIs is crucial for effective treatment planning, particularly for less experienced clinicians, as their radiopacity often overlaps with that of RCT materials. A key focus of this study was the challenge of distinguishing between these similarly radiopaque structures, a topic not extensively explored in prior research [3].

Once a FEI is detected, the treatment approach will depend on the affected tooth and canal, the location of the instrument separation, the amount of remaining contaminated material, and the potential damage to the dental structure if removal is attempted [18]. If accessible in the coronal or middle third, retrieval may be attempted using ultrasonic tips or a specialized device [19]. When removal risks excessive dentin loss or perforation, bypassing the fragment allows for successful obturation [20]. In apical third cases where removal is unfeasible, sealing the canal with biocompatible materials minimizes bacterial leakage and preserves endodontic success [21]. AI-assisted detection enhances treatment planning by providing precise localization, aiding clinicians in selecting optimal management strategies [5].

To ensure data set reliability and minimize observer bias, the study employed a strict methodology, exemplified by a high ICC threshold. By introducing an additional layer of complexity, this study required the models to differentiate between multiple scenarios within a single radiograph, including RCT and/or FEIs.

This study utilized YOLOv8 and Mask R-CNN, capitalizing on their respective strengths and differences in dental diagnostics (Table 3). YOLOv8 exhibited exceptional performance in rapid detection and classification, achieving a significantly faster inference time (14.6 ms) and an impressive mAP50 score of 0.989. These attributes may render it particularly well-suited for real-time clinical applications where both efficiency and accuracy are critical. Similarly, Mask R-CNN excelled in precise segmentation tasks, offering detailed visualization of complex dental structures, albeit with a slower inference time (88.7 ms). The slower performance of Mask R-CNN compared to YOLOv8 is not surprising, as Mask R-CNN involves more complex processes, including region proposal generation and pixel-wise segmentation, whereas YOLOv8 is optimized for real-time detection with a single network pass [16,22]. The performance of the models was further assessed by analyzing TP, FP, and FN values for both YOLO and Mask R-CNN, whereas TN values were omitted, as examples without annotations were considered unnecessary (Supplementary Materials, Figures S1 and S2).

A notable finding in this study was the statistically significant difference in mAP50 between the models, emphasizing YOLOv8's strength in detection accuracy. Future research may benefit from integrating mAP50 with complementary metrics to provide a more comprehensive understanding of model performance across varying scenarios. Additionally, exploring the impact of alternative IoU thresholds could further enhance the adaptability of object-detection systems [23,24].

While mAP50 was the only metric showing a statistically significant difference, this finding should be considered alongside other performance metrics to ensure a holistic evaluation of model capabilities. Different metrics reveal distinct capabilities of models and play equally critical roles in determining their suitability for specific clinical applications. The results underscore the importance of adopting multidimensional evaluation

frameworks to fully understand the trade-offs between speed, accuracy, and precision in object-detection models.

Table 3. Comparison of YOLOv8 and Mask R-CNN in Object Detection and Segmentation.

Feature	YOLOv8	Mask R-CNN
Detection Architecture	Single-stage detector that processes the entire image in a single pass.	Two-stage detector that first generates region proposals and then refines detections and segments objects.
Network Architecture	Uses a CNN backbone with a unified prediction head for bounding boxes and class probabilities.	Utilizes a CNN backbone with a Region Proposal Network (RPN), followed by RoI Align and separate branches for classification, box regression, and mask prediction.
Computational Efficiency	Optimized for speed and efficiency, making it suitable for real-time applications.	More computationally demanding due to its two-stage process, leading to higher precision but slower inference.
Detection and Segmentation Output	Outputs bounding boxes and class scores, with instance segmentation added after v8.	Produces bounding boxes, class labels, and masks for pixel-level segmentation.

Numerous studies highlight the effectiveness of YOLOv8 and Mask R-CNN in detecting and segmenting pathologies in dental radiographs. YOLOv8 has achieved 77.03% accuracy in classifying periodontal diseases in bitewing radiographs and 75% in PAs [25], while our study demonstrated a significantly higher accuracy of 97.4% in detecting FEIs. Similarly, the YOLOv8 m model attained 90% accuracy in diagnosing dental diseases from bitewing and orthopantomographic images, aligning with our findings [22]. Additionally, YOLOv8 reached 95.2% accuracy and 97.5% mAP50 in segmenting mandibular radiolucent lesions, comparable to our results [16]. A study has shown that while CNNs achieve higher classification performance, YOLOv8 exhibits a trade-off between speed and accuracy, with lower F1 scores in detecting apical and peri-endo combined lesions [26]. Additionally, research on the automated detection of dental conditions using YOLOv8 has reported precision and recall values exceeding 80%. However, its performance declines when applied to external data sets, highlighting challenges in generalizability [24].

Mask R-CNN studies underline its reliability in segmentation and anomaly detection. Wang et al. (2024) integrated Mask R-CNN with a neural network classifier for diagnosing periapical diseases, achieving a pixel accuracy exceeding 97% [27]. An attention-enhanced Mask R-CNN achieved 79.5% accuracy [28], similar to the accuracy observed in our FEI detection. Other studies reported 75–80% accuracy in identifying dental caries and periodontitis, highlighting its utility in dental radiographic analysis [29,30]. For FEI detection in PAs, Mask R-CNN achieved 98.8 mAP, 95.2% accuracy, and 97% F1 scores, closely aligning with our findings [5]. Furthermore, it demonstrated 100% accuracy and 97.49 mAP in tooth segmentation and numbering, underscoring its precision and clinical applicability [17].

The comparative analysis between endodontists and AI models revealed comparable levels of diagnostic accuracy, with both achieving high performance metrics. However, a notable finding was the variation observed in the diagnoses made by endodontists after

a 6-month interval compared to their initial assessments. These results indicate that while endodontists and AI models demonstrated comparable diagnostic accuracy, temporal variability was evident in endodontists' assessments over a 6-month period. This fluctuation suggests the influence of cognitive and interpretative factors over time. Such variability may be attributed to memory effects, evolving clinical judgment, or shifts in diagnostic perception. These results highlight the inherent subjectivity in human interpretation and underscore the need for standardized evaluation protocols to enhance diagnostic consistency and reliability in endodontic practice. This highlights the potential for variability in human interpretations over time, likely influenced by memory effects, shifts in judgment, or changes in clinical interpretation.

Specifically, while the overall F1 scores of the endodontists remained high, minor discrepancies were identified in certain cases, particularly in the interpretation of FEIs. This variability emphasizes the inherent subjectivity and temporal variability of human diagnoses, even among experienced practitioners.

In contrast, YOLOv8 and Mask R-CNN demonstrated consistent diagnostic performance over the same data set, unaffected by temporal or cognitive factors. This consistency highlights one of the key advantages of AI systems: their ability to provide stable and reproducible diagnostic outcomes, free from the influence of human-related factors such as fatigue or shifting perceptions.

These findings suggest that a combined approach leveraging both AI models and human expertise could enhance diagnostic accuracy. AI models can serve as reliable and consistent second opinions, helping to mitigate variability in human diagnoses and ensuring a more robust diagnostic process over time.

Despite the high success rate, certain limitations must be acknowledged. The relatively low prevalence of FEIs (37.7%) constrains the data set size, potentially limiting the models' robustness in real-world applications. However, this limitation reflects the fact that FEIs are relatively rare in clinical settings, which inherently affects data set composition [1]. Additionally, noise and artifacts in PAs, such as anatomical superimpositions and material distortions, present challenges for accurate detection and localization. These limitations emphasize the need for enhanced data quality and more robust training strategies to improve model performance under diverse clinical conditions.

AI, shaped by training data, can reflect inherent biases, necessitating human oversight to ensure fairness. While transparency fosters trust, it also poses privacy risks, highlighting the need for a balance between openness and data protection [31,32]. A significant challenge lies in bridging the gap between AI research and clinical application, as most studies remain experimental. The absence of standardized data protocols and inconsistent labeling further exacerbate biases, limiting AI's reliability in clinical settings [33,34]. Reliance on AI as a stand-alone diagnostic system without oversight is ethically contentious and may have legal implications. Addressing these concerns requires clinicians to ensure patient safety, acquire ethical AI usage skills, and advocate for robust legal frameworks to guide AI integration [32]. Ultimately, developing unbiased, transparent AI systems with robust human oversight is essential for ensuring ethical implementation and maximizing clinical utility.

Future research should focus on expanding the data set to include a more diverse range of clinical cases. Additionally, integrating multimodal imaging techniques, such as combining PA with cone-beam computed tomography or panoramic radiographs, could provide a more comprehensive diagnostic framework. Exploring the potential of ensemble learning by combining YOLOv8 with advanced segmentation models could further enhance detection and localization performance. This study establishes a strong foundation for the clinical application of YOLOv8 and Mask R-CNN in dental diagnostics. Addressing the outlined limitations and expanding future research will be vital to fully realizing

their potential across diverse clinical contexts. Such future research will contribute to the transformative and innovative impact of AI in dentistry.

5. Conclusions

This study demonstrates the clinical potential of YOLOv8 and Mask R-CNN in detecting and localizing FEIs and RCTs with high accuracy and reliability. By leveraging the complementary strengths of these models—YOLOv8 for rapid detection and Mask R-CNN for precise segmentation—the research provides a robust framework for integrating advanced AI models into routine dental diagnostics. The inclusion of observer comparisons further demonstrates the practical applicability of these AI models by benchmarking their performance against clinician expertise.

This study serves as a step forward in the integration of AI into dental practice, offering innovative solutions to long-standing diagnostic challenges. Integrating these models into clinical workflows can enhance treatment planning and patient outcomes, particularly for less experienced clinicians. Expanding research in this field will be critical for maximizing the transformative potential of AI in dentistry and ensuring its ethical and effective implementation in diverse clinical settings.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics15060653/s1, Figure S1: YOLO Model Detection Results; Figure S2: Mask R-CNN Model Detection Results.

Author Contributions: Conceptualization, İ.Ç. and E.D.Ç.; Methodology, İ.Ç., E.D.Ç., and E.Ö.; Software, İ.Ç., E.D.Ç., and E.Ö.; Validation, E.Ö.; Formal analysis, İ.Ç., E.D.Ç., and E.Ö.; Investigation, İ.Ç., E.D.Ç., and E.Ö.; Resources, İ.Ç. and E.D.Ç.; Data curation, İ.Ç. and E.D.Ç.; Writing—original draft preparation, İ.Ç. and E.D.Ç.; Writing—review and editing, İ.Ç. and E.D.Ç.; Visualization, E.Ö.; Supervision, İ.Ç.; Project administration, İ.Ç. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Trakya University (protocol code TÜTF-GOBAEK 2024/384 and date of approval 23 September 2024).

Informed Consent Statement: Patient names and identifying details were excluded, and all data were fully anonymized to maintain confidentiality. As the study involved retrospective analysis of anonymized radiographs, specific patient consent was not required.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. Due to ethical restrictions, access to the raw data sets is limited to ensure the confidentiality of patient information.

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

Abbreviations

The following abbreviations are used in this manuscript:

FEI	Fractured Endodontic Instrument
RCT	Root Canal Treatment
AI	Artificial Intelligence
CNN	Convolutional Neural Network
MAP	Mean Average Precision
PA	Periapical Radiograph
ICC	Inter-Class Correlation
BG	Background

IOU	Intersection Over Union
RPN	Region Proposal Network
ROI	Region of Interest
FC	Fully Connected
Conv	Convolutional
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

References

1. Panitvisai, P.; Parunnit, P.; Sathorn, C.; Messer, H.H. Impact of a retained instrument on treatment outcome: A systematic review and meta-analysis. *J. Endod.* **2010**, *36*, 775–780. [CrossRef] [PubMed]
2. Gorduysus, M.; Avcu, N. Evaluation of the radiopacity of different root canal sealers. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontology* **2009**, *108*, e135–e140. [CrossRef] [PubMed]
3. Brito, A.C.R.; Verner, F.S.; Junqueira, R.B.; Yamasaki, M.C.; Queiroz, P.M.; Freitas, D.Q.; Oliveira-Santos, C. Detection of fractured endodontic instruments in root canals: Comparison between different digital radiography systems and cone-beam computed tomography. *J. Endod.* **2017**, *43*, 544–549. [CrossRef]
4. Jayachandran, S. Digital imaging in dentistry: A review. *Contemp. Clin. Dent.* **2017**, *8*, 193–194. [CrossRef]
5. Özbay, Y.; Kazangirler, B.Y.; Özcan, C.; Pekince, A. Detection of the separated endodontic instrument on periapical radiographs using a deep learning-based convolutional neural network algorithm. *Aust. Endod. J.* **2024**, *50*, 131–139. [CrossRef]
6. Bonfanti-Gris, M.; Herrera, A.; Paraíso-Medina, S.; Alonso-Calvo, R.; Martínez-Rus, F.; Pradíes, G. Performance evaluation of three versions of a convolutional neural network for object detection and segmentation using a multiclass and reduced panoramic radiograph dataset. *J. Dent.* **2024**, *144*, 104891. [CrossRef] [PubMed]
7. Buyuk, C.; Arican Alpay, B.; Er, F. Detection of the separated root canal instrument on panoramic radiograph: A comparison of LSTM and CNN deep learning methods. *Dentomaxillofacial Radiol.* **2023**, *52*, 20220209. [CrossRef]
8. Xue, T.; Chen, L.; Sun, Q. Deep learning method to automatically diagnose periodontal bone loss and periodontitis stage in dental panoramic radiograph. *J. Dent.* **2024**, *150*, 105373. [CrossRef]
9. Anantharaman, R.; Velazquez, M.; Lee, Y. Utilizing mask R-CNN for detection and segmentation of oral diseases. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018.
10. Sivari, E.; Senirkentli, G.B.; Bostanci, E.; Guzel, M.S.; Acici, K.; Asuroglu, T. Deep learning in diagnosis of dental anomalies and diseases: A systematic review. *Diagnostics* **2023**, *13*, 2512. [CrossRef]
11. Yüksel, A.E.; Gültekin, S.; Simsar, E.; Özdemir, Ş.D.; Gündoğar, M.; Tokgöz, S.B.; Hamamcı, İ.E. Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning. *Sci. Rep.* **2021**, *11*, 12342. [CrossRef]
12. 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.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. *arXiv* **2015**, arXiv:1506.02640.
14. Available online: https://xaitk-saliency.readthedocs.io/en/latest/introduction.html#saliency-algorithms (accessed on 23 February 2025).
15. Petsiuk, V.; Das, A.; Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv* **2018**, arXiv:1806.07421.
16. Rašić, M.; Tropčić, M.; Karlović, P.; Gabrić, D.; Subašić, M.; Knežević, P. Detection and Segmentation of Radiolucent Lesions in the Lower Jaw on Panoramic Radiographs Using Deep Neural Networks. *Medicina* **2023**, *59*, 2138. [CrossRef]
17. Tekin, B.Y.; Ozcan, C.; Pekince, A.; Yasa, Y. An enhanced tooth segmentation and numbering according to FDI notation in bitewing radiographs. *Comput. Biol. Med.* **2022**, *146*, 105547.
18. McGuigan, M.; Louca, C.; Duncan, H. Clinical decision-making after endodontic instrument fracture. *Br. Dent. J.* **2013**, *214*, 395–400. [CrossRef]
19. Souyave, L.; Inglis, A.; Alcalay, M. Removal of fractured endodontic instruments using ultrasonics. *Br. Dent. J.* **1985**, *159*, 251–253. [CrossRef]
20. Souter, N.J.; Messer, H.H. Complications associated with fractured file removal using an ultrasonic technique. *J. Endod.* **2005**, *31*, 450–452. [CrossRef]
21. Hülsmann, M.; Schinkel, I. Influence of several factors on the success or failure of removal of fractured instruments from the root canal. *Dent. Traumatol.* **1999**, *15*, 252–258. [CrossRef]

22. Razaghi, M.; Komleh, H.E.; Dehghani, F.; Shahidi, Z. Innovative Diagnosis of Dental Diseases Using YOLO V8 Deep Learning Model. In Proceedings of the 2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP), Tehran, Iran, 6–7 March 2024.
23. Çelik, B.; Çelik, M.E. Root dilaceration using deep learning: A diagnostic approach. *Appl. Sci.* **2023**, *13*, 8260. [CrossRef]
24. Mureșanu, S.; Hedeșiu, M.; Iacob, L.; Eftimie, R.; Olariu, E.; Dinu, C.; Jacobs, R.; on behalf of Team Project Group. Automating Dental Condition Detection on Panoramic Radiographs: Challenges, Pitfalls, and Opportunities. *Diagnostics* **2024**, *14*, 2336. [CrossRef]
25. Yavuz, M.B.; Sali, N.; Bayrakdar, S.K.; Ekşi, C.; İmamoğlu, B.S.; Bayrakdar, İ.Ş.; Çelik, Ö.; Orhan, K. Classification of Periapical and Bitewing Radiographs as Periodontally Healthy or Diseased by Deep Learning Algorithms. *Cureus* **2024**, *16*, e60550. [CrossRef] [PubMed]
26. Wu, P.Y.; Mao, Y.C.; Lin, Y.J.; Li, X.H.; Ku, L.T.; Li, K.C.; Chen, C.A.; Chen, T.Y.; Chen, S.L.; Tu, W.C.; et al. Precision Medicine for Apical Lesions and Peri-Endo Combined Lesions Based on Transfer Learning Using Periapical Radiographs. *Bioengineering* **2024**, *11*, 877. [CrossRef] [PubMed]
27. Wang, K.X.; Zhang, S.B.; Wei, Z.Y.; Fang, X.L.; Liu, F.; Han, M.; Du, M. Deep learning-based efficient diagnosis of periapical diseases with dental X-rays. *Image Vis. Comput.* **2024**, *147*, 105061. [CrossRef]
28. Guo, Y.; Guo, J.; Li, Y.; Zhang, P.; Zhao, Y.-D.; Qiao, Y.; Liu, B.; Wang, G. Rapid detection of non-normal teeth on dental X-ray images using improved Mask R-CNN with attention mechanism. *Int. J. Comput. Assist. Radiol. Surg.* **2024**, *19*, 779–790. [CrossRef]
29. Jayasinghe, H.; Pallepitiya, N.; Chandrasiri, A.; Heenkenda, C.; Vidhanaarachchi, S.; Kugathasan, A.; Rathnayaka, K.; Wijekoon, J. Effectiveness of Using Radiology Images and Mask R-CNN for Stomatology. In Proceedings of the 2022 4th International Conference on Advancements in Computing (ICAC), Colombo, Sri Lanka, 9–10 December 2022.
30. Widyaningrum, R.; Candradewi, I.; Aji, N.R.A.S.; Aulianisa, R. Comparison of Multi-Label U-Net and Mask R-CNN for panoramic radiograph segmentation to detect periodontitis. *Imaging Sci. Dent.* **2022**, *52*, 383–391. [CrossRef]
31. Abbas, N.M.; Solomon, D.G.; Bahari, M.F. A review on current research trends in electrical discharge machining (EDM). *Int. J. Mach. Tools Manuf.* **2007**, *47*, 1214–1228. [CrossRef]
32. Roganović, J.; Radenković, M.; Miličić, B. Responsible use of artificial intelligence in dentistry: Survey on dentists' and final-year undergraduates' perspectives. *Healthcare* **2023**, *11*, 1480. [CrossRef]
33. Gianfrancesco, M.A.; Tamang, S.; Yazdany, J.; Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **2018**, *178*, 1544–1547. [CrossRef]
34. Mörch, C.; Atsu, S.; Cai, W.; Li, X.; Madathil, S.; Liu, X.; Mai, V.; Tamimi, F.; Dilhac, M.; Ducret, M. Artificial intelligence and ethics in dentistry: A scoping review. *J. Dent. Res.* **2021**, *100*, 1452–1460. [CrossRef]

diagnostics

MDPI

Article

Mixture of Expert-Based SoftMax-Weighted Box Fusion for Robust Lesion Detection in Ultrasound Imaging

Se-Yeol Rhyou, Minyung Yu and Jae-Chern Yoo *

Department of Electrical and Computer Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 440-746, Republic of Korea; fbtpduf@naver.com (S.-Y.R.); 6unhuiw@gmail.com (M.Y.)

* Correspondence: yoojc@skku.edu; Tel.: +82-31-299-4591; Fax: +82-31-290-7948

Abstract: Background/Objectives: Ultrasound (US) imaging plays a crucial role in the early detection and treatment of hepatocellular carcinoma (HCC). However, challenges such as speckle noise, low contrast, and diverse lesion morphology hinder its diagnostic accuracy. **Methods:** To address these issues, we propose CSM-FusionNet, a novel framework that integrates clustering, SoftMax-weighted Box Fusion (SM-WBF), and padding. Using raw US images from a leading hospital, Samsung Medical Center (SMC), we applied intensity adjustment, adaptive histogram equalization, low-pass, and high-pass filters to reduce noise and enhance resolution. Data augmentation generated ten images per one raw US image, allowing the training of 10 YOLOv8 networks. The mAP@0.5 of each network was used as SoftMax-derived weights in SM-WBF. Threshold-lowered bounding boxes were clustered using Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and outliers were managed within clusters. SM-WBF reduced redundant boxes, and padding enriched features, improving classification accuracy. **Results:** The accuracy improved from 82.48% to 97.58% with sensitivity reaching 100%. The framework increased lesion detection accuracy from 56.11% to 95.56% after clustering and SM-WBF. **Conclusions:** CSM-FusionNet demonstrates the potential to significantly improve diagnostic reliability in US-based lesion detection, aiding precise clinical decision-making.

Keywords: hepatocellular carcinoma; ultrasound; SoftMax-weighted box fusion; clustering; CNN; deep learning

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 24 January 2025
Revised: 25 February 2025
Accepted: 26 February 2025
Published: 28 February 2025

Citation: Rhyou, S.-Y.; Yu, M.; Yoo, J.-C. Mixture of Expert-Based SoftMax-Weighted Box Fusion for Robust Lesion Detection in Ultrasound Imaging. *Diagnostics* **2025**, *15*, 588. https://doi.org/10.3390/diagnostics15050588

1. Introduction

Liver cancer is a major cause of cancer-related mortality [1–3], ranking as the third leading cause of cancer deaths worldwide in 2022, with 865,000 new cases and 757,948 deaths reported [4]. Particularly, primary liver cancer is predominantly composed of HCC [5], which accounts for 75–85% of cases, and intrahepatic cholangiocarcinoma, representing 10–15%. Globally, chronic infection with HBV or HCV is associated with 21% to 55% of HCC cases [6,7]. The high mortality rate of liver cancer can be attributed to its late-stage detection, which significantly limits therapeutic options and reduces the likelihood of favorable outcomes [8,9]. Luckily, early identification of HCC allows for surgical resection, providing a favorable prognosis, with 5-year survival rates exceeding 70% [10,11]. Therefore, regular HCC screening is essential for individuals at risk to ensure early detection of HCC [12–14].

Among screening methods for HCC, ultrasonography is a non-invasive, cost-effective, and widely accessible imaging technique that provides real-time liver visualization without ionizing radiation [15]. These attributes make it ideal for regular screening and surveillance

in high-risk populations, particularly in low- and middle-income countries where liver cancer incidence is disproportionately high [16,17]. Despite its advantages, the efficacy of US depends on the operator's expertise and the quality of the equipment [18–21]. Additionally, the detection of small lesions (<2 cm) remains challenging, necessitating complementary methods such as serum biomarkers or advanced imaging techniques for confirmatory diagnosis [22].

Nevertheless, combining ultrasonography with other non-invasive approaches holds significant potential for improving the early detection of liver cancer. Advanced imaging techniques are currently being developed to improve the detectability of HCC and enhance the characterization of HCC nodules [23–25]. Recently, emerging artificial intelligence technologies, featuring powerful brain-like algorithms in the field of medical imaging, have significantly contributed to improving the accuracy of diagnoses [13–19]. Numerous artificial intelligence approaches based on deep learning have been proposed to address the challenges [26–32] and issues related to ultrasonography diagnostic accuracy of HCC as well as the expertise required [33–37].

Ryu et al. (2021) [33] proposed a joint segmentation and classification system for hepatic lesions in ultrasound images, utilizing a shared encoder with two branches. The input combined a grayscale image with Euclidean distance maps of foreground and background clicks provided by the user. The system achieved 89.8% accuracy for classification and 90.4% accuracy for joint segmentation and classification. The segmentation network employed bilinear interpolation for upsampling, reducing parameters and mitigating overfitting compared to FCNs. The classification branch, based on a VGG-16 architecture, predicted lesion types from shared convolutional features. While the integration of segmentation and classification demonstrated potential, the performance exhibited limitations, particularly in fully leveraging joint learning. The approach highlights the challenges of combining these tasks effectively in medical imaging.

Zhao et al. (2023) [34] developed a lightweight neural network, USC-Enet, designed for small-scale medical image datasets, incorporating an attention mechanism to address overfitting issues common with small datasets. Using 2168 images, the model achieved a sensitivity of 0.915. To improve performance, the study streamlined the network structure for transfer learning, reduced network parameters to prevent overfitting, and combined image features with clinical data using a random forest classifier. This end-to-end model integrated Convolutional Neural Network (CNN) feature maps with clinical data, such as age, gender, hepatitis history, and tumor markers like alpha-fetoprotein, but sensitivity score remains suboptimal.

Poreddy et al. (2024) [35] proposed a classification model for focal liver lesions by applying the discrete Haar wavelet transform to each frame of a focal liver lesions video, decomposing each frame into multiple subbands. Singular value decomposition was performed on each subband, and the maximum value among the columns of the Singular Value Decomposition matrices was extracted as a frame-level statistical feature. These features were averaged across all video frames and fed into a decision tree classifier. Experimental results achieved an accuracy of 97.18%, surpassing conventional methods. The authors chose the decision tree due to its robustness on small datasets and ability to capture simple relationships without overfitting, leading to superior performance compared to other classifiers.

Chaiteerakij et al. (2024) [36], in a retrospective study using 26,288 ultrasound images from 5444 patients, developed and evaluated an AI-assisted system for detecting and classifying seven types of focal liver lesions, including HCC. They employed YOLOv5, which predicts multiple bounding boxes per grid cell to handle objects of varying sizes and aspect ratios. To remove duplicates, non-maximum suppression was applied, retaining

only the bounding box with the highest confidence score in each group. By integrating YOLOv5 with the Darknet architecture and Cross-Stage Hierarchical Networks (CSPNet), and pretraining on the COCO dataset, the system achieved an 84% detection rate for HCC (among all lesions) and an overall accuracy of 94%. Despite its high performance, the model's computational complexity may limit its adoption in healthcare settings where computing resources are constrained.

In this study, we present CSM-FusionNet, a novel deep learning-based fusion model designed to integrate Clustering, SoftMax-weighted Box Fusion, and Mixture of Experts. The proposed framework is structured around four core functionalities, each contributing distinct advantages to the overall system. A concise description of these components is provided below to highlight their specific roles and significance within the network.

1. Image Preprocessing: Due to the inherent characteristics of US images, significant speckle noise and variations in resolution arise depending on device settings. To address these issues, four distinct image processing filters were applied to reduce noise and accommodate diverse resolutions. This approach also facilitated data augmentation, expanding a single US image into ten variations for enhanced training and analysis.
2. Lesion Detection with YOLOv8: To detect lesions within the liver region, we utilized YOLOv8, one of the most widely adopted object detection models. By leveraging the ten augmented US images, we constructed ten individual object detection networks. The mAP@0.5 of each network was measured to serve as weights for the subsequent bounding box fusion process.
3. Bounding Box Optimization: This stage represents the most technically complex and critical component of the framework. A low detection threshold is applied across the ten networks to generate multiple bounding boxes for regions suspected of containing lesions. Clustering is then performed on the resulting bounding boxes to identify distinct lesion-distributed regions. Subsequently, these multiple boxes are consolidated into a single representative box for each region using SM-WBF, an adaptation of the Weighted Box Fusion (WBF) method tailored to our approach. To mitigate potential information loss during the fusion process, padding is applied to each final bounding box as a concluding step.
4. Lesion Classification: The bounding boxes generated through the optimization process are ultimately classified into three categories: benign, malignant, and error. Each US image may yield one or multiple lesion boxes, all of which are subjected to classification. Among the classified results, boxes labeled as "error" are disregarded. In cases where both benign and malignant classifications are present within the same image, the lesion is conservatively categorized as malignant to prioritize diagnostic sensitivity.

Our technology holds the potential to achieve high accuracy and efficiency in the detection and analysis of lesions using ultrasound imaging. By employing advanced fusion techniques, such as SM-WBF and clustering algorithms, the system integrates information from multiple networks to deliver reliable diagnostic results. This enables healthcare professionals to detect even subtle lesions, reduces diagnostic time, and supports optimal decision-making for personalized treatment. Furthermore, the automated nature of the technology enhances resource efficiency and enables precise diagnostics even in regions with limited medical infrastructure, thereby improving healthcare accessibility. Ultimately, this innovation emphasizes the importance of early diagnosis and preventive care, contributing to increased patient survival rates and improved quality of life.

2. Materials and Methods

2.1. Dataset Preparation

The liver US images used in this study were acquired using the Siemens ACUSON Sequoia 512 system (Siemens Healthineers, Erlangen, Germany), which operates at a frequency range of 3 to6 MHz, with 256 gray levels and a maximum depth of 36 cm. These images were collected under data use agreements and approved by the Institutional Review Board (IRB) of SMC, one of South Korea's leading hospitals (SMC-2020-10-178-002). The dataset was annotated with two categories: benign and malignant. Since accurate diagnosis using only US images is challenging, the final labels were determined by SMC's expert physicians after completing all necessary diagnostic procedures, including CT, MRI, and biopsy. The annotation for each patient's US images was based on these comprehensive findings. Given the goal of our study to develop a screening tool, lesions were categorized simply into benign or malignant, rather than employing more granular labeling. Figure 1 provides examples of benign and malignant cases from our dataset.

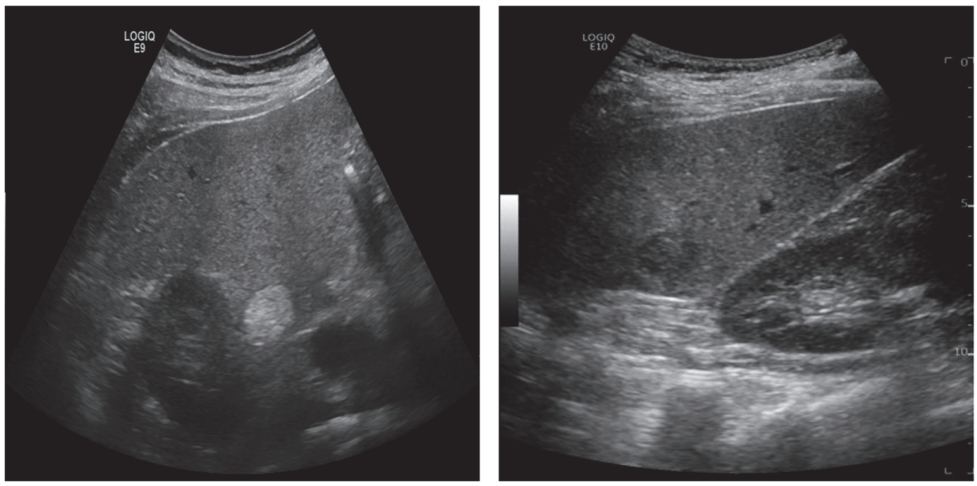


Figure 1. Example of US images from dataset. (**a**) Benign, (**b**) Malignant.

The dataset consisted of a total of 1054 ultrasound images, collected from patients aged 48 to 82 years. Among these, 746 images were labeled as benign, and 308 images were labeled as malignant. For model development and evaluation, the dataset was divided into training, validation, and test sets in a 6:2:2 ratio.

2.2. Proposed Deep Learning Neural Network: CSM-FusionNet

As illustrated in Figure 2, this flowchart outlines the framework of our proposed network, CSM-FusionNet. The input images are US images collected from SMC. These images undergo preprocessing to address the dynamic options and characteristics of US equipment. To this end, we applied a combination of filters, including a low-pass Gaussian filter, a high-pass sharpening spatial filter, an intensity adjustment filter, and an adaptive histogram equalization filter. By combining these filters, each original image was augmented into ten variations, which were then processed through ten distinct object detection models to identify potential lesions. The resulting detections produced numerous bounding boxes, requiring an ensemble approach to identify the true lesion locations. To achieve this, we utilized DBSCAN, a clustering algorithm, and a gate network inspired by Mixture of Experts, to perform SM-WBF. This process refined the bounding boxes to isolate regions most likely to contain actual lesions. During clustering, bounding boxes significantly larger or smaller than the average size of their respective clusters were filtered out as outliers. The detected

lesions were then classified using a CNN, categorizing each into benign, malignant, or error classes. The error class was specifically introduced to exclude incorrectly detected lesions. This comprehensive approach allows the system to effectively handle the inherent diversity of US images while maintaining high diagnostic accuracy.

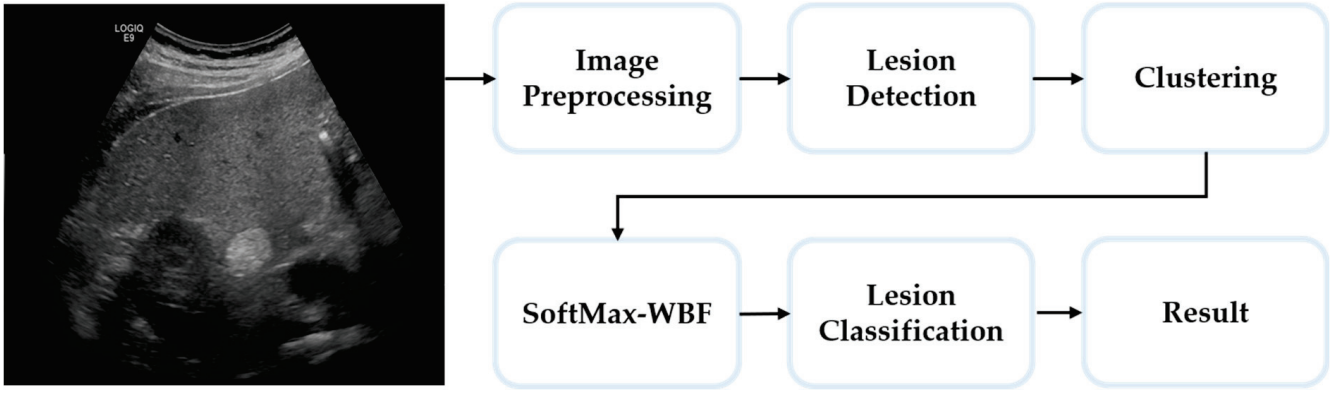


Figure 2. Flowchart of our proposed network: CSM-FusionNet.

Section 2.3 describes the method for generating ten augmented images from a single ultrasound image using a combination of preprocessing filters. In Section 2.4, the process of lesion detection is outlined, where each of the ten augmented images is analyzed using individual object detection models, and the mAP@0.5 values for each network are calculated. Section 2.5 explains the normalization of weights based on the mAP@0.5 values and introduces the clustering-based SM-WBF method for optimizing bounding boxes. Finally, Section 2.6 details the lesion classification process, where the refined bounding boxes are categorized into benign, malignant, or error classes.

2.3. Dataset Preprocessing

The preprocessing of US images is the first step in our framework. Due to the inherent characteristics of ultrasound equipment, US images often contain noise and exhibit varying resolutions depending on dynamic settings. To generalize these differences, we applied multiple filters to augment the images. A total of four filters were used, with alpha and beta values incorporated for preprocessing. First, a contrast enhancement process was applied using the intensity adjustment filter and adaptive histogram equalization filter, generating two images from the original US image. Additionally, low-pass and high-pass filters were applied, with variations introduced through alpha and beta values, resulting in two distinct images for each filter type. By combining these four filters, a total of ten augmented images were generated. Further details of this process are illustrated in Figure 3.

As shown in Figure 3, the images processed with the high-pass filter appeared significantly sharper, while those processed with the low-pass filter exhibited a more blurred effect compared to the original image. This approach allowed us to effectively handle variations in resolution and depth differences inherent in ultrasound imaging. The alpha and beta values used for preprocessing were set empirically, with 1.5 and 0.5 selected during our experiments.

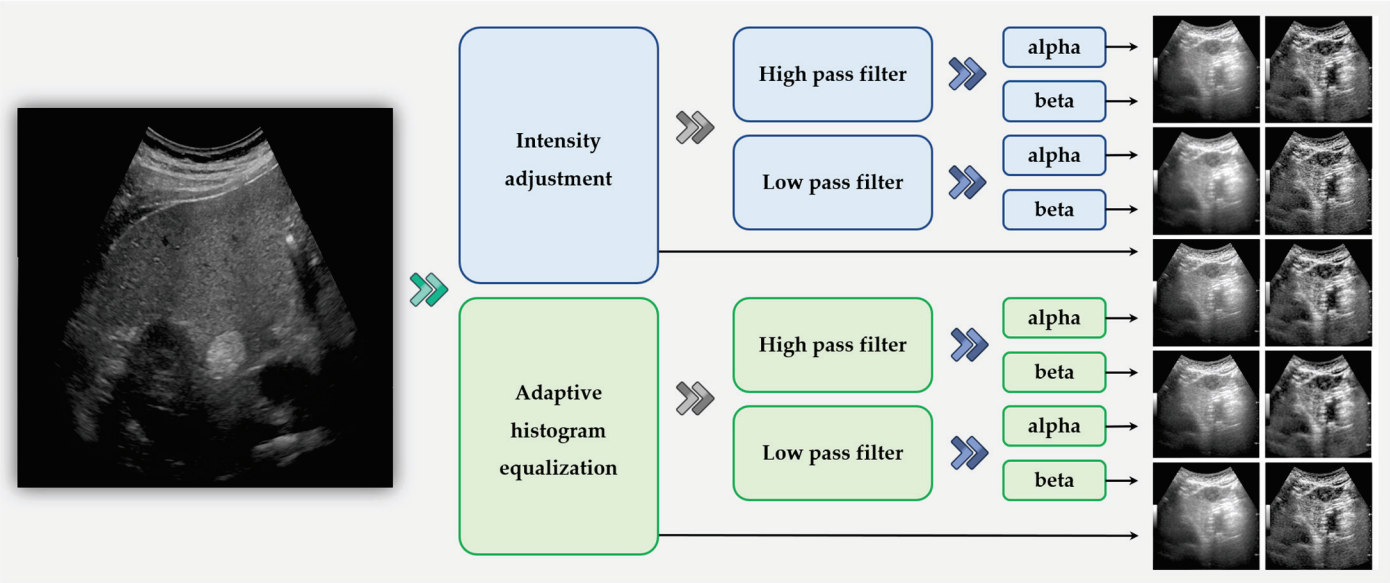


Figure 3. The original US image serves as the input image, and through the application of four distinct filters combined with alpha and beta values, a total of ten augmented images are generated.

2.4. Lesion Detect with YOLOv8

Since ultrasound imaging is frequently used for real-time diagnostics, fast lesion detection is crucial. YOLOv8 offers a more lightweight architecture and optimized computation compared to YOLOv5, enabling faster inference speed. For this reason, we chose YOLOv8 for our approach. As illustrated in Figure 4, we employed YOLOv8 to develop the lesion detector and trained a total of ten networks using ten different datasets. A low threshold was applied during the detection process to generate a larger number of bounding boxes. While a higher threshold might improve accuracy metrics, it often results in missing smaller lesions, particularly those in early stages, or lesions that deviate significantly from the average size. Such omissions would render the system less effective in clinical applications, despite high accuracy metrics. By applying a lower threshold, each network produced between zero and five bounding boxes, which were then utilized in the subsequent steps.

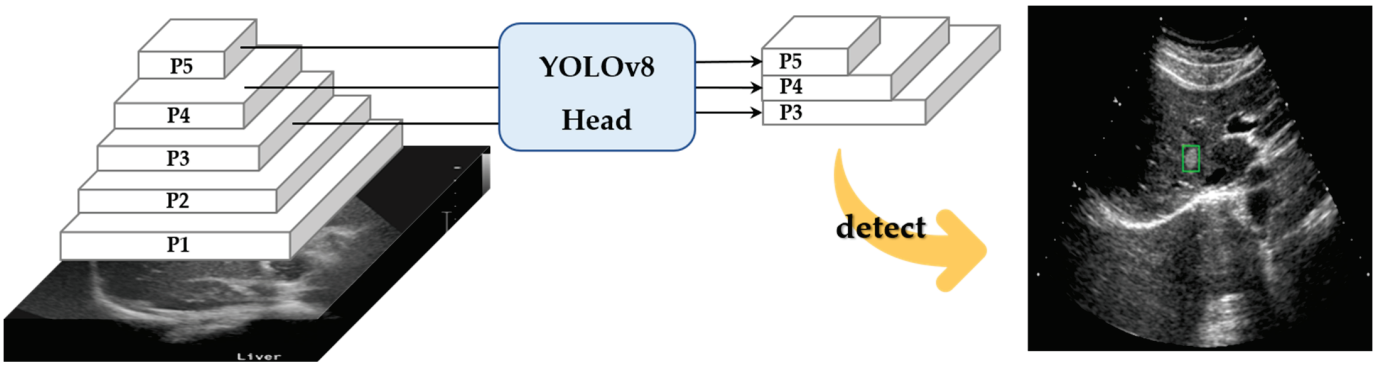


Figure 4. Lesion suspected regions, green bounding box, detected by YOLOv8.

2.5. Bounding Box Optimization by Clustering, SM-WBF and Padding

In Section 2.4, numerous bounding boxes are generated by ten YOLOv8 networks. These bounding boxes may focus on a single location, concentrate on multiple areas, or, in some cases, fail to detect any region. To address all these scenarios, clustering was employed to identify regions where bounding boxes are concentrated, followed by an ensemble process to consolidate multiple boxes into a single representative box for each region. The detailed methodology is outlined as follows.

2.5.1. Clustering Using DBSCAN

When utilizing multiple networks for bounding box detection, issues arise where bounding boxes for the same object may be generated in slightly different locations or where multiple overlapping boxes are created. In this study, ten networks were employed, and a low threshold was set to enhance detection sensitivity. This approach resulted in a significant number of overlapping bounding boxes. To address this, DBSCAN was applied for clustering. DBSCAN is a density-based clustering algorithm that groups closely located bounding boxes into clusters based on their center coordinates (x_{center}, y_{center}). Bounding boxes that are sufficiently close are assigned to the same cluster, while in traditional DBSCAN, boxes in sparsely populated areas would be considered noise. However, in this study, such boxes are treated as important data and reassigned to new clusters. For clustering, the distance criterion (*eps*) was set to 0.1, and the minimum number of samples was set to 1 to ensure that at least one bounding box could form a cluster.

As a result, each cluster contained between one and five bounding boxes. Unlike traditional DBSCAN, all bounding boxes, including those that might be classified as noise, were retained as clusters for further analysis. The clustered results were then passed to the next stage of the process.

2.5.2. Calculate SoftMax Weights Based on the mAP@0.5 Score

To integrate the bounding box detection results, this study employed a Mixture of Experts (MoE) approach to calculate weights that reflect the performance of each network. MoE combines the outputs of multiple expert models to produce optimal results, with a mechanism that dynamically adjusts the contribution of each expert. The performance of each expert is evaluated based on its reliability and relevance to the problem, maximizing the quality of the outcome.

In this study, the metric used to quantify the importance of each expert was the mAP@0.5 (Mean Average Precision at Intersection of Union (IoU) threshold 0.5) of each network. This metric provides a numerical representation of detection performance, where a higher mAP value indicates more reliable results. However, directly using mAP values does not effectively capture the non-linear differences in performance across networks. To address this, SoftMax normalization was applied to dynamically compute the relative importance of each network.

SoftMax normalization is closely aligned with the expert selection mechanism in MoE. By reflecting performance differences among networks, it ensures that higher-performing networks contribute more significantly to the final detection results. This approach effectively implements the principles of MoE, leveraging SoftMax to prioritize networks that generate more accurate and reliable predictions.

i. Performance-based weight assignment reflects the relative performance differences among networks, ensuring that networks with higher mAP values are assigned greater importance.
ii. Expressing normalized contribution normalizes the weights of all networks, clearly illustrating the relative contributions of each network to the outcome.
iii. Non-linear influence enhancement amplifies the contribution of high-performing networks non-linearly as the performance gap increases, ensuring a stronger influence from superior networks.

The SoftMax weights calculated in this manner are applied to the bounding box detection integration process based on the expert selection principle of the MoE. SoftMax normalization is defined by the following equation:

$$SoftMax\_w_i = \frac{e^{mAP_i}}{\sum_{j=1}^{N} e^{mAP_j}} \quad (1)$$

where N is the total number of YOLO networks.

2.5.3. Filter out Outlier Bounding Boxes

We employed ten networks and lowered the detection threshold to identify all regions suspected of being lesions. As a result, the bounding box detection outputs may include outliers that are either excessively large or small. Such outliers significantly deviate from the average size of their respective clusters and can reduce the reliability of the results. To address this, the bounding boxes within each cluster were refined based on their dimensions, removing those identified as outliers before proceeding to the integration process. Outlier removal was performed using the width and height of the bounding boxes within each cluster. Bounding boxes that deviated beyond a predefined outlier threshold from the average size were excluded. This process minimized the impact of extreme values on the subsequent SoftMax-weighted box fusion, thereby improving the quality of the final bounding boxes. Specifically, the average width and height of the bounding boxes in each cluster were defined as $\overline{w}$ and $\overline{h}$, respectively. A bounding box with width w and height h was considered valid if it satisfied the following conditions:

$$(1 - \text{threshold}) \times \overline{w} \leq w < (1 + \text{threshold}) \times \overline{w},\ (1 - \text{threshold}) \times \overline{h} \leq h < (1 + \text{threshold}) \times \overline{h} \quad (2)$$

where the threshold represents the allowable deviation ratio, set to 0.5 in this study.

Through this process, outliers were removed from all bounding boxes within each cluster, and the refined data were used in the subsequent integration stage. This outlier removal step is essential to ensure the reliability and accuracy of the results by preventing size distortions in bounding boxes and producing more consistent outcomes.

2.5.4. Bounding Box Fusion with SM-WBF

After clustering the bounding box detection results and removing outlier boxes, SM-WBF is applied to generate a single representative box for each cluster. This method focuses on merging the bounding boxes within each cluster identified through DBSCAN in Section 2.5.1, removing redundancies, and producing an optimal result. SM-WBF specifically incorporates the SoftMax weights calculated in the earlier stage to implement a fusion strategy that reflects the performance of each network.

SM-WBF computes the weighted average of the center coordinates (x_{center}, y_{center}) and the dimensions (width w, height h) of each bounding box, using the SoftMax weights. This ensures that the detection results from networks with higher mAP values have a greater influence on the final box. By incorporating SoftMax normalization, the relative importance of all networks is considered, closely aligning with the expert selection mechanism of MoE. Unlike traditional WBF methods, which rely on confidence scores, our approach treats all detected bounding boxes as equally valid, excluding confidence scores from the fusion

process. During the box integration process, the coordinates and dimensions of the boxes within each cluster are computed using the following equations:

$$\begin{aligned} \hat{x}_{center} &= \frac{\sum_{i=1}^{N} SoftMax\_w_i \times x_{center,i}}{\sum_{i=1}^{N} SoftMax\_w_i}, \hat{y}_{center} = \frac{\sum_{i=1}^{N} SoftMax\_w_i \times y_{center,i}}{\sum_{i=1}^{N} SoftMax\_w_i} \\ \hat{width} &= \frac{\sum_{i=1}^{N} SoftMax\_w_i \times width_i}{\sum_{i=1}^{N} SoftMax\_w_i}, \hat{height} = \frac{\sum_{i=1}^{N} SoftMax\_w_i \times height_i}{\sum_{i=1}^{N} SoftMax\_w_i} \end{aligned} \quad (3)$$

where $\hat{x}_{center}$ and $\hat{y}_{center}$ represent the center coordinates of the integrated representative bounding box, and $\hat{width}$ and $\hat{height}$ denote the width and height of the integrated box. N is the number of bounding boxes within the cluster. $SoftMax\_w_i$ is the SoftMax weight calculated in the previous step, and x_{center}, y_{center}, $width_i$, $height_i$ are the coordinates and dimensions of the *i*-th bounding box within the cluster.

The key innovation of the proposed SM-WBF method lies in the application of mAP-based SoftMax weights, ensuring that the results from higher-performing networks contribute more significantly to the final bounding box computation. This approach minimizes the influence of incorrect detections from lower-performing networks while optimizing the result by comprehensively integrating the information from all bounding boxes within a cluster. By incorporating the principles of MoE, the method bases its final decisions on the relative reliability of each expert model (network), thereby enhancing the overall accuracy of the results. The representative bounding box generated through this process is subsequently used in the next stages, including outlier removal and padding application, to further refine the detection outcome.

2.5.5. Add Padding to Bounding Boxes

In ultrasound imaging, the contrast between the tumor contour and the surrounding liver tissue provides critical clinical information. Therefore, preserving the detected tumor boundary is essential for accurate analysis and diagnosis. However, during the process of generating bounding boxes through clustering, there is a risk of partially cropping the tumor boundary, which can undermine the reliability of the detection results. To prevent this and retain additional information, such as color contrast around the tumor, padding was applied to the bounding boxes.

Padding involves extending the boundaries of the bounding box to include the detected object and its surrounding area. The padding size was defined in pixel units, and the YOLO-format bounding boxes were converted into pixel coordinates to facilitate the process. After extending each boundary by a specified number of pixels, the padded bounding box was converted back to the YOLO format to maintain compatibility with the detection system. Padding was applied using the following formula:

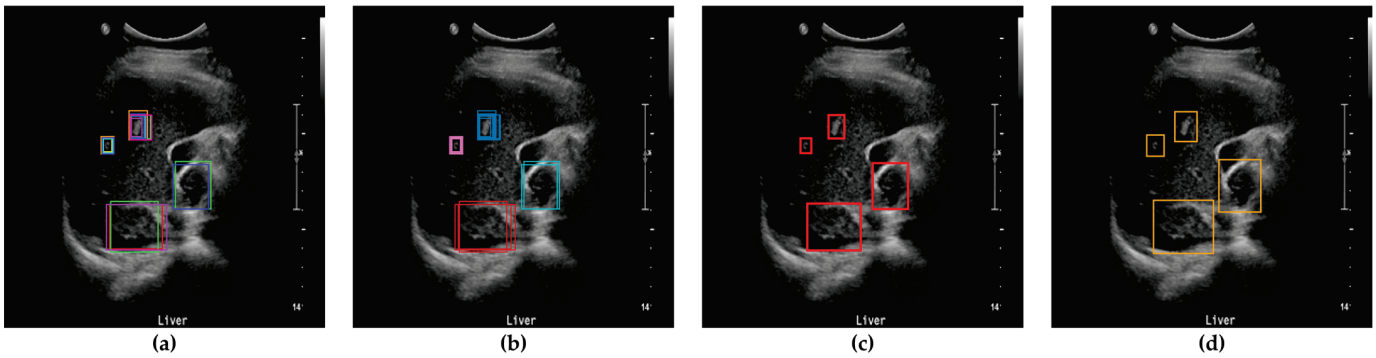
$$x_{left} = x_{min} - p,\ x_{right} = x_{max} + p,\ y_{left} = y_{min} - p,\ y_{left} = y_{max} + p \quad (4)$$

where x_{min}, x_{max}, y_{min}, y_{max} represent the original bounding box coordinates, and p denotes the number of pixels added to each boundary.

The padding size was optimized based on the characteristics of ultrasound imaging and the requirements for tumor analysis. By applying padding, the bounding box encompassed not only the detected object's boundaries but also its surrounding tissue. This approach preserved essential information for analyzing the contrast between the tumor boundary and the liver tissue, preventing boundary loss during the clustering process. As a result, the application of padding significantly improved the reliability of bounding box-based detection.

To evaluate the accuracy of lesion detection after applying clustering, SM-WBF, and padding, a novel measurement method was introduced. The primary goal was to ensure that all regions suspected of being lesions were identified, minimizing the chances of

missing any lesions. As illustrated in Figure 5, all detected lesion-suspected regions were compared with the ground truth, and if at least one box matched, the detection was considered successful. A match was determined by comparing the IoU between the detected box and the ground truth box including padding. An IoU of 0.9 or higher was used as the threshold for a match, ensuring high precision in detection evaluation (Figure 6).



(a) (b) (c) (d)

Figure 5. Determination of bounding boxes using clustering, SM-WBF, and padding. (**a**) All bounding boxes detected by the ten networks. (**b**) Clustering of bounding boxes into four regions using DBSCAN. (**c**) Application of SM-WBF with SoftMax weights to the clustered regions. (**d**) Addition of padding to the bounding boxes in (**c**).

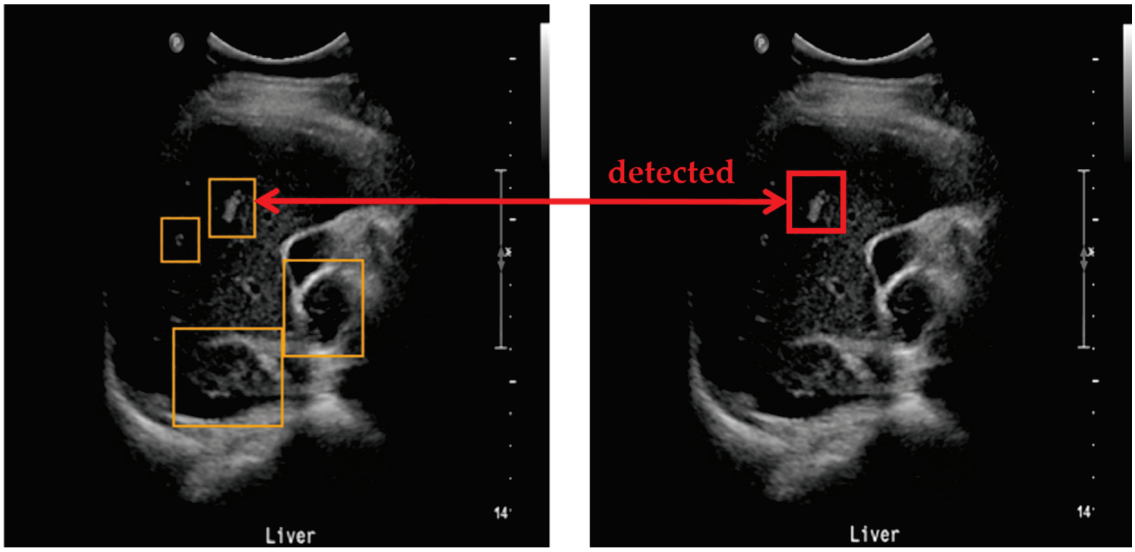


Figure 6. Method for measuring lesion detection accuracy. Among the four detected bounding boxes, at least one box has an IoU of 0.9 or higher with the ground truth, and, therefore, the detection is considered successful.

2.6. Lesion Classification

The final step is lesion classification. Up to this point, multiple processes such as clustering and SM-WBF were employed to accurately locate and classify the lesions. The lesion classification model provides predictions for the class of each detected lesion through YOLO, and this information supports clinical decision-making. The lesions were categorized into three classes: benign, malignant, and error.

The inclusion of the error class was essential due to the lowered detection threshold used during the YOLO process, which aimed to detect all potentially suspicious lesions. After the clustering and SM-WBF processes reduced the bounding boxes to a single box per region, the classification output for that box became the final result. In cases where more than one bounding box remained, the following criteria were applied: if all bounding boxes,

excluding the error class, were classified as benign, the final result was benign; however, if even one bounding box was classified as malignant, the result was determined to be malignant. This decision-making process ensured that any potentially malignant lesions were not overlooked, thereby increasing sensitivity and enhancing the clinical value of the proposed approach.

3. Result

Our proposed CSM-FusionNet was implemented using MATLAB R2023b on a computer equipped with a GeForce RTX 3090 GPU with 24 GB of memory. Liver ultrasound images, annotated by experts, were collected from Samsung Medical Center to evaluate the performance of the proposed model. As demonstrated in the following experimental results, CSM-FusionNet effectively reduced false detections caused by variations in lesion size and other noise factors. Moreover, it addressed the limitations of traditional confidence-based WBF by introducing the novel SM-WBF method. Detailed quantitative results are provided below.

3.1. Result of Bounding Box Optimization by Clustering, SM-WBF and Padding

In this study, clustering, SM-WBF, and padding were applied to optimize bounding boxes. This optimization process aimed to remove redundancy among detected bounding boxes, refine outliers, and preserve the contour information of lesions, thereby enhancing the reliability and accuracy of the final detection results. Using DBSCAN, bounding boxes detecting the same lesion were grouped into clusters. Subsequently, bounding boxes within each cluster were refined by identifying outliers based on their dimensions $\hat{width}$ and $\hat{height}$. Boxes deviating more than 50% from the average size of the cluster were considered outliers and reduced. This step was essential to prevent distortion in cluster size and ensure stability during the integration process.

After clustering and outlier reduction, SM-WBF was applied to generate a single representative bounding box for each cluster. SM-WBF calculated weights for each bounding box by normalizing the mAP@0.5 values of the ten networks using the SoftMax function. These SoftMax weights were then used to combine the center coordinates and dimensions of the bounding boxes within each cluster through a weighted average approach. This method ensured that bounding boxes from higher-performing networks contributed more significantly to the result. Details of this process are illustrated in Figure 7.

3.2. Performance of Lesion Detection

To evaluate the accuracy of the bounding boxes generated through clustering, SM-WBF, and padding, we calculated the detection success rate on the test data. A detection was considered successful if at least one detected bounding box matched the ground truth. The matching criteria were based on the IoU between the detected box and the ground truth box with the same padding applied. If the IoU was 0.9 or higher, the boxes were deemed to match. Detailed results are presented in the table below.

Table 1 compares the results across four different scenarios, presenting mAP@0.5 scores, the number of correctly detected lesions in the test data, and the resulting accuracy. When a single YOLOv8 model was applied to the original US images without any preprocessing, the mAP@0.5 was 0.6128, and the model detected 101 lesions out of 180 test cases, achieving an accuracy of 56.11%. After applying our preprocessing steps to generate ten networks, clustering the bounding boxes, and performing SM-WBF, the method detected 172 out of 180 test cases, with an accuracy of 95.56%. In our evaluation using 180 test images, we compared the detection performance of YOLOv5, YOLOv8, and YOLOv11. YOLOv5 detected 167 lesions, YOLOv8 detected 172 lesions, and YOLOv11 detected 171 lesions,

demonstrating that YOLOv8 achieved the best detection performance. It is important to note that YOLO has many versions, and the most suitable version varies depending on the dataset. The accuracy values presented here are not generalizable but are specific to our dataset.

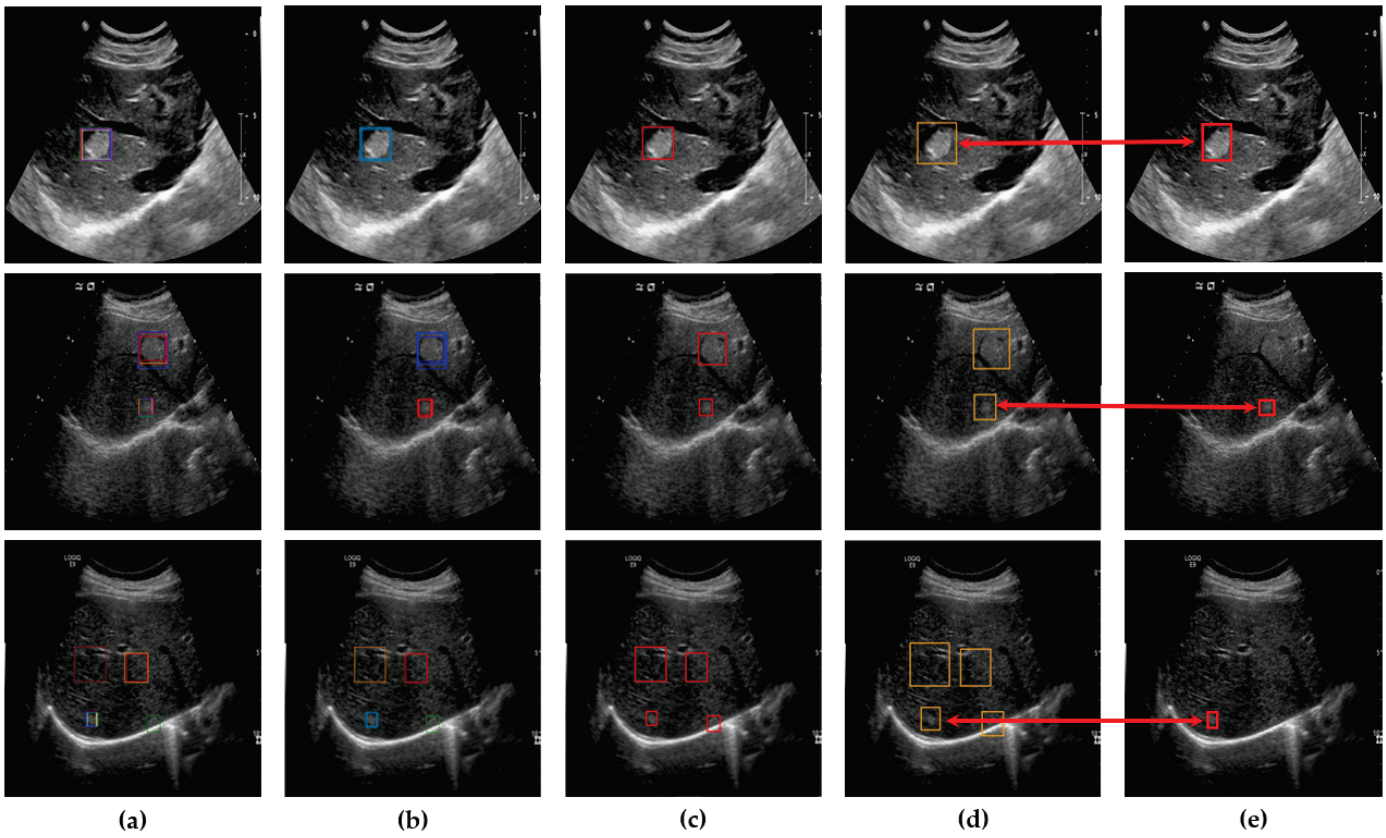


Figure 7. (**a**) An image with all bounding boxes detected by YOLOv8 overlaid, and the results of bounding box optimization through (**b**) clustering, (**c**) SM-WBF, and (**d**) padding. (**e**) Ground truth bounding box.

Table 1. The mAP@0.5 scores for each network and the detection results on the test data.

Network		mAP@0.5	Detect	Accuracy
Original US image	with one YOLOv8	0.6128	101/180	0.5611
WBF	with 10 Network	0.7453	112/180	0.6222
WBF + clustering		0.5038	163/180	0.9056
SM-WBF + clustering		0.5172	172/180	0.9556

Although the mAP@0.5 score was relatively low, this was because the clustering process preserved all potential bounding box regions, ensuring comprehensive lesion detection. However, the primary focus of our method is not achieving a higher mAP@0.5 but maximizing the capability to detect actual lesions, which demonstrates its performance was outstanding.

The bounding box optimization process contributed to enhancing the reliability and quality of the detection results by eliminating redundancies through clustering, refining outliers, and applying padding. The final bounding boxes encompassed the tumor contours without loss, providing a stable input for lesion classification models. This significantly improved the accuracy of the diagnostic support system.

3.3. Performance of Lesion Classification

The lesions identified through SM-WBF were classified using a CNN. Among various models, we selected EfficientNet-b0, known for its efficiency and performance, for comparison. As shown in Figure 8, the lesions were classified into three classes: benign, malignant, and error. For training, the image size was resized to 300 × 300, and the input size of EfficientNet-b0 was modified from its default 224 × 224 to 300 × 300. This adjustment was made to ensure accurate classification, as reducing the image size excessively could compromise the model's performance.

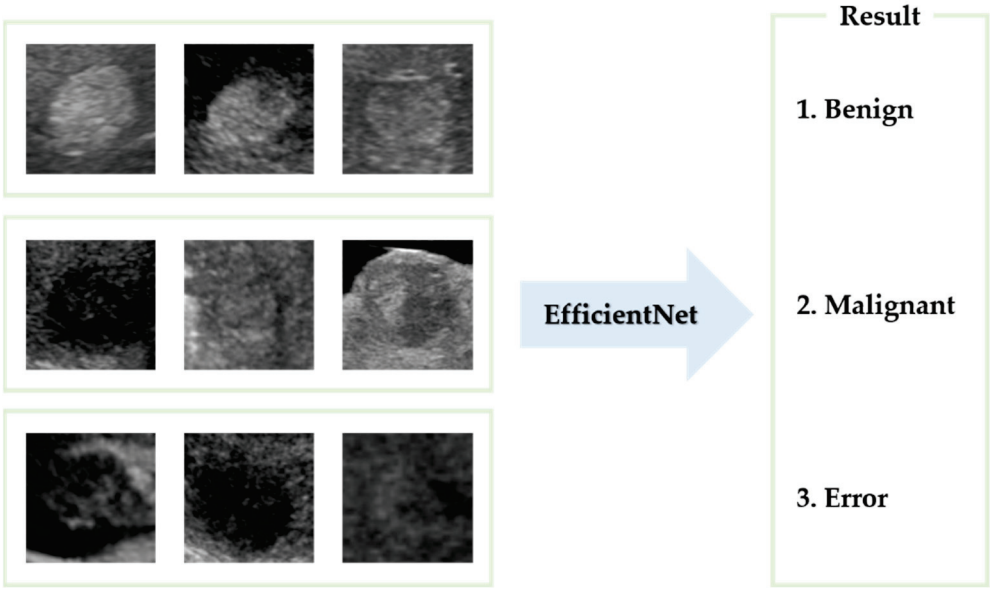


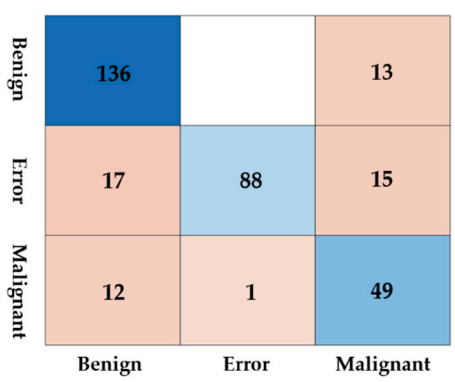
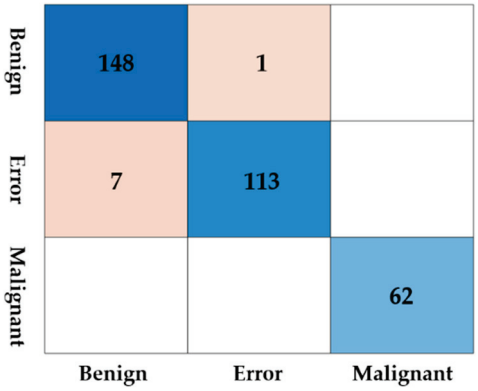
Figure 8. The lesions were classified into three classes using EfficientNet-b0.

The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a momentum of 0.9. Additionally, the training data were shuffled at each epoch to introduce randomness among the data points, facilitating faster convergence towards the optimal solution.

If multiple boxes were detected in a single image, all boxes were classified. Boxes classified as error were ignored, and the output was determined as benign if only benign boxes were present. However, if even one box was classified as malignant, the output was set to malignant. Due to the nature of medical screening, where high sensitivity is more clinically significant than specificity, this approach was adopted to maximize practical utility for end users. Additionally, a 5-pixel padding was applied to enhance the visibility of lesion contours and facilitate comparisons between the internal color of the lesion and the surrounding liver tissue.

As shown in Table 2, the difference with and without padding was substantial. Without padding, the lesion contours resulting from clustering could become ambiguous, and comparing the color of the surrounding liver tissue with the lesion itself becomes nearly impossible. When padding was applied, these comparisons were feasible, leading to improved results. The detailed reasons for this will be discussed further in the discussion section.

Table 2. The accuracy, sensitivity, specificity, and confusion matrix depending on the presence or absence of padding. Blue tones represent correct answers, while red tones represent incorrect answers.

	Without Padding	With Padding
Accuracy	0.8248	0.9758
Sensitivity	0.7903	1.0000
Specificity	0.8327	0.9703
Confusion Matrix		

Using a CNN, we performed inference on the 172 bounding boxes detected by SM-WBF. The dataset consisted of 114 benign classes, 58 malignant classes, and 48 error lesion boxes. The inference results are summarized in Table 3. Out of a total of 220 bounding boxes, 218 were correctly classified, achieving an accuracy of 99.09%, which is highly satisfactory. Moreover, the sensitivity, a critical metric for screening tests, was perfect 100%. Specificity was 98.24%, with two benign lesions misclassified as the error class. Importantly, as evidenced by the test results, no cases of misclassification of benign or malignant lesions due to error boxes occurred, alleviating a major concern.

Table 3. The accuracy, sensitivity, specificity, and confusion matrix for 172 test datasets. Blue tones represent correct answers, while red tones represent incorrect answers.

Test with 172 Test Datasets			
Accuracy	0.9909		
Sensitivity	1.000	Confusion Matrix	
Specificity	0.9824		

The results in Table 3 highlight the robustness and reliability of the proposed neural network model as an effective screening tool for liver cancer detection. Additionally, the model can assist in identifying patients who require further examination, demonstrating its practical utility in clinical applications.

Rhyou et al. [37] proposed HCC-Net, which extracts lesions using YOLOv5 and applies wavelet transform. Each component was assigned ten different weight values, followed by inverse wavelet transform, generating a total of ten images. These outputs

were concatenated into a new 10-channel dataset for classification, achieving a classifier sensitivity of 0.9732. In comparison, our proposed method achieved a sensitivity of 1.0000, demonstrating its superior ability to detect malignant lesions. This higher sensitivity indicates that our approach is more effective in minimizing false negatives, making it highly valuable for clinical screening applications where missing a potential malignancy could have serious consequences.

4. Discussion

Our methodology has three significant points. First, due to the inherent characteristics of ultrasound images, such as considerable speckle noise and variations in resolution depending on device settings, we employed various image processing filters to address these challenges. Using contrast enhancement filters like intensity adjustment and adaptive histogram equalization, as well as low-pass and high-pass filters, we generated 10 different frequency-domain images from a single ultrasound image. This data augmentation approach allowed us to create 10 networks, each with its own mAP@0.5 score, which were utilized in the SM-WBF process.

Second, we incorporated clustering and adapted the traditional WBF into SM-WBF to better suit our network. In conventional WBF, the confidence scores of bounding boxes are used as weights for fusion, effectively assigning greater importance to higher-confidence boxes. However, in our system, every bounding box contains critical information, and applying standard WBF risked losing valuable data. Additionally, we aimed to include all suspected lesions, necessitating the use of clustering. By employing clustering and SM-WBF, we achieved higher lesion detection accuracy while preserving critical information.

Lastly, we applied padding to retain richer features. The importance of comparing colors lies in its diagnostic significance. For screening where histopathological data are unavailable, color plays a vital role. Malignant lesions often exhibit distinctly different tones and patterns compared to surrounding normal tissues, whereas benign lesions generally appear uniform in color. Irregular internal color patterns suggest malignancy. Similarly, contours are critical; malignant lesions typically have irregular, spiculated, or serrated edges, reflecting their invasive nature. These lesions often display poorly defined boundaries due to infiltration into surrounding tissues, while benign lesions usually have smooth, rounded, or oval shapes with well-defined borders, often encapsulated. Such morphological differences based on contours and internal features are essential for distinguishing malignant from benign lesions. By applying padding, we ensured that these features were preserved, leading to significant performance improvements in detection.

However, the methodology also has limitations. The proposed approach involves data augmentation with 10 filters, inference on each US image using YOLOv8, clustering, outlier reduction, SM-WBF, and CNN-based classification, leading to high computational costs that are unsuitable for real-time operation. This trade-off was deemed acceptable for cancer screening, where accuracy is prioritized over real-time performance. As a future direction, model optimization for real-time applications will be explored. Furthermore, while our methodology demonstrates significant improvements in lesion detection, further validation on larger and more diverse datasets is necessary to confirm its robustness across different clinical environments. The variability in ultrasound imaging conditions, including differences in equipment, operators, and patient characteristics, may affect the generalizability of the proposed approach. To address this, future studies will focus on testing the model with external datasets from multiple institutions to assess its reliability and adaptability in real-world clinical applications. Additionally, setting a low threshold to capture all potential lesions led to the generation of unnecessary boxes, potentially creating

redundant clusters during the clustering process. Further research will focus on refining the threshold values to minimize such occurrences.

5. Conclusions

By 2025, it is projected that over one million individuals will receive a liver cancer diagnosis annually. HCC constitutes approximately 75–85% of these cases. Fortunately, when detected at an early stage, HCC can be effectively managed before causing significant liver damage, emphasizing the critical need for efficient screening protocols for high-risk populations. However, fully automating the classification of liver cancer remains a significant challenge due to the inherent limitations of US imaging. These challenges include speckle noise, poor contrast between tumors and surrounding tissues, as well as the diverse morphology and echogenicity of lesions, all of which hinder their clinical applicability. Recently, various computer-aided diagnostic systems, including those leveraging deep learning techniques, have been developed to address these limitations and enhance hepatic lesion detection and classification.

This study presents "CSM-FusionNet", a model achieving clinically acceptable accuracy through the targeted application of image processing filters, clustering, SM-WBF, and padding. Raw US images were collected from one of South Korea's leading general hospitals, SMC. To overcome speckle noise inherent in ultrasound imaging and variations in resolution due to different settings, four types of filters were applied: intensity adjustment filter, adaptive histogram equalization filter, low-pass filter, and high-pass filter. These approaches not only reduced noise but also accounted for diverse resolutions, resulting in data augmentation that generated 10 US images from a single original image. Due to the limited nature of data acquisition for HCC, this method provided dual benefits by expanding the dataset. Each of the 10 US images was processed using the YOLOv8 model, producing 10 trained networks. The mAP@0.5 score of each network was measured and used to calculate weights for SM-WBF through SoftMax normalization.

Subsequently, bounding boxes were generated with a low threshold to identify multiple regions suspected of lesions, followed by clustering to segment these regions. Afterward, outliers were removed from each cluster, and SM-WBF was applied to merge multiple bounding boxes into a single one, reducing computational complexity. Padding was then added to the resulting bounding boxes to capture more features. The accuracy when using only original US images was 56.11%, which increased to 90.56% after applying clustering and WBF. Further application of SM-WBF elevated the accuracy to 95.56%, detecting lesions in 172 out of 180 test samples. Finally, classification was performed on the detected lesions, where padding also played a significant role. Without padding, the classification accuracy was 82.48%, while with padding, it increased to 97.58%, achieving 100% sensitivity—critical for computer aided diagnosis (CAD) systems.

Our technology holds significant potential to enhance the accuracy and efficiency of lesion detection and analysis using ultrasound imaging. By applying advanced fusion techniques like SM-WBF and clustering algorithms, it integrates information from multiple networks to deliver reliable diagnostic results. This enables clinicians to detect even minute lesions that might otherwise be overlooked, reduces diagnostic time, and supports better decision-making for personalized treatment. Furthermore, it enhances the efficiency of healthcare systems through automation and enables precise diagnostics even in resource-limited areas, improving healthcare equity. Ultimately, our technology underscores the importance of early diagnosis and preventative management, contributing to improved survival rates and enhanced quality of life for patients.

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

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of SAMSUNG MEDICAL CENTER (SMC-2020-10-178-002 and 28 November 2020).

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

Data Availability Statement: Data available on request due to restrictions.

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

Abbreviations

The following abbreviations are used in this manuscript:

US	Ultrasound
HCC	Hepatocellular Carcinoma
SM-WBF	SoftMax-Weighted Box Fusion
SMC	Samsung Medical Center
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
CNN	Convolutional Neural Network
YOLO	You Only Look Once
WBF	Weighted Box Fusion
IRB	Institutional Review Board
MoE	Mixture of Experts
mAP	Mean Accuracy Precision
IoU	Intersection of Union
CAD	Computer Aided Diagnosis

References

1. Brown, Z.J.; Tsilimigras, D.I.; Ruff, S.M.; Mohseni, A.; Kamel, I.R.; Cloyd, J.M.; Pawlik, T.M. Management of Hepatocellular Carcinoma: A Review. *JAMA Surg.* **2023**, *158*, 410–420. [CrossRef] [PubMed]
2. Devarbhavi, H.; Asrani, S.K.; Arab, J.P.; Nartey, Y.A.; Pose, E.; Kamath, P.S. Global Burden of Liver Disease: 2023. *J. Hepatol.* **2023**, *79*, 516–537. [CrossRef] [PubMed]
3. Siegel, R.L.; Giaquinto, A.N.; Jemal, A. Cancer statistics, 2024. *CA A Cancer J. Clin.* **2024**, *74*, 12–49. [CrossRef] [PubMed]
4. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2024**, *74*, 229–263. [CrossRef]
5. Tümen, D.; Heumann, P.; Gülow, K.; Demirci, C.-N.; Cosma, L.-S.; Müller, M.; Kandulski, A. Pathogenesis and Current Treatment Strategies of Hepatocellular Carcinoma. *Biomedicines* **2022**, *10*, 3202. [CrossRef]
6. de Martel, C.; Georges, D.; Bray, F.; Ferlay, J.; Clifford, G.M. Global burden of cancer attributable to infections in 2018: A worldwide incidence analysis. *Lancet Glob. Health* **2020**, *8*, 180–190. [CrossRef] [PubMed]
7. Rumgay, H.; Ferlay, J.; de Martel, C.; Georges, D.; Ibrahim, A.S.; Zheng, R.; Wei, W.; Lemmens, V.E.; Soerjomataram, I. Global, regional and national burden of primary liver cancer by subtype. *Eur. J. Cancer* **2022**, *161*, 108–118. [CrossRef] [PubMed]
8. Cancer of the Liver Italian Program (CLIP) Investigators. A new prognostic system for hepatocellular carcinoma: A retrospective study of 435 patients. *Hepatology* **1998**, *28*, 751–755. [CrossRef] [PubMed]
9. Llovet, J.M.; Kelley, R.K.; Villanueva, A.; Singal, A.G.; Pikarsky, E.; Roayaie, S.; Lencioni, R.; Koike, K.; Zucman-Rossi, J.; Finn, R.S. Hepatocellular carcinoma (primer). *Nat. Rev. Dis. Primers* **2021**, *7*, 6. [CrossRef] [PubMed]
10. Zhang, B.H.; Yang, B.H.; Tang, Z.Y. Randomized controlled trial of screening for hepatocellular carcinoma. *J. Cancer Res. Clin. Oncol.* **2004**, *130*, 417–422. [CrossRef] [PubMed]
11. Parra, N.S.; Ross, H.M.; Khan, A.; Wu, M.; Goldberg, R.; Shah, L.; Mukhtar, S.; Beiriger, J.; Gerber, A.; Halegoua-DeMarzio, D. Advancements in the Diagnosis of Hepatocellular Carcinoma. *Int. J. Transl. Med.* **2023**, *3*, 51–65. [CrossRef]

12. Kuo, S.C.; Lin, C.N.; Lin, Y.J.; Chen, W.Y.; Hwang, J.S.; Wang, J.D. Optimal Intervals of Ultrasonography Screening for Early Diagnosis of Hepatocellular Carcinoma in Taiwan. *JAMA Netw. Open* **2021**, *4*, e2114680. [CrossRef] [PubMed]
13. Zhao, C.; Nguyen, M.H. Hepatocellular Carcinoma Screening and Surveillance: Practice Guidelines and Real-Life Practice. *J. Clin. Gastroenterol.* **2016**, *50*, 120–133. [CrossRef]
14. Tsuchiya, N.; Sawada, Y.; Endo, I.; Saito, K.; Uemura, Y.; Nakatsura, T. Biomarkers for the Early Diagnosis of Hepatocellular Carcinoma. *World J. Gastroenterol.* **2015**, *21*, 10573–10583. [CrossRef] [PubMed]
15. Hu, H.; Zhao, Y.; He, C.; Qian, L.; Huang, P. Ultrasonography of Hepatocellular Carcinoma: From Diagnosis to Prognosis. *J. Clin. Transl. Hepatol.* **2024**, *12*, 516–524. [CrossRef] [PubMed]
16. Hall, A.J.; Wild, C.P. Liver Cancer in Low and Middle Income Countries. *BMJ* **2003**, *326*, 994–995. [CrossRef] [PubMed]
17. Ahn, J.C.; Lee, Y.T.; Agopian, V.G.; Zhu, Y.; You, S.; Tseng, H.R.; Yang, J.D. Hepatocellular Carcinoma Surveillance: Current Practice and Future Directions. *Hepatoma Res.* **2022**, *8*, 10. [CrossRef]
18. Rodgers, S.K.; Fetzer, D.T.; Seow, J.H.; McGillen, K.; Burrowes, D.P.; Fung, C.; Udare, A.S.; Wilson, S.R.; Kamaya, A. Optimizing US for HCC Surveillance. *Abdom. Radiol.* **2024**. [CrossRef] [PubMed]
19. Lee, J.H.; Jeong, Y.K.; Park, K.B.; Park, J.K.; Jeong, A.K.; Hwang, J.C. Operator-Dependent Techniques for Graded Compression Sonography to Detect the Appendix and Diagnose Acute Appendicitis. *Am. J. Roentgenol.* **2005**, *184*, 91–97. [CrossRef] [PubMed]
20. Rosario, P.W. Ultrasonography for the Follow-Up of Patients with Papillary Thyroid Carcinoma: How Important Is the Operator? *Thyroid* **2010**, *20*, 833–835. [CrossRef] [PubMed]
21. Michał, B.; Grzegorz, S.; Cezary, S.; Piotr, K.; Łukasz, M.; Rafał, P.; Bogna, Z.; Krzysztof, Z.; Piotr, S.; Andrzej, N. Transfer Learning with Deep Convolutional Neural Network for Liver Steatosis Assessment in Ultrasound Images. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1895–1903.
22. Sparchez, Z.; Craciun, R.; Caraiani, C.; Horhat, A.; Nenu, I.; Procopet, B.; Sparchez, M.; Stefanescu, H.; Mocan, T. Ultrasound or Sectional Imaging Techniques as Screening Tools for Hepatocellular Carcinoma: Fall Forward or Move Forward? *J. Clin. Med.* **2021**, *10*, 903. [CrossRef]
23. Candita, G.; Rossi, S.; Cwiklinska, K.; Fanni, S.C.; Cioni, D.; Lencioni, R.; Neri, E. Imaging Diagnosis of Hepatocellular Carcinoma: A State-of-the-Art Review. *Diagnostics* **2023**, *13*, 625. [CrossRef]
24. Rodriguez De Santiago, E.; Tellez, L.; Guerrero, A.; Albillos, A. Hepatocellular Carcinoma After Fontan Surgery: A Systematic Review. *Hepatol. Res.* **2021**, *51*, 116–134. [CrossRef] [PubMed]
25. Addissouky, T.A.; Sayed, I.E.T.E.; Ali, M.M.; Wang, Y.; Baz, A.E.; Khalil, A.A.; Elarabany, N. Latest Advances in Hepatocellular Carcinoma Management and Prevention Through Advanced Technologies. *Egypt. Liver J.* **2024**, *14*, 2. [CrossRef]
26. Rhyou, S.Y.; Yoo, J.C. Aggregated Micropatch-Based Deep Learning Neural Network for Ultrasonic Diagnosis of Cirrhosis. *Artif. Intell. Med.* **2023**, *139*, 102541. [CrossRef] [PubMed]
27. Wei, Q.; Tan, N.; Xiong, S.; Luo, W.; Xia, H.; Luo, B. Deep Learning Methods in Medical Image-Based Hepatocellular Carcinoma Diagnosis: A Systematic Review and Meta-Analysis. *Cancers* **2023**, *15*, 5701. [CrossRef] [PubMed]
28. Nishida, N.; Yamakawa, M.; Shiina, T.; Mekada, Y.; Nishida, M.; Sakamoto, N.; Nishimura, T.; Iijima, H.; Hirai, T.; Takahashi, K.; et al. Artificial Intelligence (AI) Models for the Ultrasonographic Diagnosis of Liver Tumors and Comparison of Diagnostic Accuracies Between AI and Human Experts. *J. Gastroenterol.* **2022**, *57*, 309–321. [CrossRef] [PubMed]
29. Chen, J.; Zhang, W.; Bao, J.; Wang, K.; Zhao, Q.; Zhu, Y.; Chen, Y. Implications of Ultrasound-Based Deep Learning Model for Preoperatively Differentiating Combined Hepatocellular-Cholangiocarcinoma from Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma. *Abdom. Radiol.* **2024**, *49*, 93–102. [CrossRef]
30. Li, M.D.; Li, W.; Lin, M.X.; Lin, X.X.; Hu, H.T.; Wang, Y.C.; Ruan, S.M.; Huang, Z.R.; Lu, R.F.; Li, L.; et al. Systematic Comparison of Deep-Learning Based Fusion Strategies for Multi-Modal Ultrasound in Diagnosis of Liver Cancer. *Neurocomputing* **2024**, *603*, 128257. [CrossRef]
31. Rhyou, S.Y.; Yoo, J.C. Cascaded Deep Learning Neural Network for Automated Liver Steatosis Diagnosis Using Ultrasound Images. *Sensors* **2023**, *21*, 5304. [CrossRef] [PubMed]
32. Rhyou, S.Y.; Cho, Y.J.; Yoo, J.C.; Hong, S.H.; Bae, S.H.; Bae, H.J.; Yu, M.Y. Automatic Lower-Limb Length Measurement Network (A3LMNet): A Hybrid Framework for Automated Lower-Limb Length Measurement in Orthopedic Diagnostics. *Electronics* **2024**, *14*, 160. [CrossRef]
33. Ryu, H.; Shin, S.Y.; Lee, J.Y.; Lee, K.M.; Kang, H.J.; Yi, J.H. Joint Segmentation and Classification of Hepatic Lesions in Ultrasound Images Using Deep Learning. *Eur. Radiol.* **2021**, *31*, 8733–8742. [CrossRef]
34. Zhao, T.; Zeng, Z.; Li, T.; Tao, W.; Yu, X.; Feng, T.; Bu, R. USC-ENet: A High-Efficiency Model for the Diagnosis of Liver Tumors Combining B-Mode Ultrasound and Clinical Data. *Health Inf. Sci. Syst.* **2023**, *11*, 15. [CrossRef]
35. Poreddy, A.K.R.; Lingamaiah, S.C.; Krishna, T.B.; Kokil, P. Focal Liver Lesion Classification Based on Statistical Variations of Discrete Haar Wavelet Transform and Singular Value Decomposition. *IEEE Sens. Lett.* **2024**, *8*, 6008604. [CrossRef]

36. Chaiteerakij, R.; Ariyaskul, D.; Kulkraisri, K.; Apiparakoon, T.; Sukcharoen, S.; Chaichuen, O.; Pensuwan, P.; Tiyarattanachai, T.; Rerknimitr, R.; Marukatat, S. Artificial Intelligence for Ultrasonographic Detection and Diagnosis of Hepatocellular Carcinoma and Cholangiocarcinoma. *Sci. Rep.* **2024**, *14*, 20617. [CrossRef] [PubMed]
37. Rhyou, S.Y.; Yoo, J.C. Automated Ultrasonography of Hepatocellular Carcinoma Using Discrete Wavelet Transform-Based Deep-Learning Neural Network. *Med. Image Anal.* **2025**, *101*, 103453. [CrossRef]

diagnostics

MDPI

Article

Detection of Gallbladder Disease Types Using a Feature Engineering-Based Developed CBIR System

Ahmet Bozdag [1], Muhammed Yildirim [2,*], Mucahit Karaduman [3], Hursit Burak Mutlu [2], Gulsah Karaduman [4] and Aziz Aksoy [5]

[1] Department of General Surgery, School of Medicine, Firat University, Elazığ 23119, Turkey; abozdag@firat.edu.tr
[2] Department of Computer Engineering, Malatya Turgut Ozal University, Malatya 44210, Turkey; burakmutlu44@gmail.com
[3] Department of Software Engineering, Malatya Turgut Ozal University, Malatya 44210, Turkey; mucahit.karaduman@ozal.edu.tr
[4] Department of Computer Engineering, Firat University, Elazığ 23119, Turkey; gkaraduman@firat.edu.tr
[5] Department of Bioengineering, Malatya Turgut Ozal University, Malatya 44200, Turkey; aziz.aksoy@ozal.edu.tr
* Correspondence: muhammed.yildirim@ozal.edu.tr

Abstract: Background/Objectives: Early detection and diagnosis are important when treating gallbladder (GB) diseases. Poorer clinical outcomes and increased patient symptoms may result from any error or delay in diagnosis. Many signs and symptoms, especially those related to GB diseases with similar symptoms, may be unclear. Therefore, highly qualified medical professionals should interpret and understand ultrasound images. Considering that diagnosis via ultrasound imaging can be time- and labor-consuming, it may be challenging to finance and benefit from this service in remote locations. **Methods:** Today, artificial intelligence (AI) techniques ranging from machine learning (ML) to deep learning (DL), especially in large datasets, can help analysts using Content-Based Image Retrieval (CBIR) systems with the early diagnosis, treatment, and recognition of diseases, and then provide effective methods for a medical diagnosis. **Results:** The developed model is compared with two different textural and six different Convolutional Neural Network (CNN) models accepted in the literature—the developed model combines features obtained from three different pre-trained architectures for feature extraction. The cosine method was preferred as the similarity measurement metric. **Conclusions:** Our proposed CBIR model achieved successful results from six other different models. The AP value obtained in the proposed model is 0.94. This value shows that our CBIR-based model can be used to detect GB diseases.

Keywords: artificial intelligence; gallstone diseases; CBIR; carcinoma; cholecystitis

Academic Editors: Hiam Alquran and Wan Azani Mustafa

Received: 21 January 2025
Revised: 22 February 2025
Accepted: 23 February 2025
Published: 25 February 2025

Citation: Bozdag, A.; Yildirim, M.; Karaduman, M.; Mutlu, H.B.; Karaduman, G.; Aksoy, A. Detection of Gallbladder Disease Types Using a Feature Engineering-Based Developed CBIR System. *Diagnostics* **2025**, *15*, 552. https://doi.org/10.3390/diagnostics15050552

1. Introduction

Gallbladder (GB) disease, a frequent pathology, must be diagnosed accurately and early on. The GB is an intraperitoneal organ located on the right lower surface of the liver. The gallbladder is located close to the duodenum, pancreas, and transverse colon. Anatomically, it is divided into three parts: the fundus and body, infundibulum, and neck. By holding and releasing bile produced by the liver, the gallbladder contributes significantly to the digestion process. Bile is a fluid produced by the liver that aids the digestive system and is discharged into the duodenum through the bile duct system for digestion and fat

absorption [1]. Bile acids, phospholipids, lecithin, cholesterol, and bilirubin are the main constituents of bile.

Cholelithiasis (gallstones) is the most prevalent GB condition, affecting about 10–15% of the adult population [2]. Nine pathological problems have been shown among the most common gallbladder diseases today. These include GB perforation, polyps and cholesterol crystals, gallbladder wall thickening, GB adenomyomatosis, gallstones, cholecystitis, gangrenous cholecystitis, carcinoma, and issues with the intraabdominal and retroperitoneum [3]. When the concentrations of the components that comprise bile surpass their solubility points, gallstones develop. Excessive cholesterol secretion relative to lecithin and bile salts causes cholesterol gallstone formation. However, it is known that gallstones are primarily due to genetic predisposition [4]. The risk also increases due to age, diabetes, and obesity. In Native Americans, prevalence rates of 30% in men and 60% in women have been shown [5]. The main symptoms of gallbladder disease are bloating, fever, frequent retching, change in skin color (jaundice), and pain in the upper abdomen, as shown in Figure 1 [6].

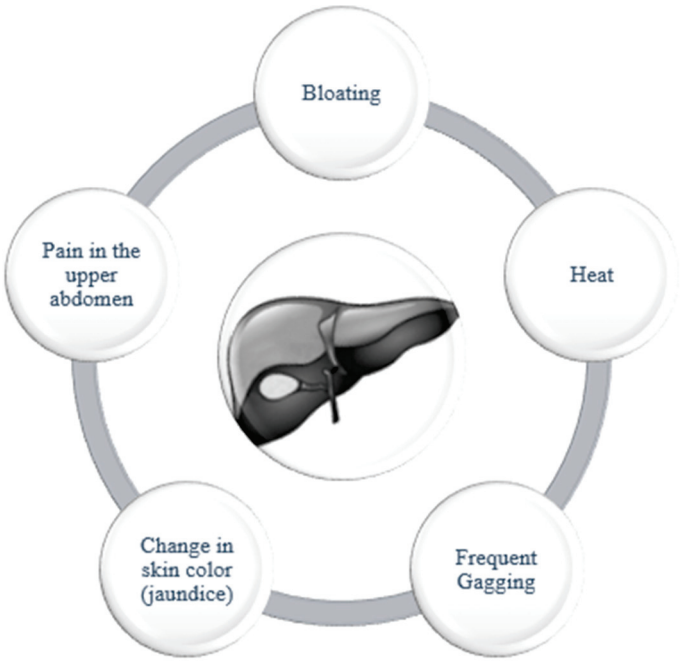


Figure 1. Main symptoms of gallbladder disease.

Before gallbladder surgery, the risk of choledocholithiasis should be assessed because gallbladder and common bile duct stones are frequently detected combined. Retrograde endoscopic cholangiography is used to remove stones from the common bile duct, and a cholecystectomy is performed if choledocholithiasis is verified by ultrasound, endoscopic ultrasound, or magnetic resonance cholangiography [7]. Gallstones are a major global health concern because, according to the recent research, they may put patients at risk for various illnesses, like cancer, cardiovascular disease, and even greater mortality [8–10]. Because of its aggressive nature and lack of effective treatment options, gallbladder carcinoma (GBC), the most prevalent malignant tumor of the biliary tract, has a terrible prognosis. The early detection of GBC is a great challenge for physicians, and most GBCs are incidentally detected during cholecystectomy procedures for gallstones [11]. Gallbladder polyps (GBPs) are elevated lesions of the gallbladder mucosa that extend into the gallbladder lumen. The prevalence of GBPs varies by region and ethnicity, ranging from 0.3% to 9.5%. Pathologically, GBPs can be separated into non-neoplastic polyps, such as gallbladder adenomyosis, cholesterol polyps, and inflammatory polyps, and neoplastic polyps, which are linked to adenomas and adenocarcinomas [12].

Because of the organ's intricate anatomy and the variety of medical states that can affect it, diagnosing GB illnesses can be difficult. With the rapid growth of medical science and technology in recent years, artificial intelligence (AI) techniques have the potential to reduce human efforts significantly [13]. Based on ultrasound images (UIs), the AI field of deep learning (DL) is an instructive medical tomography technique that can aid in the early identification of GB disease. In order to assist medical personnel in identifying and categorizing different GB pathological anomalies, ML and deep learning (DL) approaches have recently demonstrated a notable capacity to analyze medical images [3]. By automating the analysis process and boosting diagnostic accuracy based on extensive datasets and sophisticated algorithms, these methods have the promising potential to improve GB diagnosis [14]. This study used a CBIR-based model to simultaneously detect nine gallbladder diseases and determine the disease type using UIs. The CBIR model is the search and analysis of different content-based image features.

CBIR is used to identify various schemas (modalities) in many images. Compared to a query image, CBIR lets clinicians retrieve pertinent images from an extensive database, minimizing time-consuming manual searches and aiding in diagnosis [15]. Based on a query image, CBIR seeks to identify related images from a big database. Numerous medical imaging domains are being actively researched in this discipline [16]. In addition to the deep learning architectures already mentioned, the suggested CBIR system offers interaction information to allow the user to choose which disease in the slice is being queried. CBIR can more effectively match images in this query for successful retrieval; that is, it can return more pertinent images. The CBIR system suggested in this paper can automatically produce pertinent image characteristics from well-annotated image datasets. It can also retrieve images to collect similar images and previous cases to obtain comprehensive information about the patient's conditions.

Class consistency in the top-level images is a common way to evaluate a CBIR system's quality. CBIR can improve the efficiency of time-consuming clinical workflow operations. The use of automated assistance systems, such as CBIR, in diagnosing numerous illnesses is growing in significance [16].

Images used in medicine are complicated. Due to the complex imaging data and the modest distinctions between disease states, automated techniques for objective lesion characterization are required. This approach can speed up radiology workflows and improve the overall quality of healthcare. Therefore, a CBIR-based model was developed, which produced successful results in GB diagnosis.

1.1. Releated Works

There are some studies in the literature for the detection of GB diseases. Unlike our study, deep learning-based methods were used in most of these studies. Lo et al. (2024) [17] collected a total of 2827 abdominal CT slices from computed tomography (CT) images to recognize the seven most relevant organs in the abdomen. DenseNet and transformer-based methods were preferred for the automatic detection of organs. Accuracy values in the range of 94–99% were achieved in the models used. Gupta et al. (2024) [18] compared the diagnostic performance of gallbladder cancer (GBC) with deep learning (DL)- based CNN architectures and skilled radiologists. The study included 565 patients, 334 of whom had gallstones and gallbladder diseases. When it came to detecting GBCs on the US CNN (0.836–0.945), Radiologist1 (0.733–0.891), and Radiologist2 (0.761–0.909), the DL-based method performed as well as or better than the expert radiologists. Zhou et al. (2024) [19] retrospectively trained a previously trained deep learning-based smartphone application to help diagnose biliary atresia from ultrasonographic gallbladder images using 3659 original sonographic gallbladder images and 51,226 smartphone photographs. A novice radiolo-

gist and an experienced pediatric radiologist also tested a new model. The new model's diagnostic performance was more consistent and on par with that of seasoned pediatric radiologists. Yu et al. (2021) [20] recorded 89,000 abdominal US images from 2386 patients in the hospital database to detect and localize gallstones and cholecystitis with acceptable separation and speed. Using SSD-FPN-ResNet-50 and MobileNet V2 architectures, they determined accuracies of 0.92 and 0.94, respectively. Dadjouy and H. Sajedi (2024) [21] used the Faster R-CNN and YOLOv8 fusion method for gallbladder detection in ultrasound images to diagnose gallbladder cancer. Although the Faster R-CNN could estimate highly accurate bounding boxes, it also produced multiple bounding boxes that misidentified the background. In contrast, YOLO correctly estimated the location of the bounding boxes. This was achieved with 90.16% and 82.79% accuracies with Faster R-CNN and YOLOv8 fusion methods. Esen et al. (2024) [22] showed that the gradient boosting technique achieved the highest accuracy (85.42%) in predicting gallstones using a dataset consisting of Bioimpedance and the laboratory data of 319 individuals, 161 gallstone patients, and 158 healthy controls. Chattopadhyay et al. (2005) [23] detected gallbladder abnormalities with 92.3% accuracy from ultrasound scan (USS) images obtained from 751 cholecystectomy patients. Pang et al. (2019) [24] determined the diagnosis of cholelithiasis with 90.8% accuracy in the analysis of 1300 CT images of cholelithiasis patients with the CNN architecture (MobileNetV1, SSD, YOLOv2, and original SSD (with VGG-16)). Jang et al. (2021) [25] analyzed 1039 endoscopic ultrasound (EUS) images in patients with gallbladder polypoid lesions using the CNN architecture ResNet50. The rates were 57.9%, 96.5%, 77.8%, 91.6%, and 89.8% for EUS-AI in the differential diagnosis of neoplastic and non-neoplastic GB polyps. Veena et al. (2022) [26] analyzed the data from computerized tomography (CT) images of gallbladder diseases with SSD-efficient-Det, Faster R-CNN, and Mask R-CNNs models. They achieved 0.938 accuracy in the classification with Mask R-CNN architecture. In the study by Song et al. (2019) [27], the dataset included 5350 images from 726 patients who segmented CT images of gallbladder stones. In this study, the researchers achieved 91.68% success using DC-GAN and CNN architecture.

1.2. Contribution and Novelty

- It is seen that CNN-based architectures are primarily used in the literature to detect GB diseases. CNN-based architectures cannot produce successful results, especially in studies where the number of classes and data are high, and the training of the architectures takes a very long time in studies where large datasets are used.
- Considering that the dataset used in the study carried out for the detection of GB diseases has nine classes, we believe that this dataset is suitable for CBIR systems.
- Considering that the early detection of GB diseases is vital, it is very important to be able to detect these diseases at an early stage with computer-aided models.
- A new CBIR system has been developed to detect GB diseases early and highlight the success of CBIR systems in detecting GB diseases.
- In the developed system, feature extraction from different architectures and similarity measurement metrics were tested. As a result, features obtained from three different models were concatenated. At this stage, different features of the same image were used together.
- The results were also obtained with different CNN architectures and textural-based features to test the model's performance. Different metrics were used to test the performance of the architectures.
- As a result, the proposed CBIR-based model obtained the most successful results in the detection of GB diseases.

1.3. Organization of Paper

In the rest of the article, the structure of the dataset used in the study, methods, similarity measurement metrics, feature extraction techniques, and the developed model are presented. Then, the experimental results, discussion, and conclusion sections are presented.

2. Materials and Methods

2.1. Datasets of Ultrasound Images of Gallbladder Diseases

Ultrasound images of the gallbladder organ taken from inside the digestive system made up the dataset. JPEG images of the digestive system were included in this dataset. Siemens Acuson X700, Philips Affiniti 70, Philips CX50, and Canon Viamo c100 ultrasound machines were used to create the images [28,29].

2.2. Data Processing and Data Collection

To maintain uniformity throughout the dataset, every image was enlarged to 900 × 1200 pixels with a 600 px horizontal resolution and 600 px vertical resolution, even though the original images had 450 × 600 pixels, a 3:4 aspect ratio, and 150 px horizontal and 150 px vertical resolutions. In total, 10,692 IU images from 1782 patients were used (Table 1) [29].

Table 1. Gallbladder disease type dataset details.

Gallbladder Disease Type	Female	Male	Number of Patients	Number of Images
Gallstones	137	84	221	1326
Intraabdominal and retroperitoneum problems	110	85	195	1170
Cholecystitis	102	89	191	1146
Membranous and gangrenous cholecystitis	109	95	204	1224
Perforation	95	82	177	1062
Polyps and cholesterol crystals	99	71	170	1020
Adenomyomatosis	108	86	194	1164
Carcinoma	155	110	265	1590
Various causes of GB wall thickening	92	73	165	990
Total	1.007	775	1.782	10.692

Sample images from the dataset are presented in Figure 2.

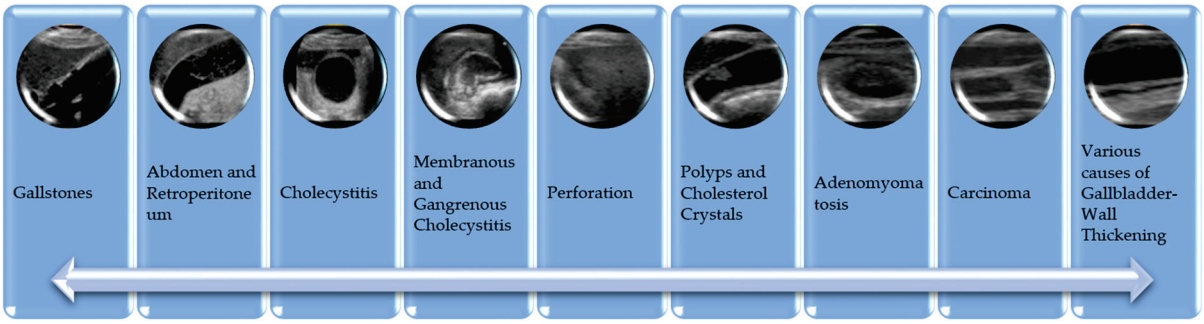


Figure 2. Gallbladder disease pathology IU images.

2.3. CBIR System Developed for the Detection of Gallbladder Disease Types

The feature map of the images related to the proposed model was extracted. This process produced a feature vector of 10,692 × 1000 corresponding to 10,692 images in the

dataset. The proposed model combined feature maps extracted from three different models, aiming to work with different features of the same image. Finally, the combined feature maps were adopted as a feature extraction model in the developed CBIR system. The diagram of the suggested CBIR system is presented in Figure 3.

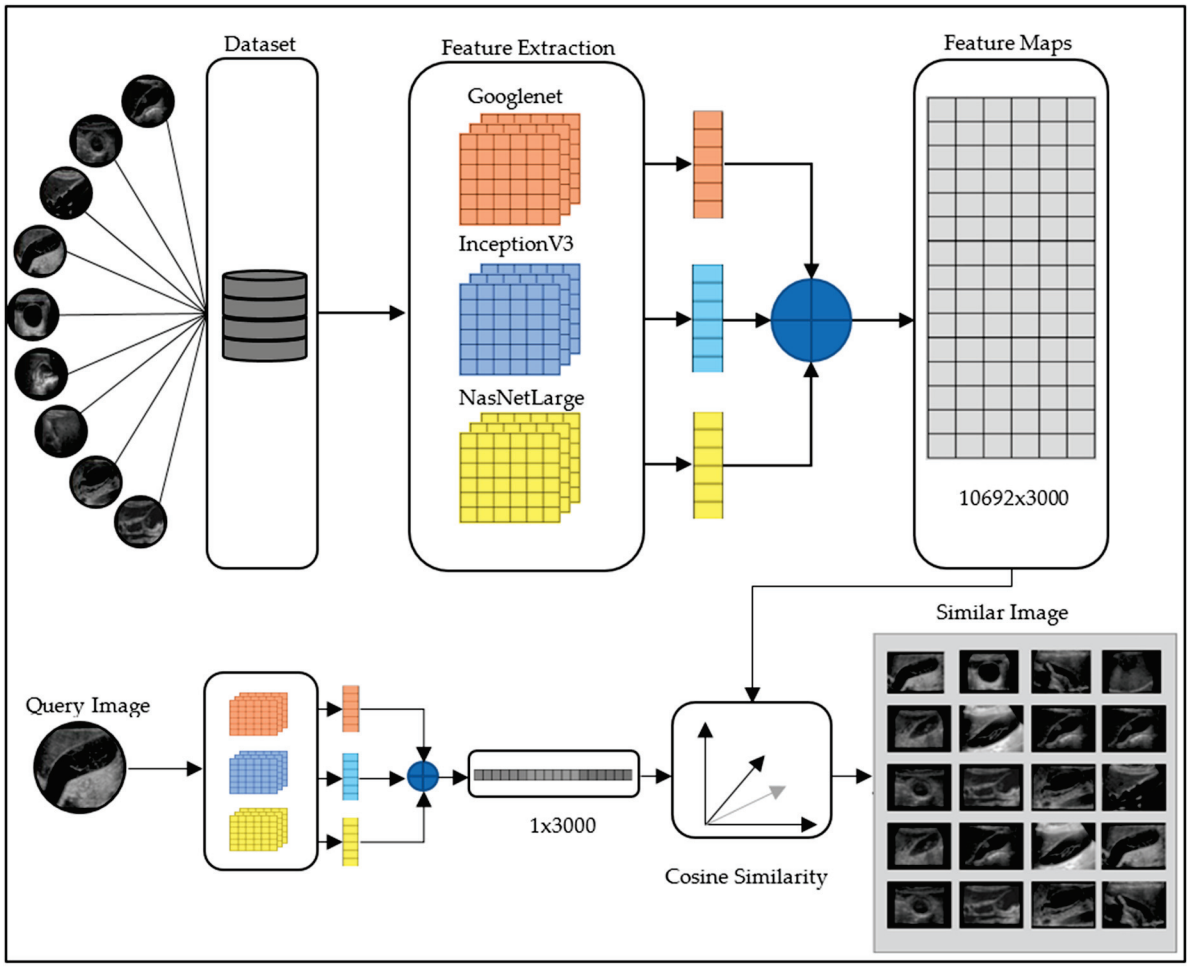


Figure 3. Developed CBIR system.

Image retrieval was conducted using the feature maps generated by the CBIR-based systems. In the proposed CBIR system, the feature map of the query image was extracted using the suggested model, and its performance was compared with other CNN architecture- and textural-based methods. The comparison was performed using cosines similarity measurement techniques, and the evaluation was carried out by analyzing the precision–recall (P-R) curve using an 11-point interpolated retrieval curve. In CNN architectures, Googlenet, InceptionV3, NasNetLarge, DenseNet201, and LBP and HOG architectures in machine learning classifiers were compared with the proposed model. The proposed model produced more successful results compared to the models used in this study.

2.4. Cosine Similarity Measure

The cosine similarity calculated for two vectors is the ratio of the calculation obtained from the product of the cosine angle of these two vectors. When the cosine similarity value

of two vectors was 1, it was decided that they are similar. Equation (1) was used for the cosine similarity calculation [30].

$$\text{Cos}\alpha = \frac{A \cdot B}{|A| \cdot |B|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

In Equation (1), A represents the weight of each feature of vector A, and B represents the weight of each feature in vector B. According to the cosine similarity rule, the smaller the angle for comparing two vectors, the greater the degree of similarity. Twenty related images in the dataset were accessed by the suggested CBIR model, as seen in Figure 4. From the dataset's classes, one image was chosen at random, and the 20 related images that were retrieved are displayed in order. The images accessed in the correct class and those accessed in the wrong are indicated with class labels.

Figure 4. Examples of the queried image with the proposed CBIR systems.

Figure 4 shows 20 images taken from the dataset by querying a random image belonging to the gallstones class. As a result of the query, images 1–12, 14, 15, 17, and 18 are in the actual class, while images 13, 16, 19, and 20 are in different classes.

3. Application Results

The MATLAB 2024b program was used to obtain the findings of this research on a computer running Windows 10, 64 bit, equipped with an Intel i7 processor and 32 GB of RAM.

The order in which the images are accessed is crucial in content-based image access. Standard P-R graphs with 11 points are used for the evaluations. A P-R graph is provided for 20 randomly selected images in each class in the CBIR system that we created. The P-R graph employed the cosine approach, one of the techniques for measuring similarity.

The interpolated 11-point sequential access was assessed in this study using the average precession value and the P-R curve. The data collection consists of nine classes. The average P-R curve was obtained by querying each of the nine classes' images separately.

Twenty images were retrieved from the CBIR system after each class's images were queried independently. By analyzing the images accessed and queried in each class, the average P-R curve was produced. A total of 10,692 images in the dataset were queried after the P-R curves of nine classes were assessed independently. The average P-R curve of the 20 images retrieved in each query was then computed and assessed.

Figure 5a shows the P-R curve for the gallstones class of six different architectures and the proposed model. The cosine method is used as the distance measurement metric. Since there are 1326 images in the gallstones class, 26,520 images are accessed. Figure 5b plots the curves of six different architectures and the proposed model for the images in the abdomen class. The cosine method is used as the distance measurement metric. Since there are 1170 images in the abdomen class, 23,400 images are accessed.

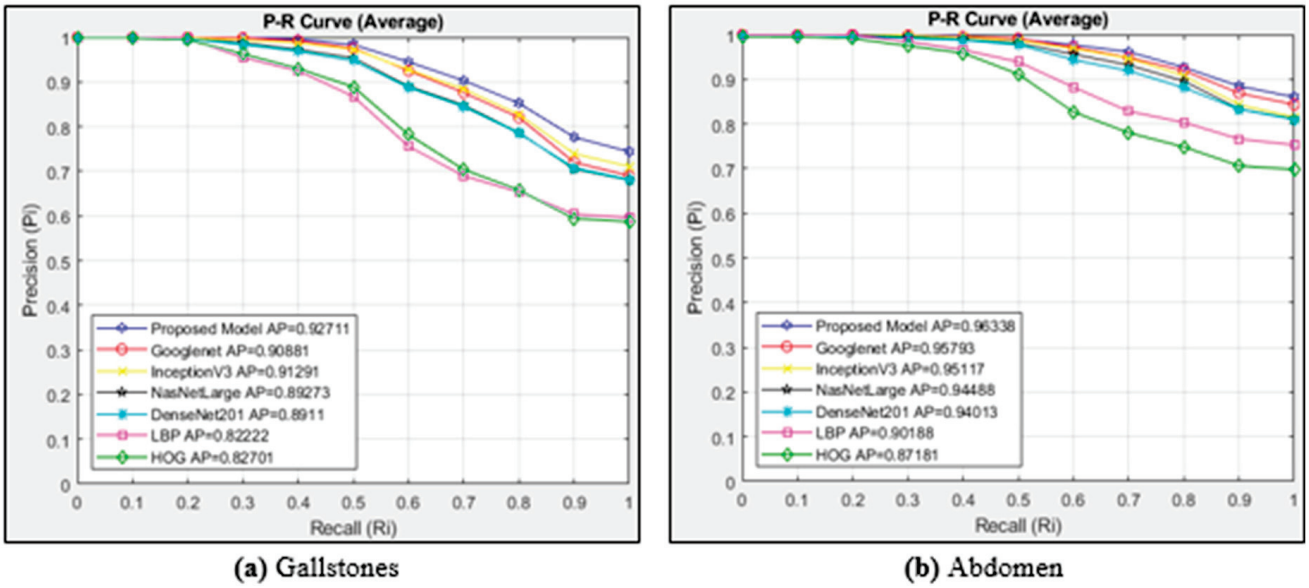


Figure 5. Average P-R curves for the classes of (**a**) gallstones and (**b**) abdomen.

In Figure 5a, the P-R curves are calculated for the gallstones class using the cosine similarity measurement method for different models and the proposed model. While the proposed model achieved the highest AP value in the gallstones class, the least successful architecture was LBP. In the gallstones class, CNN-based methods achieved more successful results than textural-based methods. The AP value of the proposed model was 0.92711. The proposed model was successful in all cases in the recall {0. . .1} range in the gallstones class compared to the other six models. The models showed a high performance in cases up to Ri 0.2. When Figure 5b is examined, it can be seen that the proposed model shows the best performance in the abdomen class with an AP value of 0.96338. The textural-based HOG method performed the worst in the abdomen class, with an AP value of 0.87181. CNN architectures also produced more successful results than textural-based architectures in the abdomen class. Googlenet, InceptionV3, NasNetLarge, and DenseNet201 architectures achieved similar performances in the abdomen class.

Figure 6a shows the P-R curves obtained for the cholecystitis class using the cosine distance measurement metric in different models. Since there are 1146 images in the cholecystitis class, 22,920 images are accessed. Figure 6b plots the curves of six different architectures and the proposed model for the images in the membranous class. The cosine method is used as the distance measurement metric. Since there are 1224 images in the membranous class, 24,480 images are accessed.

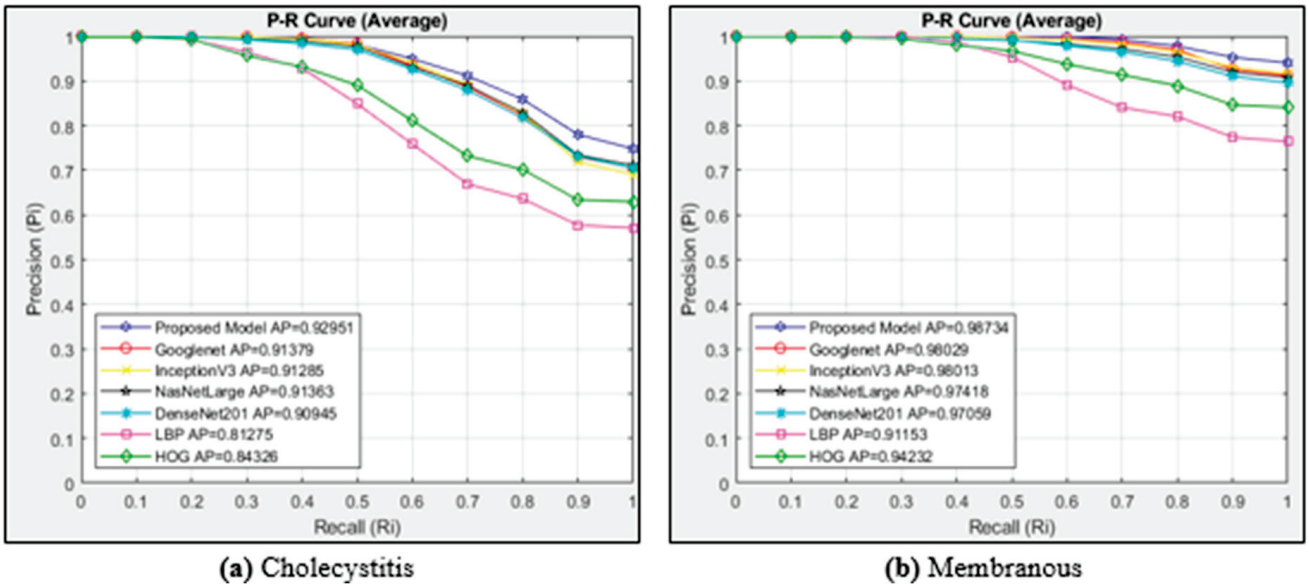


Figure 6. Average P-R curves for the classes of (**a**) cholecystitis and (**b**) membranous.

In Figure 6a, the P-R curves are calculated for the cholecystitis class using the cosine similarity measurement method for different models and the proposed model. While the proposed model achieved the highest AP value in the cholecystitis class, the least successful architecture was LBP. In the cholecystitis class, CNN-based methods achieved more successful results than textural-based methods. The AP value of the proposed model was 0.92951. The proposed model was successful in all cases in the recall {0...1} range in the cholecystitis class compared to the other six models. The models showed high performances in cases up to Ri 0.2. When Figure 6b is examined, it can be seen that the proposed model shows the best performance in the membranous class with an AP value of 0.98734. The textural-based LBP method performed the worst in the membranous class, with an AP value of 0.91153. CNN architectures also produced more successful results than textural-based architectures in the membranous class. NasNetLarge, DenseNet201, Googlenet, InceptionV3, HOG, and LBP architectures followed the proposed model.

The P-R curve for the perforation class images is shown in Figure 7a. This figure plots the curves of six different architectures and the proposed model. Since there are 1062 images in the perforation class, 21,240 images are accessed. Figure 7b plots the curves of six different architectures and the proposed model for the images in the polypose class. The cosine method is used as the distance measurement metric. Since there are 1020 images in the polypose class, 20,400 images are accessed.

Figure 7a plots P-R curves for the perforation class using the cosine distance measurement method. The proposed model obtained the highest AP value in the perforation class. The AP value of the proposed model is 0.93358. The least successful model in the perforation class is LBP, with an AP value of 0.81454. The highest AP value after the proposed model was achieved with the Googlenet architecture. The proposed model achieved a performance close to 1 in the range of 0–0.4. In Figure 7b, the most successful model in the polypose class was the proposed model with an AP value of 0.9678. Similar values were obtained in the other four CNN architectures used for comparison purposes. Textural-based models achieved lower AP values. In the proposed model, Ri decreased after the value of 0.5. It remained close to 1 until this interval.

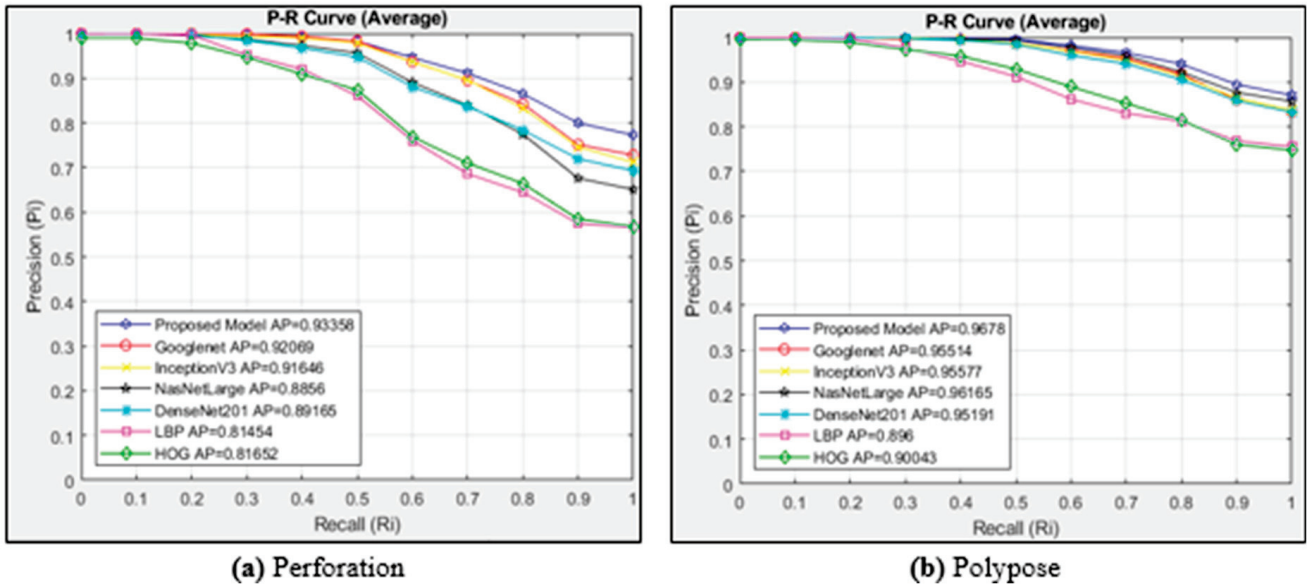


Figure 7. Average P-R curves for the classes of (**a**) perforation and (**b**) polypose.

The P-R curve for the adenomyomatosis class is shown in Figure 8a. The curves of six different architectures and the suggested model are plotted in this figure. A total of 23,280 images are viewed since the adenomyomatosis class contains 1164 images. The curves of six different architectures and the suggested model for the images in the carcinoma class are plotted in Figure 8b. The metric for measuring distance is the cosine technique. A total of 31,800 images are viewed since the carcinoma class contains 1590 images.

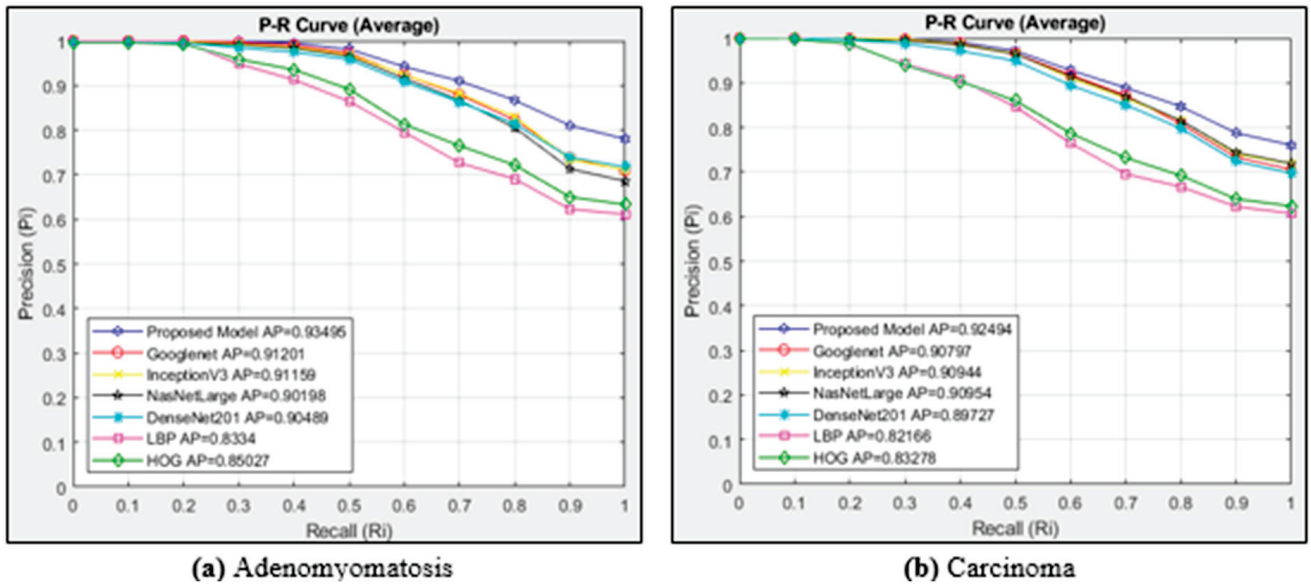


Figure 8. Average P-R curves for the classes of (**a**) adenomyomatosis and (**b**) carcinoma.

According to the P-R curve in Figure 8a, the most successful model among the models used for image retrieval in the adenomyomatosis class is the proposed model with an AP value of 0.93495. Googlenet and InceptionV3 followed the proposed model with very close AP values. The proposed model was more successful than the other models in the Ri 0–1 range. The least successful models in the adenomyomatosis class were textural-based models. LBP was the least successful model in this class, with an AP value of 0.8334. When Figure 8b is examined, among the models used for image access in the carcinoma

class, the most successful is the proposed model with an AP value of 0.92494. Googlenet, NasNetLarge, and InceptionV3 followed the proposed model with very close AP values. The proposed model was more successful than the other models in the Ri 0–1 range. These models performed in a close range at Ri 0–0.3. The least successful models in the carcinoma class were textural-based models. LBP was the least successful model in this class, with an AP value of 0.82166.

The P-R curves obtained using the cosine metric for the various class of six different architectures and the proposed model are presented in Figure 9a. A total of 19,800 images are viewed since the various classes contain 990 images. The P-R curves of all images in the nine classes in the dataset are shown in Figure 9b. This figure plots the curves of six different architectures and the proposed model. When the general performance curve is plotted, the cosine similarity measurement metric is used as the distance measurement metric. Since there are 10,692 images in the dataset, 213,840 images were accessed. In all steps, 20 images were accessed for each image.

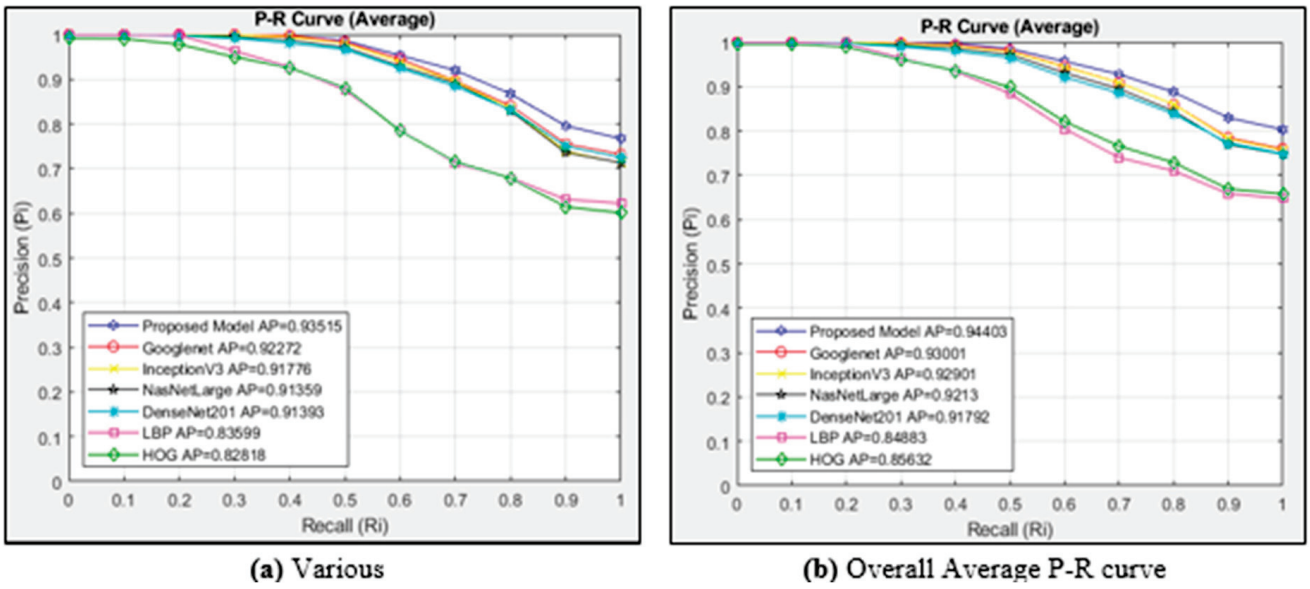


Figure 9. Average P-R curves for the classes of (**a**) various and (**b**) overall average P-R curves.

Figure 9a compares the models and the proposed model using the cosine similarity measurement method. Our suggested model performed the best in the CBIR system, with AP = 0.93515 in the various classes of images. With an AP of 0.82818, the HOG architecture performed the worst in the various class average. The suggested model outperforms the existing models in image access in every scenario within the recall {0..1} range, according to an analysis of the P-R graph. Compared to other structures, the HOG architecture exhibits very little success and is similar to the LBP architecture.

When the images in the classes gallstones, abdomen and retroperitoneum, cholecystitis, membranous and gangrenous cholecystitis, perforation, polyps and cholesterol crystals, adenomyomatosis, carcinoma, and various causes of gallbladder wall thickening were evaluated separately in our CBIR system, our proposed model achieved success in all classes. InceptionV3, NasNetLarge, Googlenet, DenseNet201, LBP, and HOG architectures were less successful than our proposed model in all classes.

As in all classes, our suggested model performs better in our suggested CBIR system than alternative models when the overall average of all classes is analyzed in Figure 9b. Our suggested model's InceptionV3, NasNetLarge, and Googlenet architectures perform

similarly when employed alone, but less successfully when it comes to image retrieval. In terms of overall success, the LBP architecture performed the worst.

In Figure 9b, our proposed model, which consists of the combination of Googlenet + InceptionV3 + NasNetLarge architectures, is more successful in similar image retrieval with the feature extraction method and the proposed CBIR system using the cosine similarity measurement metric, where AP = 0.94403. Class-based performance measurement metrics are presented in Table 2.

Table 2. Class-based performance metrics of the models (AP).

Diseases	Proposed Model	Googlenet	InceptionV3	NasNetLarge	DenseNet201	LBP	HOG
Gallstones	0.92711	0.90881	0.91291	0.89273	0.8911	0.82222	0.82701
Abdomen and Retroperitoneum	0.96338	0.95793	0.95117	0.94488	0.94013	0.90188	0.87181
Cholecystitis	0.92951	0.91379	0.91285	0.91363	0.90945	0.81275	0.84326
Membranous and Gangrenous Cholecystitis	0.98734	0.98029	0.98013	0.97418	0.97059	0.91153	0.94232
Perforation	0.93358	0.92069	0.91646	0.8856	0.89165	0.81454	0.81652
Polyps and Cholesterol Crystals	0.9678	0.95514	0.95577	0.96165	0.95191	0.896	0.90043
Adenomyomatosis	0.93495	0.91201	0.91159	0.90198	0.90489	0.8334	0.85027
Carcinoma	0.92494	0.90797	0.90944	0.90954	0.89727	0.82166	0.83278
Various Causes of Gallbladder Wall Thickening	0.93515	0.92272	0.91776	0.91359	0.91393	0.83599	0.82818
General Average	0.94403	0.93001	0.92901	0.9213	0.91792	0.84883	0.85632

4. Discussion

GB disease is a common pathology that requires accurate and early diagnosis for optimal medical treatment. The early diagnosis and appropriate treatment of gallbladder diseases are important for preventing complications. A healthy lifestyle, a balanced diet, and regular health check-ups can reduce the risk of developing these diseases. A delay in the diagnosis process or misdiagnosis can cause deterioration in the patient's condition. Today, the use of AI, ML, and DL techniques in predicting disease progression, identifying abnormalities, and estimating mortality rates associated with GB diseases has increased rapidly in the last decade [3,31]. In the proposed CBIR system, 10,692 US images obtained from gallbladder diseases were analyzed and compared with textural-based models LBP and HOG, CNN architectures DenseNet201, Googlenet, InceptionV3, and NasNetLarge models, and disease verification was achieved with an average of 94.4%. Some studies on the subject and the proposed CBIR-based model are presented in Table 3.

Due to the cost of the training process of CNN-based models, CBIR systems are very important in studies with large datasets and many classes. When Table 2 is examined, it can be seen that the proposed CBIR-based model produces successful results. It is also possible to mention some limitations of the study. One of the limitations of this study is the use of data obtained from a single center. Another limitation is that the proposed CBIR system has been tested on a single dataset. The model may need to be tested on other datasets to prove its effectiveness. In subsequent studies, it is possible to include data and experts from different centers and obtain more general and successful results.

Table 3. Literature review.

Paper	Year	Model/Method/ Architecture	Dataset	Images	Accuracy (%)
Dadjouy and Sajedi, 2024 [21]	2024	Faster R-CNN and YOLOv8	Gallbladder Cancer Ultrasound (GBCU)	The GBCU dataset includes 1255 ultrasound images from 218 patients	90.16%, 82.79%
Chattopadhyay et al., 2005 [23]	2005	Ultrasound scan (USS)	Polypoidal lesions in the gall bladder (PLG): ultrasound scan (USS)	751 cholecystectomies	Gallbladder abnormality (specificity) 92.3%
Esen et al., 2024 [22]	2024	RF, MLP, LR, NB, DT, KNN, GB, AdaBoost, and XGBoost	Bioimpedance and laboratory data	319 samples, 161 gallstone patients, and 158 healthy controls	85.42%
Yu et al., 2021 [20]	2021	SSD-FPN-ResNet-50 and MobileNet V2	89,000 abdominal US images taken from 2386 patients	Abdominal US images (>89,000)	0.92 and 0.94
Pang et al., 2019 [24]	2019	MobileNetV1, SSD, YOLOv2, and original SSD (with VGG-16)	CT images of 100 patients with cholelithiasis	A total of 1300 CT images of cholelithiasis	90.8%
Jang et al., 2021 [25]	2021	ResNet50	Endoscopic ultrasound (EUS):	1039 EUS images: polypoid lesions of the gallbladder (GB)	57.9%, 96.5%, 77.8%, 91.6%, 89.8%
Veena et al., 2022 [26]	2022	SSD-efficient-Det, Faster R-CNN, and Mask R-CNN models	computerized tomography (CT) images	60	Mask R-CNN has a value of 0.938
Song et al., 2019 [27]	2019	DC-GAN and CNN	Segmenting CT images of gallstones	The dataset includes 5350 images from 726 patients	91.68%
Proposed Model	2025	CBIR-based system	Gallbladder diseases	10.692 IU images	94.4%

5. Conclusions

This study proposed a CBIR-based classification model for diagnosing gallbladder diseases using UI datasets. By comparing feature extraction methods, including machine learning architectures like LBP and HOG, and CNN architectures, such as DenseNet201, Googlenet, InceptionV3, and NasNetLarge, the model demonstrated a superior performance with the cosine similarity metric. The evaluation included class-specific and overall dataset analyses using interpolated 11-point PR curves and AP calculations for performance measurements. The proposed model achieved the best performance across nine gallbladder disease classes and the entire dataset. Specifically, the cosine similarity-based CBIR system yielded notable AP scores, such as 0.92711 for gallstones, 0.96338 for abdomen and retroperitoneum, and 0.98734 for membranous and gangrenous cholecystitis. Overall, the model demonstrated an impressive AP of 0.94403 for the entire dataset, outperforming other architectures in retrieval accuracy. In conclusion, the proposed CBIR system with the cosine similarity effectively enhances image retrieval and classification for gallbladder diseases, showcasing its potential for clinical applications and advancing diagnostic methodologies.

Author Contributions: Conceptualization, M.Y., M.K., H.B.M., G.K. and A.A.; Methodology, A.B., M.Y., M.K., H.B.M., G.K. and A.A.; Software, M.K., H.B.M. and G.K.; Validation, A.B., M.Y., M.K., H.B.M., G.K. and A.A.; Formal analysis, A.B., M.K., H.B.M., G.K. and A.A.; Investigation, A.B., M.K., H.B.M., G.K. and A.A.; Resources, A.B., M.K., H.B.M., G.K. and A.A.; Data curation, A.B., M.Y., M.K., H.B.M., G.K. and A.A.; Writing—original draft, A.B., M.Y., M.K., H.B.M., G.K. and A.A.; Writing—review & editing, A.B., M.Y., M.K., H.B.M., G.K. and A.A.; Visualization, M.K., H.B.M. and G.K.; Supervision, A.B., M.Y., M.K., G.K. and A.A.; Project administration, A.B., M.Y., M.K., G.K. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A publicly available dataset was used in this article.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

References

1. Vendrami, C.L.; Magnetta, M.J.; Mittal, P.K.; Moreno, C.C.; Miller, F.H. Gallbladder carcinoma and its differential diagnosis at MRI: What radiologists should know. *Radiographics* **2021**, *41*, 78–95. [CrossRef] [PubMed]
2. Shabanzadeh, D.M.; Sørensen, L.T.; Jørgensen, T. Determinants for gallstone formation--a new data cohort study and a systematic review with meta-analysis. *Scand. J. Gastroenterol.* **2016**, *51*, 1239–1248. [CrossRef]
3. Obaid, A.M.; Turki, A.; Bellaaj, H.; Ksantini, M.; AlTaee, A.; Alaerjan, A. Detection of gallbladder disease types using deep learning: An informative medical method. *Diagnostics* **2023**, *13*, 1744. [CrossRef] [PubMed]
4. Wittenburg, H.; Lammert, F. Genetic predisposition to gallbladder stones. *Semin. Liver Dis.* **2007**, *44*, 109–121. [CrossRef] [PubMed]
5. Ahmed, A.S.; Ahmed, S.S.; Mohamed, S.; Salman, N.E.; Humidan, A.A.M.; Ibrahim, R.F.; Salim, R.S.; Elamir, A.A.M.; Hakim, E.M.; Humidan, A.A.M., Jr.; et al. Advancements in Cholelithiasis Diagnosis: A Systematic Review of Machine Learning Applications in Imaging Analysis. *Cureus* **2024**, *16*, e66453. [CrossRef]
6. Dvora, B. Gallbladder Cancer. 2024. Available online: https://tamc.co.il/en/article/gallbladder-cancer (accessed on 15 January 2025).
7. Lammert, F.; Miquel, J.-F. Gallstone disease: From genes to evidence-based therapy. *J. Hepatol.* **2008**, *48*, S124–S135. [CrossRef] [PubMed]
8. Shabanzadeh, D.M.; Skaaby, T.; Sørensen, L.T.; Jørgensen, T. Screen-detected gallstone disease and cardiovascular disease. *Eur. J. Epidemiol.* **2017**, *32*, 501–510. [CrossRef] [PubMed]
9. Zheng, Y.; Xu, M.; Heianza, Y.; Ma, W.; Wang, T.; Sun, D.; Albert, C.M.; Hu, F.B.; Rexrode, K.M.; Manson, J.E.; et al. Gallstone disease and increased risk of mortality: Two large prospective studies in US men and women. *J. Gastroenterol. Hepatol.* **2018**, *33*, 1925–1931. [CrossRef] [PubMed]
10. Wang, X.; Yu, W.; Jiang, G.; Li, H.; Li, S.; Xie, L.; Bai, X.; Cui, P.; Chen, Q.; Lou, Y.; et al. Global Epidemiology of Gallstones in the 21st Century: A systematic review and Meta-analysis. *Clin. Gastroenterol. Hepatol.* **2024**, *22*, 1586–1595. [CrossRef]
11. Burud, I.A.S.; Elhariri, S.; Eid, N. Gallbladder carcinoma in the era of artificial intelligence: Early diagnosis for better treatment. *World J. Gastrointest. Oncol.* **2025**, *17*, 99994. [CrossRef] [PubMed]
12. Lee, S.R.; Kim, H.O.; Shin, J.H. Reasonable cholecystectomy of gallbladder polyp--10 years of experience. *Asian J. Surg.* **2019**, *42*, 332–337. [CrossRef] [PubMed]
13. Li, B.; Hou, B.; Yu, W.; Lu, X.; Yang, C. Applications of artificial intelligence in intelligent manufacturing: A review. *Front. Inf. Technol. Electron. Eng.* **2017**, *18*, 86–96. [CrossRef]
14. Nia, N.G.; Kaplanoglu, E.; Nasab, A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discov. Artif. Intell.* **2023**, *3*, 5.
15. Barata, C.; Santiago, C. Improving the explainability of skin cancer diagnosis using CBIR. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 550–559.
16. Wickstrøm, K.K.; Østmo, E.A.; Radiya, K.; Mikalsen, K.Ø.; Kampffmeyer, M.C.; Jenssen, R. A clinically motivated self-supervised approach for content-based image retrieval of CT liver images. *Comput. Med. Imaging Graph.* **2023**, *107*, 102239. [CrossRef]

17. Lo, C.-M.; Wang, C.-C.; Hung, P.-H. Interactive content-based image retrieval with deep learning for CT abdominal organ recognition. *Phys. Med. Biol.* **2024**, *69*, 45004. [CrossRef]
18. Gupta, P.; Basu, S.; Rana, P.; Dutta, U.; Soundararajan, R.; Kalage, D.; Chhabra, M.; Singh, S.; Yadav, T.D.; Gupta, V.; et al. Deep-learning enabled ultrasound based detection of gallbladder cancer in northern India: A prospective diagnostic study. *Lancet Reg. Heal. Asia* **2024**, *24*, 100279. [CrossRef] [PubMed]
19. Zhou, W.; Ye, Z.; Huang, G.; Zhang, X.; Xu, M.; Liu, B.; Zhuang, B.; Tang, Z.; Wang, S.; Chen, D.; et al. Interpretable artificial intelligence-based app assists inexperienced radiologists in diagnosing biliary atresia from sonographic gallbladder images. *BMC Med.* **2024**, *22*, 29. [CrossRef]
20. Yu, C.-J.; Yeh, H.J.; Chang, C.C.; Tang, J.H.; Kao, W.Y.; Chen, W.C.; Huang, Y.J.; Li, C.H.; Chang, W.H.; Lin, Y.T.; et al. Lightweight deep neural networks for cholelithiasis and cholecystitis detection by point-of-care ultrasound. *Comput. Methods Programs Biomed.* **2021**, *211*, 106382. [CrossRef] [PubMed]
21. Dadjouy, S.; Sajedi, H. Gallbladder Cancer Detection in Ultrasound Images based on YOLO and Faster R-CNN. In Proceedings of the 2024 10th International Conference on Artificial Intelligence and Robotics (QICAR), Singapore, 15–17 November 2024; pp. 227–231.
22. Esen, I.; Arslan, H.; Esen, S.A.; Gül, M.; Kültekin, N.; Özdemir, O. Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data. *Medicine* **2024**, *103*, e37258. [CrossRef] [PubMed]
23. Chattopadhyay, D.; Lochan, R.; Balupuri, S.; Gopinath, B.R.; Wynne, K.S. Outcome of gall bladder polypoidal lesions detected by transabdominal ultrasound scanning: A nine year experience. *World J. Gastroenterol. WJG* **2005**, *11*, 2171. [CrossRef]
24. Pang, S.; Wang, S.; Rodríguez-Patón, A.; Li, P.; Wang, X. An artificial intelligent diagnostic system on mobile Android terminals for cholelithiasis by lightweight convolutional neural network. *PLoS ONE* **2019**, *14*, e0221720. [CrossRef]
25. Jang, S.I.; Kim, Y.J.; Kim, E.J.; Kang, H.; Shon, S.J.; Seol, Y.J.; Lee, D.K.; Kim, K.G.; Cho, J.H. Diagnostic performance of endoscopic ultrasound-artificial intelligence using deep learning analysis of gallbladder polypoid lesions. *J. Gastroenterol. Hepatol.* **2021**, *36*, 3548–3555. [CrossRef]
26. Veena, A.; Gowrishankar, S. Context based healthcare informatics system to detect gallstones using deep learning methods. *Int. J. Adv. Technol. Eng. Explor.* **2022**, *9*, 1661.
27. Song, T.; Meng, F.; Rodriguez-Paton, A.; Li, P.; Zheng, P.; Wang, X. U-next: A novel convolution neural network with an aggregation u-net architecture for gallstone segmentation in ct images. *IEEE Access* **2019**, *7*, 166823–166832. [CrossRef]
28. Turki, A.; Obaid, A.M.; Bellaaj, H.; Ksantini, M.; AlTaee, A. UIdataGB: Multi-Class ultrasound images dataset for gallbladder disease detection. *Data Br.* **2024**, *54*, 110426. [CrossRef] [PubMed]
29. Turki, A.; Mahdi Obaid, A.; Bellaaj, H.; Ksantini, M.; Altaee, A. Gallblader Diseases Dataset. *Mendeley Data* **2024**. [CrossRef]
30. Lahitani, A.R.; Permanasari, A.E.; Setiawan, N.A. Cosine similarity to determine similarity measure: Study case in online essay assessment. In Proceedings of the 2016 4th International conference on cyber and IT service management, Bandung, Indonesia, 26–27 April 2016; pp. 1–6.
31. Serban, D.; Badiu, D.C.; Davitoiu, D.; Tanasescu, C.; Tudosie, M.S.; Sabau, A.D.; Dascalu, A.M.; Tudor, C.; Balasescu, S.A.; Smarandache, C.G. Systematic review of the role of indocyanine green near-infrared fluorescence in safe laparoscopic cholecystectomy. *Exp. Ther. Med.* **2022**, *23*, 187. [CrossRef] [PubMed]

diagnostics

MDPI

Article

An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification

Madallah Alruwaili [1,*,†] and Mahmood Mohamed [2,†]

[1] Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Aljouf, Saudi Arabia
[2] Department of Information Systems and Technology, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza 12613, Egypt; mahmoodissr@cu.edu.eg
* Correspondence: madallah@ju.edu.sa
† These authors contributed equally to this work.

Abstract: Background: Medical diagnosis for skin diseases, including leukemia, early skin cancer, benign neoplasms, and alternative disorders, becomes difficult because of external variations among groups of patients. A research goal is to create a fusion-level deep learning model that improves stability and skin disease classification performance. **Methods:** The model design merges three convolutional neural networks (CNNs): EfficientNet-B0, EfficientNet-B2, and ResNet50, which operate independently under distinct branches. The neural network model uses its capability to extract detailed features from multiple strong architectures to reach accurate results along with tight classification precision. A fusion mechanism completes its operation by transmitting extracted features to dense and dropout layers for generalization and reduced dimensionality. Analyses for this research utilized the 27,153-image Kaggle Skin Diseases Image Dataset, which distributed testing materials into training (80%), validation (10%), and testing (10%) portions for ten skin disorder classes. **Results:** Evaluation of the proposed model revealed 99.14% accuracy together with excellent precision, recall, and F1-score metrics. **Conclusions:** The proposed deep learning approach demonstrates strong potential as a starting point for dermatological diagnosis automation since it shows promise for clinical use in skin disease classification.

Keywords: fusion-based deep learning; EfficientNet; ResNet; skin disease classification; multi-class classification; dermatological image analysis

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 17 January 2025
Revised: 22 February 2025
Accepted: 23 February 2025
Published: 25 February 2025

Citation: Alruwaili, M.; Mohamed, M. An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification. *Diagnostics* **2025**, *15*, 551. https://doi.org/10.3390/diagnostics15050551

1. Introduction

Skin diseases refer to a subgroup of diseases that affect the skin and thus are not only diverse but also pose serious problems in the management of health care for people of all ages and from all over the world [1]. Recent developments of deep learning techniques for medical image diagnosis apply enhanced chances to increase the diagnosis and classification strategies of dermatological diseases that can help to reform the clinical approaches of skin disorders for the better health and comfort of the patients [2]. Such skin disease features indicate a high demand for accurate and time-saving diagnostic methods, including more frequent pathologies, such as eczema and acne, and severe conditions, like psoriasis and cutaneous lymphoma [3]. However, there is still the problem of inequality in the availability of specialized dermatological services, which leads to several delayed diagnoses and ways of dealing with illnesses [4]. Timely diagnosis is an important aspect in cutting down on death and enhancing treatment plans, especially with the area of focus

being melanoma, which has very low survival rates if diagnosed late. However, referrals to dermatological specialists and dermatological technologies are frequently unavailable for patients living in rural or underdeveloped regions, and this is why effective automated diagnosis systems are called for [5]. Recent improvements in deep learning, especially convolutional neural networks (CNNs), have been successfully applied in medical image diagnosis and detection with high automation and accuracy. Currently, there are several categories of state-of-the-art pre-trained CNNs, including EfficientNet [6], ResNet [7], and DenseNet [8], among others, which have proven to provide optimum results when applied to skin lesion classification. Most of these models use transfer learning to learn features that are then used in accurate predictions for medical images without the need for large datasets that are time-consuming to annotate. However, discriminative intra-class variability and inter-class similarities may be an issue for a single CNN in dermatological imaging. Several studies have attempted to solve this problem of skin disease classification with deep learning methods. For instance, Esteva et al. [9] re-tuned an InceptionV3 network and obtained dermatologist-level performance with their study on a large dataset of skin images. Likewise, Tschandl et al. [10] trained several CNNs to diagnose several skin diseases, and they appear to be effective. Nevertheless, these approaches mainly addressed the problem of a single model or homogeneous ensemble, and this can cause some drawbacks when handling imbalanced or noisy databases. Such challenges can easily be solved using a fusion-based method that merges the strengths of the various models. Meštrović et al. [11] investigated the mutual relationship of urinary tract infection development between the gut and vaginal and urinary microbiomes to develop new risk assessment tools and therapeutic approaches. This research examines microbial disruptions and their consequences for avoiding UTIs and treating existing infections. Mujahid et al. [12] investigated the detection of pneumonia in X-ray images through deep learning models, which include Inception-V3, VGG16, and ResNet50. According to experimental findings, the combination of Inception-V3 with the CNN ensemble reached 99.73% in recall and 99.29% in accuracy, which shows promising potential for pneumonia diagnosis. Bordin et al. [13] evaluated diagnostic procedures to detect Helicobacter pylori by dividing them into invasive and non-invasive methods. The combination of endoscopic imaging progress with artificial intelligence boosts real-time diagnosis as the urea breath test alongside the stool antigen test functions as the main diagnostic method for Helicobacter pylori detection. Antibiotic resistance assessment requires the combination of molecular methods together with gastric biopsy cultures. Thomas et al. [14] examined how bacterial populations in the mouth affect both oral infections and chronic systemic health conditions, including caries and periodontal disorders, as well as systemic problems. The authors introduced innovative diagnostic instruments with therapeutic methods based on bacterial recolonization and host modulation therapy, which supports the importance of combined oral–general health professional cooperation.

One of the challenges encountered in this research is class imbalance, especially in skin disease datasets, where some classes have much less data than others, e.g., melanoma is much less represented than benign conditions [15]. Due to the skewed nature, these traditional deep learning models are trained in a way to favor major classes, and hence the performance metrics are off on the minority of classes. As discussed above, the latest studies [16] tried to address this problem with data augmentation and class-weighted loss functions. However, the problem of achieving a balanced overall classification performance has not yet been solved, which requires more effective approaches. Müller et al. [17] have presented an integration technique, including stacking and augmenting, that has been shown to hold great promise in improving the medical image classification systems' reliability and precision. Li et al. [18] introduced a multimodal medical image fusion approach

combining CNN and supervised learning to address limitations in traditional single-image fusion methods. The proposed method enhances fusion quality, detail clarity, and efficiency, demonstrating state-of-the-art performance across various evaluation metrics. Although these fusion techniques appear to achieve higher accuracy, the regularization methods like dropout and batch normalization for improving the generalization performance were not very well incorporated.

In this research, we put forward a fusion-based deep learning model of skin disease classification based on the combination of EfficientNet-B2, EfficientNet-B0, and ResNet50 architectures. These pre-trained models are then employed to extract static and highly informative features of skin disease images. The extracted features are normalized using batch normalization to improve the learning process. Combining these features, the model receives different feature representations of the features, which solves the problem of variations within one class and similarities between different classes. The proposed model includes several dense and dropout layers to better visualize the feature maps and cut down on overfitting. Dropout, on the other hand, helps prevent the model from over-relying on some of the features to improve the overall generalization of the model. These strategies work in conjunction with our proposed model to handle imbalanced datasets with better accuracy in minority classes. This kind of fusion-based approach serves as the solution to all the above-mentioned previous studies, which appeared to use only a single piece of architecture or, at best, did not use complex regularization methods.

Thus, our work contributes to the area by bridging certain known gaps in skin disease classification. The proposed fusion model is superior to a single-model-based approach in that it capitalizes on more than one pre-trained architecture and their complementary qualities to improve outcomes. Furthermore, we also use batch normalization and dropout techniques to reduce overfitting and handle class imbalance problems. A detailed experimental evaluation validates the proposed model on a standard skin disease dataset used in previous studies and shows that it outperforms existing techniques in terms of accuracy. The key contributions of this article are as follows:

1. Propose a fusion-based deep learning model combining EfficientNet-B2, EfficientNet-B0, and ResNet50 for accurate skin disease classification.
2. A feature fusion strategy with batch normalization and dropout layers is introduced to improve generalization and robustness.
3. The proposed model effectively addresses class imbalance and enhances minority class predictions.
4. Comprehensive evaluations on a benchmark skin disease dataset validate the model's superiority over existing approaches.

The rest of the paper, Section 2, describes the literature review of skin disease classification; after that, the material and methodology are described, and the types of methods used in our study are presented in Section 3. Then, in Section 4, the results and discussions of our work are presented. In addition, Section 5 contains the conclusion of the study.

2. Literature Review

Skin disease classification and deep learning approaches have been studied in recent years with high efficiency. Significant to deep learning in dermatology, Esteva et al. [9] used the InceptionV3 architecture to do so. Even though the single-CNN model's results were as accurate as those of a dermatologist, this design choice may not hold up well in the real world. Likewise, Tschandl et al. [10] applied CNN ensembles, but some issues that turned up included computational overhead and overtraining. In [19], the authors extend their work on deploying advanced deep learning approaches: Vision Transformer (ViT) and convolutional neural networks (CNNs) to detect melanoma. With increased rates of skin

cancer and detection of melanoma essential to increase patient survival, the study assesses the performance of these models using a dermoscopic image dataset. The findings also reveal that the pre-trained Vision Transformers yielded an impressive diagnostic accuracy of 97.97%.

Venkata et al. [20] proposed a wrapper feature selection approach based on the dragonfly algorithm (DFA) to select the best feature subsets for skin cancer classification. The researchers combined DFA, where the population of dragonflies has been classified according to their fitness values and brought about the required modification to the positions of dragonflies. In this work, two CNN models were trained using EfficientNet-B2 and VGG19 models and applied data augmentation to both labeled and unlabeled datasets. The framework EfficientNet-B2 yielded excellent performance, specifically in obtaining an overall accuracy of approximately 89.55% on the test set for the classification of some types of skin diseases. It has the potential for use in other dermatological conditions and for other tasks such as psoriasis area and severity index (PASI) scoring, for which this study offers important insights for severity measures and the development of skin disease diagnosis and classification. The authors in [21] focused on deep learning for image analysis on skin disease diagnosis. Therefore, the idea of using skin disease identification brings the consideration of personalization into a potentially more workable reality, with disparate data sources and adherence to ethical AI practices. The authors in [22] described multiple skin lesion classification from dermoscopic images using a shallow model Transformer module (TM) and self-attention unit (SAU) integrated with an efficient CNN model. The proposed lightweight architecture suggests that CNN-based feature generation simplification is possible; simultaneously, the degree of accuracy preservation with dermoscopy images (DIs) and complex textures is possible. Strong performance on the ISIC-2019 and PH2 datasets demonstrates the cross-dataset transferability of the approach via experiments with large sets of other unseen DIs. In the current work, cross-fusion approaches are used to handle dermoscopy images. Also, in [23], the authors presented a new skin disease classification model that utilizes leading-edge deep learning as the key methodology to enhance performance. Based on the MobileNet-V2 architecture, it introduces squeeze-and-excitation networks, atrous spatial pyramid pooling, and the channel attention mechanism. Trained on diverse datasets, including PH2, Skin Cancer MNIST: HAM10000, DermNet, and Skin Cancer ISIC are used as datasets; resizing and mean subtraction normalization are applied. Here, the MobileNet-V2 backbone extracts hierarchical features, whereas ASPP combines multi-scale contextual information to produce a feature map. A study on self-attention-based mechanisms promotes the extraction of inter-channel relationships and contextual information, positively affecting feature discrimination. The model yields an accuracy of 98.6%. Moreover, in [24], the author proposed ResNet50-LSTM, a combination of the ResNet50 deep model and the LSTM classification model, to overcome the shortcomings of existing models. The deep networks are cascaded with the transfer learning technique, which handles the sequential data and captures the structural content of the lesion textures. The findings showed that the third case of ResNet50-LSTM-TL had stunning accuracy, higher than 99.09%, making it even more potent than other deep learning models in detecting different skin cancers.

Although the work done in recent years has enhanced the accuracy and effectiveness of deep learning technology in classifying skin diseases, there are still some drawbacks. Most research considers binary or few categories of classification and so does not offer broad solutions for classifying skin diseases concurrently. For instance, current models that are presented in the literature fail to predict all ten standard labels of skin diseases at the same time; therefore, they cannot be implemented in real-life clinical practice. The proposed model fulfills the need for better-automated skin disease classification since

InceptionNet-v3 and other existing deep-learning models prove insufficient. Traditional approaches fall short of accurate diagnosis when used for skin condition identification since various skin diseases share visual patterns that produce diagnostic errors through misclassification errors. A proposed model improves classification results through state-of-the-art feature extraction processes that minimize misclassification errors while increasing general performance. An optimized deep learning architecture powerfully differentiates different dermatological conditions, which makes the model beneficial for clinical decision support applications alongside telemedicine systems.

3. Materials and Methods

In this section, we describe the models used in our proposed model, such as ResNet50, EfficientNetB0, and B2, and the rationale behind their selection and fusion. ResNet-50 demonstrated choice as a feature extractor because its deep residual connections address gradient vanishing problems, thus resulting in effective feature transmission. The combination of EfficientNet variants B0 and B2 needed integration because they demonstrated high parameter efficiency together with a balanced trade-off between network depth and width while running fewer computations. The proposed fusion model selects strengths from multiple network architectures that will use different feature structures to boost classification results. CancerNet is expected to deliver improved overall performance by uniting standard models because it improves both generalization capabilities and misclassification resistance when classifying visually similar skin diseases.

3.1. ResNet50

There is a CNN model called ResNet50, which stands for Residual Network-50, aimed at resolving difficulties with training very deep neural networks, mostly the vanishing gradient drawbacks. As proposed by He et al. [7], ResNet50, as shown in Figure 1, has 50 layers and is based on what is known as residual learning with shortcut connections. These shortcut connections allow the network to jump over one or more of them and perform an identity function, thereby making learning much easier. Unlike MobileNet, ResNet50 learns not direct mappings but residual functions, which should make the backpropagation step smoother. This enables the establishment of deeper models as far as the innovation helps in training the same without compromising on its performance, which is much different from the original architectures like VGG16 and AlexNet.

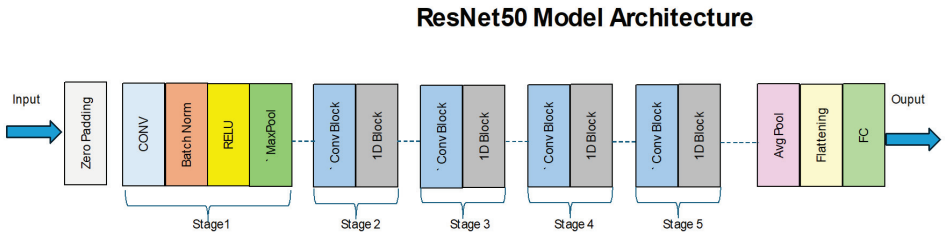


Figure 1. ResNet50 model architecture.

Figure 1 depicts ResNet-50's architectural design, adopting deep residual learning to extract features with added protection against gradient vanishing. The initial architecture starts with an input layer, and it completes zero padding for spatial dimension protection. The first stage contains Conv layers for feature extraction, after which Batch Norm helps stabilize training before the model applies ReLU and Max Pool for down-sampled feature maps. The architecture includes stages 2 through 5 with convolution combined with identity blocks, which apply learnable filters within convolutional blocks but let gradient flow pass through shortcut connections. The framework implements an average pooling (average

pool) together with a flattening layer to complete its operation. The average pooling functionality helps decrease dimensions before the final layer transforms feature maps into one-dimensional vectors. The output predictions emerge from the last fully connected layer, which concludes the classification process. The systematic design of ResNet-50 enables both performance in organizing hierarchical models and operational efficiency.

ResNet50 has residual blocks where 1×1, 3×3, and again 1×1 layers are used for the reduction of dimensionality, feature extraction, and expansion, respectively. In the same manner, ResNet50 decreases computational costs and Cal energies via bottleneck designs without compromising the network's ability to contain sufficient representational capacities. Because of this learning of hierarchical features, CNNs have been adopted in image classification chores, performing impressively well on benchmark image datasets, such as ImageNet. Besides the classification context, ResNet50 has also been used in many other computer vision tasks, such as object detection, image segmentation, and medical image analysis [25,26]. Due to its modularity and efficiency, the authors mentioned above noted that it has become the base for many other derived architectures, primarily ResNet and DenseNet, due to its residual learning capability [8]. Nonetheless, ResNet50 has its drawbacks: Although residual learning helps to reduce gradient problems, the depth of the network poses a problem of computational and memory overhead, especially in environments where such resources are scarce. Fine-tuning deep networks such as ResNet50 may be very sensitive and compelled by optimization considerations and hyperparameters' settings. However, on balance, ResNet50 is the most important accomplishment in deep learning because it learns deep networks well and generalizes well in many contexts. It has paved the way for the improvement of subsequent deep learning frameworks in computer vision, which marks the study as a strong foundation in modern computer vision [7,8].

3.2. Efficient Net

Tan and Le [6] proposed a novel CNN called EfficientNet, improving both efficiency and accuracy through a proper scaling of depth, width, and resolution of the input data. As opposed to previous works, EfficientNet applies a compound scaling factor to these dimensions so that all are positively scaled while at the same time maximizing computational complexity. With EfficientNet-B0 as the baseline model, launched from neural architecture search (NAS), which maximizes the architecture density while providing extremely high efficiency. Even though it has merely 5.3 M parameters, EfficientNet-B0, as shown in Figure 2, provides high results in image classification tasks, is energy efficient, and even surpasses heavy-hitter models, such as ResNet50 on the ImageNet dataset. It comprises mobile inverted bottleneck convolution (MBConv) blocks, depth-wise convolutions, and SWISH activation functions, all of which enable the model to pick fine-grained feature representation with much less computational cost.

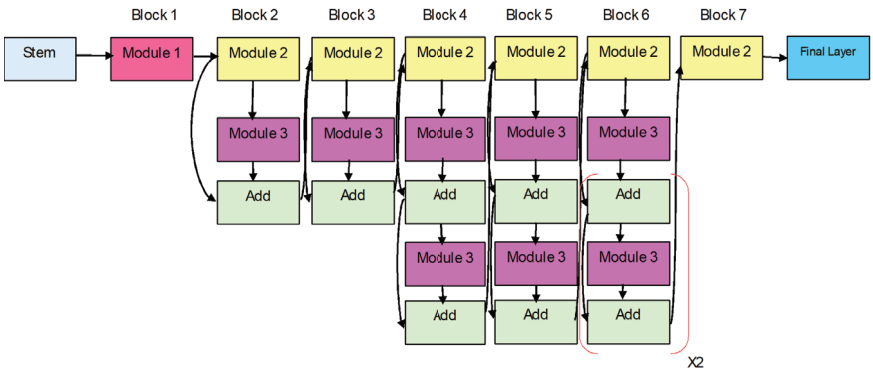


Figure 2. EfficientNet-B0 Model Architecture.

The EfficientNetB0 design shows the network architecture used for efficient and scalable image classification duties, as shown in Figure 2. As the base model of the EfficientNet family, EfficientNetB0 applies a compound scaling methodology for equal scaling of network depth, width, and resolution. The model design permits both high-accuracy performance and efficient operation. Modularity refers to the network design, which exhibits multiple numbered components extending from Module 1 through Module 584. The modular approach in the network design represents individual network elements, such as convolutional layers or activation functions, as well as pooling layers. The systematic and hierarchical EfficientNetB0 architecture achieves optimal performance–resource utilization through its structure/modules numbered from 1 to 584.

EfficientNet-B2 is developed from the base EfficientNet-B0 model using compound scaling, which scales network depth, width, and input resolution in a proportional manner. More precisely, B2 enlarges the input image resolution to 260 × 260, enlarges the number of channels in the convolutional block, and increases the depth of the model in comparison with B0. Such enhancements enable EfficientNet-B2 to better encode finer-scale spatial patterns yet be efficient. EfficientNet-B2 has fewer parameters than B0, with approximately 9.1 million; however, it has slightly better performance in terms of accuracy, so it can be preferred for tasks that require finer features, such as medical image classification and fine-grained classification [27]. Nevertheless, B2 enlarges its scale while remaining computationally efficient, thanks to its efficient layout design and the selection of dimension scaling.

Figure 3 depicts the EfficientNet-B2 architecture, which functions as a convolutional neural network platform optimized for scalable image classification. A modular design structure is evident through the large number of numbered modules ranging from Module 1 up to Module 586. Every separate module matches a distinct part of the network, including convolutional modules and activation functions alongside pooling blocks. The sequential arrangement of these modules explains how EfficientNetB2 follows a systematic hierarchy that ensures optimal performance alongside resource efficiency in deep learning tasks. A modification of EfficientNetB0 implements expanded scaling parameters to achieve better precision and resistance in this model.

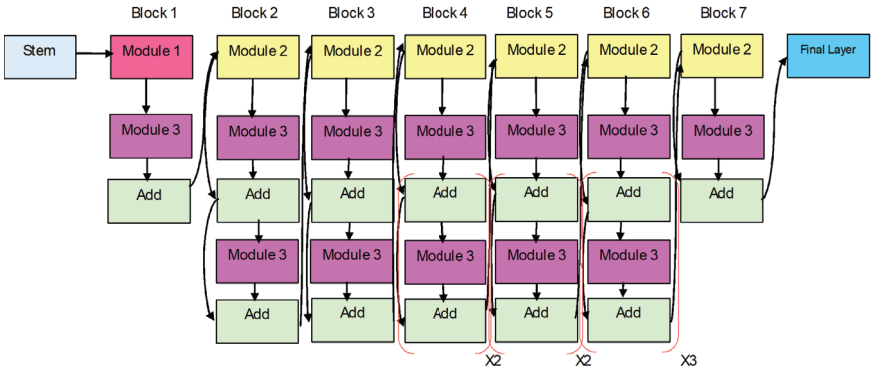


Figure 3. EfficientNet-B2 model architecture.

Recent members of the EfficientNet family, namely B0 and B2, have outperformed their counterparts, leveling high scores within numerous computer vision tasks, notably ImageNet. Due to their highly accurate results with low computational complexities, they are used in a wide range of applications, from mobile applications to large-scale image recognition [28]. However, due to compound scaling, the growth of model performance is limited to the baseline architecture, and any additional increase must be obtained through computationally intensive NASs. Still, EfficientNet-B0 and B2 provide insight into ef-

fective model scaling and have succeeded in efficient deep learning solutions in various domains [6,29].

In the proposed model, we select the EfficientNetB0 and EfficientNetB2 models because they achieve a perfect trade-off between resource efficiency and performance, which makes them ideal for various applications that work within limited computing capacity. The two EfficientNet versions establish a foundation through EfficientNetB0, which provides the compact architectural design with compound scaling as well as the enhanced accuracy capabilities of EfficientNetB2 without additional computational complexity. The orientation stems from requirements to use accuracy-strong and deployment-friendly models for practical use, yet higher variants B7 and B8 demonstrate insufficient performance despite requiring more resources. The implementation of B0 and B2 allows for an applicable solution that provides both versatility and computational efficiency for different operational domains.

3.3. Proposed Fusion Model

The proposed fusion model for skin disease classification illustrated in Figure 4 operates as an architecture for automated skin disease classification. The proposed approach employs three convolutional neural networks named EfficientNet-B0, EfficientNet-B2, and ResNet50 that extract multiple feature patterns from medical images. Each CNN performs separate analyses on the input images while building deep hierarchical features until reaching a holistic representation of the features. The fused features move through fully connected layers until the SoftMax layer makes a final classification among the ten skin disease categories.

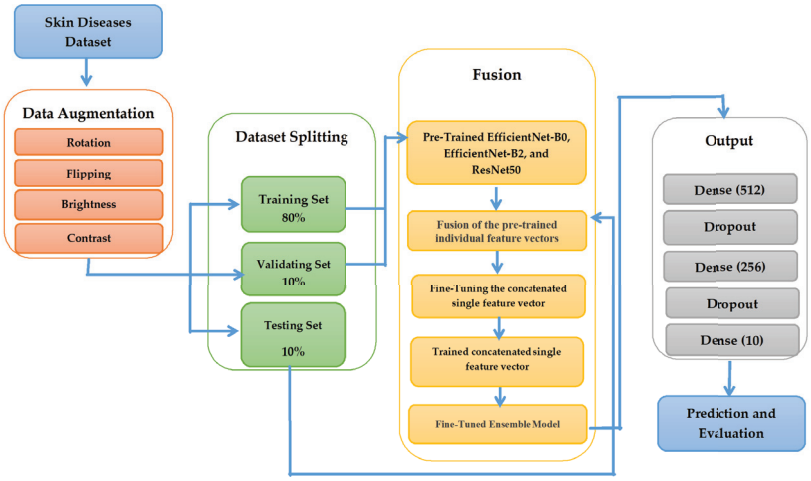


Figure 4. Proposed fusion model.

3.3.1. Data Preprocessing

The data processing stage starts by applying preprocessing and augmentation techniques to the skin disease dataset before continuing to the training process. Data augmentation adopts techniques such as rotation, flipping, brightness, and contrast adjustment. Additionally, rotation serves to provide real-world simulation effects alongside flipping mechanisms that perform horizontal and vertical transformations for increasing data variety. Adjustments in brightness levels and contrast values enable the model to understand different levels of illumination conditions more effectively. The group of augmentation techniques works together to improve how the model detects skin diseases across different setting conditions. Data augmentation enables the subdivision of the dataset into three parts, where 80% serves for training purposes and 10% functions for validation, while the remaining 10% serves for testing purposes. The model optimizes its parameters during

training with the data from the training set, and the validation set confirms hyperparameter adjustments and stops overfitting scenarios. The testing set provides the final model evaluation, so performance metrics exactly gauge actual real-world generalization performance.

3.3.2. Feature Fusion and Classification

The main component of the model combines three pre-trained powerful deep learning representatives known as EfficientNet-B0, EfficientNet-B2, and ResNet50 for feature extraction through fusion. The input images undergo high-level feature representation processing by each model to detect various elements of skin disease patterns. The classification procedure combines model-generated feature vectors through concatenation to attain an enriched feature set that receives additional optimization for classification purposes. Using this feature fusion technique enhances the compilation of robust features compared to standalone model usage. The classification process begins after the fused feature vector goes through a dense layer sequence that contains numerous fully connected blocks. The final classification head contains three linearly dense layers with 512, then 256, and finally 10 that use SoftMax activation. Between each dense layer exists a Dropout layer. During training, the dropout layers work as a regularization tool that creates better generalization through random neuron deactivation. Finally, the output layer performs categorical classification aimed at predicting the nature of skin disease. The SoftMax activation function gives probability scores for each class, and then the model can predict the most probable label of the input images. The types of skin diseases to be classified in the proposed fusion model include eczema, melanoma, atopic dermatitis, basal cell carcinoma (BCC), melanocytic nevi (NV), and benign keratosis-like lesions (BKL).

3.3.3. Evaluation Metrics

Multiple evaluation metrics determine the effectiveness assessment of the proposed model. The main measure of classification accuracy consists of accuracy. The evaluation of model performance includes calculations of precision, recall, and F1-score to identify how well the model distinguishes between various skin disease classes while maintaining sensitivity and specificity. Assessment through the area under the receiver operating characteristic (AUC-ROC) curve analysis allows measurement of model discrimination power between classes across different classification threshold points. The proposed model reaches better classification accuracy together with improved generalization ability through its integration of multiple pre-trained CNNs and optimized feature fusion techniques. The model's reproducibility stems from structured preprocessing procedures along with available scripts and systematic testing methods that allow real-world skin disease diagnosis applications.

4. Experimental Results and Discussion

The section presents an in-depth description of the dataset while also providing performance mathematical equations followed by experimental results and a state-of-the-art comparison discussion.

4.1. Dataset Description

The Skin Diseases Image Dataset is available on Kaggle [30] and provides a vast collection of original, visually identified skin images intended for a range of dermatological disease classifications. The dataset is 5.58 GB large. It has 10 different labels containing different skin diseases and 27,153 images of clinical images that demonstrate variations in skin color and patterns and range from mild forms of skin diseases to severe. It also enables reliable model formulation and evaluation by catering to the diverse characteristics of the dataset. The proposed model was accessed through TensorFlow and Keras tools on

an NVIDIA GPU device that provided quick model training plus evaluation processes. An Adam optimizer initiated at a 0.0001 learning rate powered all operations and hidden layers, using ReLU activation before the Softmax operation generated class probabilities. The dataset was divided into training (80%), validation (10%), and testing (10%) splits, which remained constant for all models during the comparison process. To establish the accuracy of our findings, we conducted 5-fold cross-validation, which strengthened performance assessment while reducing potential biases in the dataset.

The distribution of skin conditions contains 10 distinct categories that show their occurrence numbers in the presented Table 1. The research reveals that 1677 persons suffer from eczema, and 3140 individuals experience melanoma. The record shows that atopic dermatitis affects 1257 cases, whereas basal cell carcinoma (BCC) has 3323 cases. The dermatological condition classification has 7970 cases of melanocytic nevi (NV) as the most common and 2079 cases of benign keratosis-like lesion (BKL) as the second most common. The reports for psoriasis and lichen planus, along with related conditions, represent 2055 cases and seborrheic keratoses with other benign tumors represent 1847 cases. Of the recorded skin conditions, Tinea ringworm, Candidiasis, and other fungal infections total 1702 cases, while warts, molluscum, and other viral infections amount to 2103 cases. The dataset demonstrates how different skin-related diseases occur throughout their specified areas.

Table 1. Distribution of skin disease classes with their respective counts.

Class Name	Count
Eczema	1677
Melanoma	3140
Atopic dermatitis	1257
Basal cell carcinoma (BCC)	3323
Melanocytic nevi (NV)	7970
Benign keratosis-like lesion (BKL)	2079
Psoriasis, lichen planus, and related diseases	2055
Seborrheic keratoses and other benign tumors	1847
Tinea, ringworm, candidiasis, and other fungal infections	1702
Warts, molluscum, and other viral infections	2103

The analysis uses the hyperparameters in Table 2 to attain peak training outcomes. Because of its effectiveness on big datasets and generalization capabilities, the stochastic gradient descent (SGD) optimizer was the choice. The learning rate was set to 0.0001 to achieve gradual convergence and reduce the possibility of seeking past the ideal solution. The models were trained through the suitable multi-class classification loss function known as categorical cross-entropy. The model performance evaluation used accuracy as the main metric. The model used early stopping through validation loss monitoring to prevent overfitting using a patience threshold set to 20 epochs. The restore best weights capability was enabled to save the model parameters that exhibited peak performance for ultimate assessment.

Several approaches were used to handle the class imbalance problem in the image dataset by the proposed system. The minority classes benefited from data augmentation procedures that produced synthetic data by transforming real images with rotation effects and flipping operations in addition to brightness and contrast transformations. The combination of EfficientNet-B0, EfficientNet-B2, and ResNet50 allowed the model to extract comprehensive features from skin disease images, thereby reducing reliance on the majority class data. A weighting mechanism was included in the model's loss function structure to prioritize the training of minority classes, therefore achieving balanced performance across

all diagnosis categories. By applying these combined strategies, the model gained effective control over the class imbalance, which resulted in strong classification outcomes.

Table 2. Hyperparameters for all models.

Hyperparameter	Value
Optimizer	Stochastic Gradient Descent (SGD)
Learning Rate	0.0001
Loss Function	Categorical Cross entropy
Metrics	Accuracy
Monitor	Validation Loss
Patience	20 epochs
Restore Best Weights	TRUE

4.2. Performance Metrics

The evaluation of each classifier depended on accuracy levels while also using classification reports alongside ROC curves. The accuracy evaluation relied on dividing correctly classified instances by the total instances present. The evaluation process provided initial precision, recall (detection rate), and F1 scores for each test set class in classification reports. The comparison of model effectiveness through different classification thresholds employed ROC curves along with their area under the curve (AUC) evaluation.

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

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

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-measure} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

The model performs assessment with four categories, including true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs).

4.3. Results

Here, we report the comparative results of the deep learning models alongside the proposed model, such as EfficientNet-B2, ResNet101V2, MobileNet-v3, and InceptionNet-v3. To evaluate how well these models are for dermatological disease classification, they were examined using the Skin Diseases Image Dataset. Their performances were evaluated with the help of accuracy, precision, recall, and F1-score. The experiments for pre-trained model comparison depended on standardized execution procedures. The comparison involved maintaining consistent dataset partitioning methods along with preprocessing requirements while using the stochastic gradient descent optimizer and identical learning rates and batch sizes across all experimented models. The training duration changed according to early stopping, which used validation loss to stop training before overfitting occurred. The training duration for EfficientNet-B2 models needed only 70+ epochs for convergence, yet ResNet101V2 required 20+ epochs to achieve maximum performance, MobileNet-v3 required 60 epochs for convergence, InceptionNet-v3 required 100 epochs for convergence, and the final proposed model required 20 epochs for convergence. We explicitly added in the manuscript that all models achieved convergence under uniform experimental conditions regardless of their various epoch numbers.

Table 3 shows the classification report of testing the Skin Diseases Image Dataset for five models, including EfficientNet-B2, ResNet101V2, MobileNet-V3, InceptionNet-V3, and the proposed model. The execution time of 100 seconds represents the fastest performance of MobileNet-V3 along with a fair precision and recall score of 0.85 that also produces comparable F1-score and accuracy levels. The computational time was 170 seconds for ResNet101V2 while it delivered a performance of 0.80 across metrics. The InceptionNet-V3 model delivered moderate accuracy for skin disease diagnosis through 150 seconds of execution time, together with precision at 0.82 and F1-score and accuracy at 0.81. The proposed model delivered superior performance compared to all other models tested by reaching 99.14% across all assessment indicators within 200 seconds of execution time. Its diagnostic capabilities for identifying 10 skin disease conditions make the proposed model a powerful diagnostic instrument even though it requires slightly more computer processing time.

Table 3. Classification report of the testing dataset.

Model	Precision	Recall	F1-Score	Accuracy	Execution Time in Second
EfficientNet-B2	0.84	0.84	0.84	0.84	140
ResNet101V2	0.80	0.80	0.80	0.80	170
MobileNet-v3	0.85	0.85	0.85	0.85	100
InceptionNet-v3	0.82	0.81	0.81	0.81	150
Proposed Model	0.9914	0.9914	0.9914	0.9914	200

Figure 5 shows the performance evaluation through confusion matrices of four deep learning models, including ResNet101V2, MobileNet-V3, and InceptionNet-V3, and the proposed model appears in the figure based on their assessment in a ten-class classification scenario. The classification performance becomes visible in each confusion matrix through its presentation of properly identified and wrongly identified sample counts for different classes. Efficient Net-B2 classification accuracy reaches 84% but enables too many misdiagnoses between disparate categories across all non-diagonal elements. ResNet101V2 confirms moderate diagnostic errors between related dermatological conditions when achieving an 80% overall success rate. The classification capability of MobileNet-V3 reaches 85% accuracy because it shows more correct predictions along its main diagonal. InceptionNet-V3, with an accuracy of 81%, exhibits a slightly higher misclassification rate than MobileNet-V3 but outperforms ResNet101V2. The proposed model delivers exceptional performance by exceeding all baseline models through its 99.14% accuracy rating. The proposed model achieves nearly perfect precision, recall, and F1 score for all classes based on its strong diagonal values and low number of misclassified predictions found in its confusion matrix. The proposed model provides an effective solution for differentiating between dermatological conditions because its performance steadily improves in comparison to other models evaluated.

Figure 6 displays the accuracy changes of five deep learning models, including EfficientNet-B2, ResNet101V2, MobileNet-V3, InceptionNet-V3, and the proposed model, during their training and validation cycles across different epochs. The training accuracy of EfficientNet-B2 steadily rises to indicate 100% success, whereas its validation accuracy stabilizes at 85%, which implies weak model overfitting potential. ResNet101V2 demonstrates solid training accuracy along with validation accuracy that stops below 80%, potentially because of model generalization difficulties. MobileNet-V3 demonstrates fast learning, which results in nearly 85% validation accuracy, as well as stable training–validation accuracy differences. Training accuracy from InceptionNet-V3 achieved near-perfect results, but validation accuracy stagnated at 81%, which indicates steps toward overfitting occurred. A

high validation accuracy emerges from the proposed model, which demonstrates minimal fluctuations and suggests better generalization potential. The proposed model demonstrates excellent stability between its training accuracy and validation accuracy, which suggests more robustness together with lower overfitting than the other competing models. The research findings demonstrate that the proposed model outperforms alternative models because it shows better results in complex pattern detection combined with strong generalization capabilities.

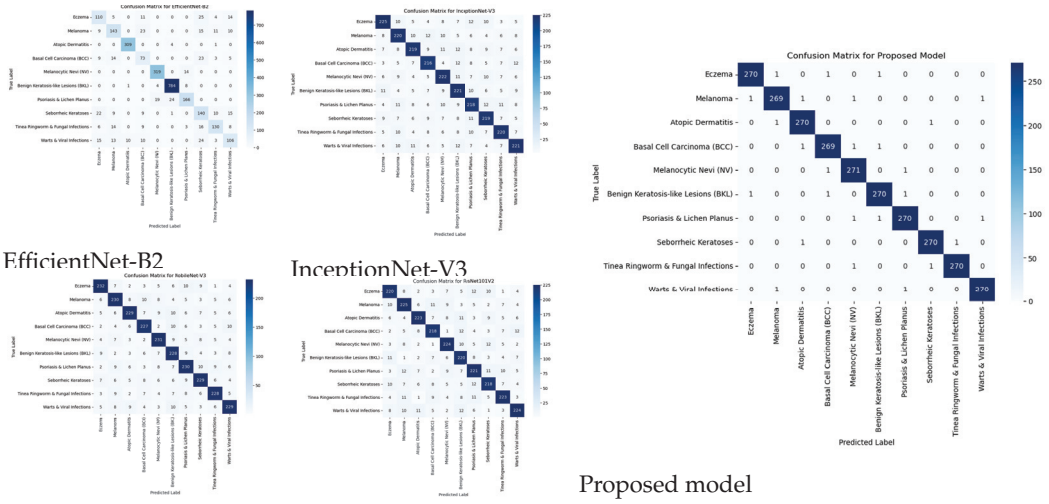


Figure 5. Confusion matrices.

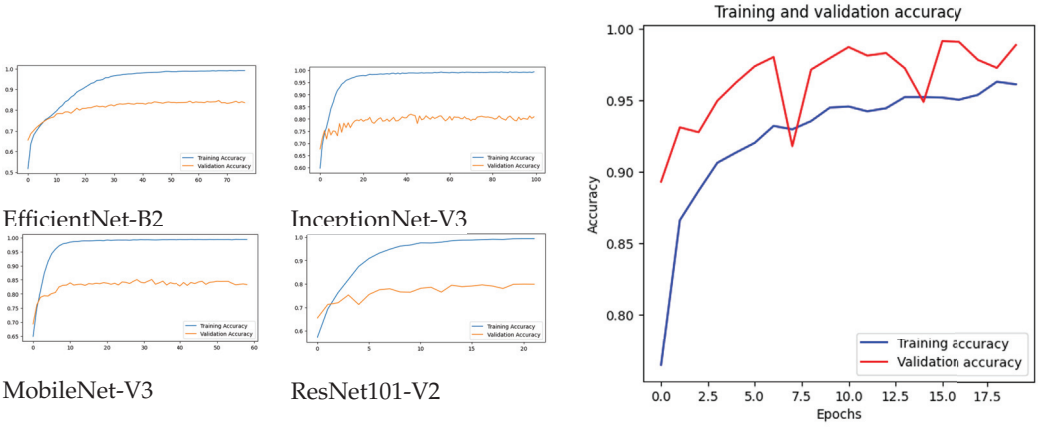


Figure 6. Accuracies for all models.

Figure 7 shows a comparison of convergence and generalization capacity exists between the training and validation loss curves of five models, including EfficientNet-B2 and ResNet101V2, along with MobileNet-V3, InceptionNet-V3, and the proposed model in a single depiction. The EfficientNet-B2 model demonstrates quick loss decreases during training and displays steady validation loss at a slightly higher level than training loss because of limited model overfitting. ResNet101V2 demonstrates training and validation loss patterns that show a significant gap that signals potential overfitting of the model. The convergence speed of MobileNet-V3 and InceptionNet-V3 enables their training loss to approach zero values as the validation loss stays steady, indicating acceptable but not peak generalizability. A distinct characteristic of the proposed model lies in its stable loss reduction pattern because training loss correlates closely with validation loss while showing excellent generalization capacity. The proposed model demonstrates lower overfitting

through its minimal training–validation loss gap; thus, it offers enhanced performance for complex data distributions.

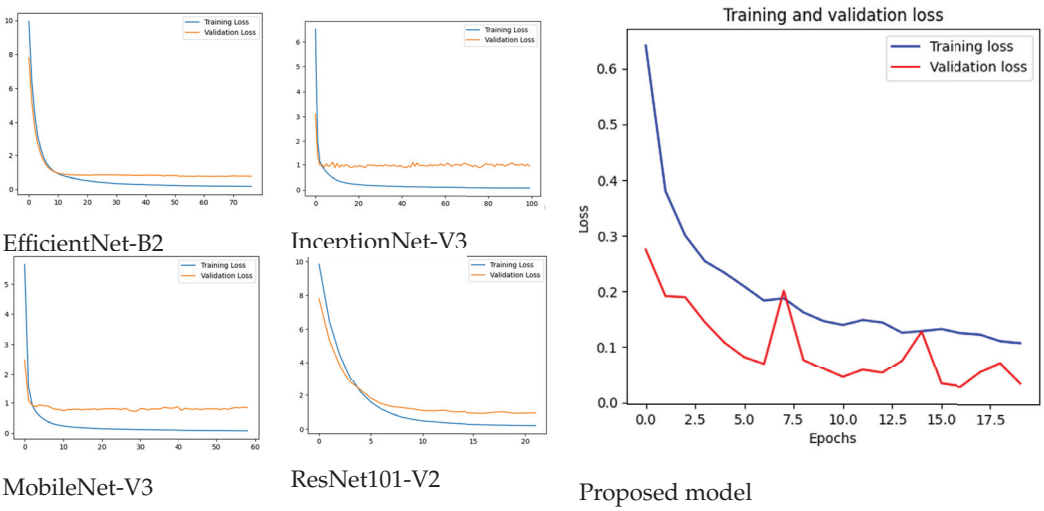


Figure 7. Losses for all models.

Figure 8 shows receiver operating characteristic (ROC) curve analysis with performance evaluation of ten distinct classes, including eczema, alongside melanoma and basal cell carcinoma. The representation of the true positive rate (TPR) vs. false positive rate (FPR) trade-offs in specific classes occurs through individual curves alongside area under the curve (AUC) metrics, which assess model distinction capability. The AUC = 1.00 perfect separation results of classes melanoma and basal cell carcinoma are distinguished from other classes such as eczema and warts molluscum, which exhibit strong but sub-perfect performance, with AUC values between 0.93 and 0.99. Random guessing stands as the reference point, while model effectiveness increases when the curves get closer to the top-left area of the graph. High classification accuracy stands out in the figure, even as performance shows varying levels across specified conditions, especially for selected classes.

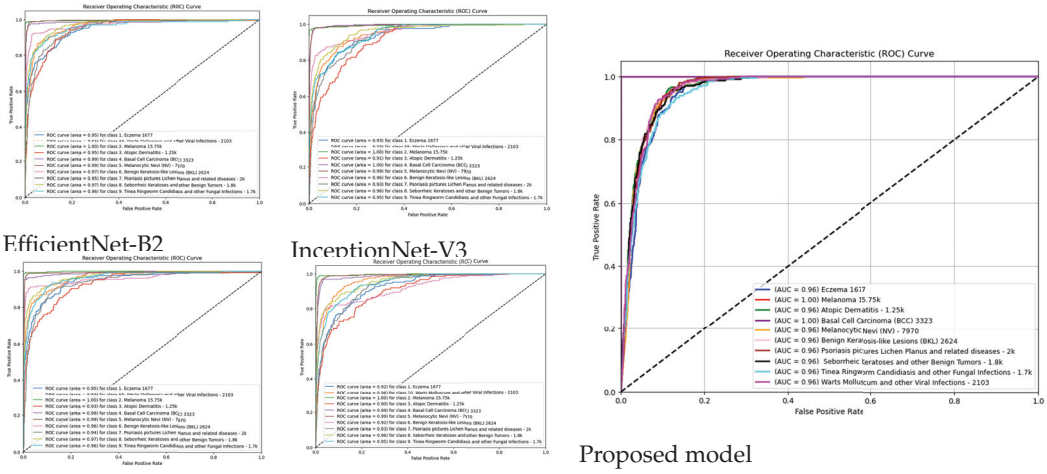


Figure 8. ROC curve for all models.

4.4. Discussion

The fusion-based deep learning model, which combined EfficientNet-B0, EfficientNet-B2, and ResNet50, reached excellent results in skin disease classification, with an accuracy of 99.14%. The dual benefit of merging multiple pre-trained architectural frameworks allowed the model to acquire different features, thereby strengthening its ability to per-

form generalizations across diverse skin conditions. The utilization of single-pretrained EfficientNet-B2 proves effective, but the multi-architecture fusion approach extracts wider features while achieving optimal performance in skin disease classification. The proposed model benefited from additional regularization techniques, which prevented overfitting while maintaining training stability to enhance its total performance.

The proposed model achieves better class imbalance performance than MobileNet-v3 and ResNet101v2. The design of MobileNet-v3 focuses on efficiency, which results in an accuracy level reaching 85%, although this result makes it inadequate for analyzing complex medical datasets needed for skin disease diagnosis. Efforts to use ResNet101v2 resulted in 80% accuracy, yet the model faced challenges when dealing with multiple diseases and similar patterns between them. The proposed fusion model delivers a comprehensive performance improvement through an accuracy of 99.14%. Through fusion, the model benefits from multiple network-learned features to achieve precise identification of skin disease variations.

The combination model demonstrates superior performance compared to Inception and EfficientNet-B2 on challenging multi-class skin disease identification tasks. The distribution–meaning reduction by convolution filters of Inception networks creates a challenge when dealing with intra-class differences in data. EfficientNet-B2 shows superior performance when it comes to both parameter utilization and processing speed, but the accuracy is 84% for detecting multi-class skin diseases. The proposed model merges EfficientNet-B0 with EfficientNet-B2 plus ResNet50 as its components to overcome these hurdles since the combination enhances performance for minority class detection and guarantees improved accuracy in all categories.

Comparative analysis shows the evolution and further developments in skin disease classification methods adopted by deep learning frameworks, as shown in Table 4. The study of Venkata et al. [16] involved the use of the EfficientNet-B2 model on the DermNet NZ Image Library and was 89.55% accurate. While this result is quite impressive, it specializes in identifying simple patterns that need some improvement when it comes to capturing complex skin disease patterns. In contrast, K. Vayadande [17] only adopted regularization techniques in the CNN framework, whereas the datasets were collected from the Kaggle site and obtained a striking accuracy of 98%. This improvement is notable for proving the idea of enhancing the CNN architecture designed for generalization and model robustness. Similarly, Rezaee [18] used capsule networks to enhance feature extraction, with an accuracy of 96.87% on the PH2 dataset, and warned that further improvement may be required to achieve greater levels of accuracy. However, the sophistication of models that have been developed in the recent past, like those by Nirupama [19] and Mavaddati [20], has led to improved benchmarks in skin disease classification. Thus, by using MobileNet-V2 and applying additional modules, such as squeeze-and-excitation networks and atrous spatial pyramid pooling, Nirupama achieved 98.6% accuracy on multiple datasets. To round off, Mavaddati developed a ResNet50-LSTM to increase accuracy by a margin; the repeatability of 99.09% proved this was possible when using both convolutional and sequential models to capture lesion texture and structural content. These include the following smart state-of-the-art techniques that the proposed model outperforms, yielding a maximum accuracy of 99.14% on datasets from Kaggle. This shows that the proposed model can be used in real-world applications since it indicates high classification accuracy while at the same time presenting a workable solution for classification methods useable for various data sizes and applicable in scalable classifications.

Table 4. Comparative analysis of the proposed model vs. state-of-the-art methods.

Reference No.	Year	Technique	Collections	Accuracy
Venkata et al. [20]	2024	EfficientNet-B2 (CNN model)	DermNet NZ Image Library	89.55
K. Vayadande [21]	2024	Regularization techniques within the CNN framework	Datasets sourced from Kaggle	98.00
Rezaee [22]	2024	Utilization of capsule networks for improved feature extraction	PH2 datasets	
Nirupama [23]	2024	MobileNet-V2	PH2 dataset, Skin Cancer MNIST: HAM10000 dataset, DermNet. dataset, and Skin Cancer ISIC dataset.	98.6
Mavaddati [24]	2025	ResNet50 and LSTM	Skin Dataset	99.09
Proposed Model	-	ResNet50 and EfficientNet(B0, B2)	Skin Datasets sourced from Kaggle	99.14

5. Conclusions

The deep learning model that fuses a classification network with two pre-trained networks shows better results for identifying skin diseases. The proposed model design effectively captures detailed visual characteristics in dermatological images through its multi-stream architecture that implements EfficientNet-B0, EfficientNet-B2, and ResNet50 components. The combination strategy permits these networks to work together using their respective strengths while resolving the unique challenges that single models encounter during both inter-class and intra-class classification. The model delivers exceptional results on the Kaggle Skin Diseases Image Dataset with 99.14% accuracy together with matching precision and recall values and an F1 score of 0.9914, which validates its diagnostic credibility. The model executes medical image analysis with deep learning-based complexity while operating within 200 seconds of execution time. The research results demonstrate how model fusion approaches can boost dermatological image classification procedures. Deep learning establishes itself as a dependable tool for diagnostic support in medical skin disease assessment by achieving nearly flawless performance. These outcomes create standards for medical imaging research that advance the practical applicability of modern deep learning approaches in healthcare.

Several important obstacles must be considered. The quality of data and its variation level determine how well the model functions. The model is now limited to the Kaggle dataset, which contains a consistent group of skin diseases or conditions, but it cannot diagnose any other pathologies, and performance is unknown on other datasets. Training and inference procedures from this system require significant computational power that becomes a major factor for adoption among resource-restricted locations, including mobile devices and rural healthcare services. The model requires external validation with various datasets before it can be used clinically because its effectiveness on external datasets remains untested. Future research must devote efforts to building a larger database that includes various skin disorders, particularly uncommon ones, to extend the model's generalization abilities. The optimized model design will make it feasible to use this system in real-time diagnostic tools alongside mobile health applications. Enhanced predictions will result from processing multiple types of patient information that include demographic features alongside clinical background records. By implementing the model into clinical decision support systems, healthcare professionals would gain more effective disease diagnosis capabilities. This study positions itself as the foundation for developing future sophisticated

dermatological diagnosis tools that will deliver better accessibility and durability through AI technology.

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

Funding: This work was funded by the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 223202.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Images dataset available at https://www.kaggle.com/datasets/ismailpromus/skin-diseases-image-dataset, (accessed on 5 January 2025).

Acknowledgments: The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 223202.

Conflicts of Interest: The authors have no conflicts of interest to report regarding the present study.

References

1. Do, T.T.; Zhou, Y.; Zheng, H.; Cheung, N.M.; Koh, D. Early melanoma diagnosis with mobile imaging. In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Chicago, IL, USA, 26–30 August 2014; pp. 6752–6757.
2. Aleem, M.; Hameed, N.; Anjum, A. m-Skin doctor: A mobile-enabled system for early melanoma skin cancer detection using support vector machine. *Comput. Biol. Med.* **2017**, *87*, 468–475.
3. Barata, C.; Marques, J.S.; Rozeira, J. The role of key point sampling on the classification of melanomas in dermoscopy images using bag-of-features. In *Iberian Conference on Pattern Recognition and Image Analysis*; Marques, J.S., De la Blanca, N.P., Pina, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7887, pp. 715–723.
4. Ashraf, R.; Afzal, S.; Rehman, A.U.; Gul, S.; Baber, J.; Bakhtyar, M.; Mehmood, I.; Song, O.Y.; Maqsood, M. Region-of-interest-based transfer learning-assisted framework for skin cancer detection. *IEEE Access* **2020**, *8*, 147858–147871. [CrossRef]
5. Akther, M.; Akter, S.H.; Sarker, S.; Aleri, J.W.; Annandale, H.; Abraham, S.; Uddin, J.M. Global burden of lumpy skin disease, outbreaks, and future challenges. *Viruses* **2023**, *15*, 1861. [CrossRef] [PubMed]
6. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 10–15 June 2019.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
8. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
9. 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]
10. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [CrossRef]
11. Meštrović, T.; Matijašić, M.; Perić, M.; Čipčić Paljetak, H.; Barešić, A.; Verbanac, D. The Role of Gut, Vaginal, and Urinary Microbiome in Urinary Tract Infections: From Bench to Bedside. *Diagnostics* **2021**, *11*, 7. [CrossRef]
12. Mujahid, M.; Rustam, F.; Álvarez, R.; Luis Vidal Mazón, J.; Díez, I.d.l.T.; Ashraf, I. Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network. *Diagnostics* **2022**, *12*, 1280. [CrossRef] [PubMed]
13. Bordin, D.S.; Voynovan, I.N.; Andreev, D.N.; Maev, I.V. Current Helicobacter pylori Diagnostics. *Diagnostics* **2021**, *11*, 1458. [CrossRef] [PubMed]
14. Thomas, C.; Minty, M.; Vinel, A.; Canceill, T.; Loubières, P.; Burcelin, R.; Kaddech, M.; Blasco-Baque, V.; Laurencin-Dalicieux, S. Oral Microbiota: A Major Player in the Diagnosis of Systemic Diseases. *Diagnostics* **2021**, *11*, 1376. [CrossRef]

15. Mosquera, C.; Ferrer, L.; Milone, D.H. Class imbalance on medical image classification: Towards better evaluation practices for discrimination and calibration performance. *Eur. Radiol.* **2024**, *34*, 7895–7903. [CrossRef] [PubMed]
16. Zhao, J.; Zhou, Y.; Zhong, R.; Li, C.; Li, X.; Qin, J.; Zhuang, M.; Wen, L.M. Addressing Class Imbalance in Skin Lesion Classification with Dynamic Multi-Output Convolutional Neural Network. Available online: https://ssrn.com/abstract=4367546 (accessed on 5 January 2025).
17. Müller, D.; Soto-Rey, I.; Kramer, F. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *IEEE Access* **2022**, *10*, 66467–66480. [CrossRef]
18. Yi, L.; Zhao, J.; Lv, Z.; Pan, Z. Multimodal medical supervised image fusion method by CNN. *Front. Neurosci.* **2021**, *15*, 638976.
19. Haghshenas, F.; Krzyżak, A.; Osowski, S. Comparative study of deep learning models in melanoma detection. In *Artificial Neural Networks in Pattern Recognition*; Suen, C.Y., Ed.; Springer: Berlin/Heidelberg, Germany, 2024; Volume 15154, pp. 104–116.
20. Venkata Sekhar, N.B.D.; Purushotham Reddy, M. Feature selection based on dragonfly optimization for psoriasis classification. *Int. J. Intell. Syst. Appl. Eng.* **2024**, *12*, 935–943.
21. Vayadande, K. Automated multiclass skin disease diagnosis using deep learning. *Int. J. Intell. Syst. Appl. Eng.* **2024**, *12*, 327–336.
22. Rezaee, K.; Zadeh, H.G. Self-attention transformer unit-based deep learning framework for skin lesions classification in smart healthcare. *Discov. Appl. Sci.* **2024**, *6*, 3. [CrossRef]
23. Nirupama, V. MobileNet-V2: An enhanced skin disease classification by attention and multi-scale features. *J. Digit. Imaging* **2024**. [CrossRef]
24. Mavaddati, S. Skin cancer classification based on a hybrid deep model and long short-term memory. *Biomed. Signal Process. Control* **2025**, *100*, 107109. [CrossRef]
25. 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]
26. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
27. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
28. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
29. Elsken, T.; Metzen, J.H.; Hutter, F. Neural architecture search: A survey. *J. Mach. Learn. Res.* **2019**, *20*, 1–21.
30. Kaggle. Skin Diseases Image Dataset. Available online: https://www.kaggle.com/datasets/ismailpromus/skin-diseases-image-dataset (accessed on 5 January 2025).

diagnostics

MDPI

Article

Deep Learning-Based Object Detection Strategies for Disease Detection and Localization in Chest X-Ray Images

Yi-Ching Cheng [1], Yi-Chieh Hung [1], Guan-Hua Huang [1,*], Tai-Been Chen [2,3,4], Nan-Han Lu [5], Kuo-Ying Liu [5] and Kuo-Hsuan Lin [6]

[1] Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[2] Department of Radiological Technology, Faculty of Medical Technology, Teikyo University, Tokyo 173-8605, Japan
[3] Infinity Co., Ltd., Taoyuan 320021, Taiwan
[4] Der Lih Fuh Co., Ltd., Taoyuan 320021, Taiwan
[5] Department of Radiology, E-Da Cancer Hospital, I-Shou University, Kaohsiung 824005, Taiwan
[6] Department of Emergency Medicine, E-Da Hospital, I-Shou University, Kaohsiung 824005, Taiwan
* Correspondence: ghuang@nycu.edu.tw; Tel.: +886-3-513-1334

Abstract: Background and Objectives: Chest X-ray (CXR) images are commonly used to diagnose respiratory and cardiovascular diseases. However, traditional manual interpretation is often subjective, time-consuming, and prone to errors, leading to inconsistent detection accuracy and poor generalization. In this paper, we present deep learning-based object detection methods for automatically identifying and annotating abnormal regions in CXR images. Methods: We developed and tested our models using disease-labeled CXR images and location-bounding boxes from E-Da Hospital. Given the prevalence of normal images over diseased ones in clinical settings, we created various training datasets and approaches to assess how different proportions of background images impact model performance. To address the issue of limited examples for certain diseases, we also investigated few-shot object detection techniques. We compared convolutional neural networks (CNNs) and Transformer-based models to determine the most effective architecture for medical image analysis. Results: The findings show that background image proportions greatly influenced model inference. Moreover, schemes incorporating binary classification consistently improved performance, and CNN-based models outperformed Transformer-based models across all scenarios. Conclusions: We have developed a more efficient and reliable system for the automated detection of disease labels and location bounding boxes in CXR images.

Keywords: chest X-rays; deep learning; few-shot object detection; object detection

Citation: Cheng, Y.-C.; Hung, Y.-C.; Huang, G.-H.; Chen, T.-B.; Lu, N.-H.; Liu, K.-Y.; Lin, K.-H. Deep Learning-Based Object Detection Strategies for Disease Detection and Localization in Chest X-Ray Images. *Diagnostics* **2024**, *14*, 2636. https://doi.org/10.3390/diagnostics14232636

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 10 October 2024
Revised: 15 November 2024
Accepted: 18 November 2024
Published: 22 November 2024

1. Introduction

Medical images play a crucial role in disease prevention, detection, and diagnosis, providing essential support for clinicians. Of the various types, chest X-ray (CXR) images are particularly valuable for detecting abnormalities in the lungs, heart, and bones, which aids in making appropriate treatment decisions. The accurate analysis of these images is highly beneficial for improving patient care. In this study, we aim to enhance diagnostic accuracy by analyzing CXR images for 12 common chest conditions, including aortic sclerosis (calcification), arterial curvature, small pulmonary nodules, pulmonary nodule shadows, tuberculosis, pulmonary fibrosis, increased lung markings, prominent hilar regions, spinal lesions, intercostal pleural thickening, cardiac hypertrophy, and the presence of heart pacemakers. Twelve conditions were selected for study due to their reliable high-quality annotations, which are crucial for developing accurate models. Additionally, they have clear visual manifestations, making them more detectable via automated analysis.

Traditionally, doctors manually detect abnormalities in chest images through visual examination, which can be influenced by personal biases and external factors, leading to

inconsistent results. During initial CXR screenings, physicians must also manually label lesion areas, a process that is time-consuming and labor-intensive. With the rapid growth in the volume of clinical image data, the workload for doctors has increased significantly. In recent years, artificial intelligence, particularly machine learning (ML), has emerged as a powerful tool for addressing such challenges. Deep learning, a subset of ML, has shown great success in computer vision tasks like image classification, segmentation, and object detection. Consequently, researchers have begun applying deep learning to medical image analysis to automate tasks such as disease diagnosis, detection, and lesion localization. These automated methods allow time savings, improving diagnostic efficiency and reducing the impact of external factors.

Several studies have applied deep learning to classify CXR images, aiming to aid in diagnosing a wide range of diseases. In particular, deep learning has been widely used to support automated diagnosis of COVID-19 from CXR images. For example, Ali et al. [1] developed a densely connected squeeze convolutional neural network (CNN) for classifying cases of COVID-19 and pneumonia with high accuracy, showcasing the potential of deep learning to enhance diagnostic reliability in the context of a pandemic. Singh et al. [2] also proposed a CNN architecture where segmentation and classification were combined to boost the classification accuracy for COVID-19-affected CXR images. Other studies have focused on identifying various types of pneumonia. Garstka and Strzelecki [3] developed a custom CNN, trained on a small dataset, to classify pneumonia types from CXR images. Additionally, recent studies have explored using deep learning to detect multiple lung diseases. For instance, Rana et al. [4] created an automated system for classifying 10 different lung diseases, utilizing a flexible CNN architecture in which graph neural networks were integrated with feedforward layers. A comprehensive analysis of deep learning applications in lung cancer diagnosis and classification is provided in a recent systematic review [5].

Object detection involves identifying and locating specific objects within an image, merging recognition with localization tasks. Through the application of modern object detection techniques, notable success has been achieved across various fields, including wildlife monitoring [6], autonomous driving [7], defect inspection [8], security surveillance [9], and face mask detection [10]. These advances have largely been driven by deep learning models, which are typically based on two architectures: convolutional neural networks (CNNs) [11–16] and self-attention-based Transformers [17]. CNN-based detectors are classified as either one-stage models, like the YOLO series [18–20], SSD [21], and RetinaNet [22], which prioritize speed, or two-stage models, such as the R-CNN family [23–25], which focus on accuracy. Recently, Transformer-based models like DETR [26] and Deformable DETR [27] have also gained popularity, reflecting ongoing innovations in the field of object detection.

Despite this progress, there are several challenges hindering the application of deep learning in medical image object detection [28–30]. For example, most deep learning object detection models are trained on the MS COCO dataset [31], which consists primarily of images unrelated to medical applications. This raises concerns about whether models trained on such datasets can perform well when applied to CXR images—a key issue this study seeks to address.

Another challenge is the costly and time-consuming process of labeling medical image data, especially for object detection tasks that require detailed annotations, such as adding bounding boxes around lesions. This task is even more difficult for CXR images, as it requires the expertise of radiologists, making the process more complex and resource-intensive [30]. As a result, the dataset used in this study contains only a limited number of training images for each disease category, with some categories having very few examples. This data scarcity poses a significant challenge for training deep learning models, for which large amounts of labeled data to avoid overfitting are required. To address this, we employ few-shot object detection methods, which are designed to recognize new (unseen)

disease categories using only a few training examples after the model has been trained on numerous examples of known (seen) categories [32–34].

In this study, we aim to improve CXR image analysis by focusing on disease labels and location bounding boxes for object detection. We explore advanced deep-learning models, incorporating few-shot techniques to enhance their performance. Additionally, we compare various deep learning methods to evaluate their strengths and weaknesses, ultimately seeking to develop a more efficient and reliable system for the automated detection of chest diseases in CXR images.

2. Related Work

Substantial progress has been made in deep learning for image classification and object detection, impacting fields like medical imaging. However, accurately detecting specific disease markers in CXRs remains challenging, especially for rare conditions with limited data. A review of the current classification and detection methods reveals several gaps and limitations that serve as motivation for further research in this area.

2.1. Classification

Significant progress has been made in network architectures for image classification. Scaling up neural networks by increasing their depth can enhance accuracy but may also lead to the vanishing gradient problem. This was addressed through skip connections using ResNet [15] to improve gradient flow in deeper networks. DenseNet [16] is an expansion of ResNet with dense connections that allow each layer to receive feature maps from all previous layers, enabling feature reuse across layers to reduce parameter count and improve efficiency. Model scaling was further optimized in developing EfficientNet (B0–B7) [35] by balancing depth, width, and resolution, achieving a strong trade-off between accuracy and computational cost.

However, while these classification models are powerful for general tasks, they are limited in their ability to localize and classify the smaller more subtle abnormalities often found in CXRs, which require precise object detection capabilities beyond merely classification.

2.2. Object Detection

Different object detection architectures, typically categorized as one- or two-stage detectors, each have strengths and weaknesses. Two-stage detectors like those derived from the R-CNN framework are generally more accurate as they utilize a refined candidate selection process that filters out negative samples early, while one-stage detectors, exemplified by the YOLO series, focus on real-time detection but often generate too many candidate boxes, causing class imbalance by overfocusing on background samples. The focal loss function introduced in RetinaNet [22] improved the one-stage detector by reducing the influence of easy samples and enhancing learning from difficult samples. RetinaNet also uses a feature pyramid network [36] to integrate features from feature maps of different scales, thereby enhancing its feature extraction capability. YOLOv3 was improved using a decoupled head design, stronger data augmentation, and a shift to an anchor-free framework in developing the real-time detector YOLOX [37], with fewer parameters and better generalization. The problem of fixed IoU thresholds was addressed using Dynamic R-CNN [38] by dynamically adjusting the threshold during training and refining the loss function based on regression label statistics. DETR [26] revolutionized object detection by employing the Transformer architecture for end-to-end detection, eliminating the need for traditional anchor boxes or region proposals, though it suffers from slow convergence and poor small object detection. These issues were addressed using Deformable DETR [27], which focuses attention on key points near a reference point, improving performance for high-resolution images and small object detection.

Although these models have been adapted for various fields, their direct application to CXR analysis remains problematic due to issues such as high computational demand, slow convergence, and difficulties in detecting small but clinically relevant features. This

highlights the need for more specialized object detection approaches that can overcome these limitations within the context of medical imaging.

2.3. Few-Shot Object Detection (FSOD)

The aim of few-shot learning is to build a model that can accurately classify images using very few training examples for specific classes. In FSOD, the categories are divided into base classes (with many training examples) and novel classes (with fewer examples). There are two stages in the training process: base training and k-shot fine-tuning. During base training, the model is only trained on base class objects, even if the images also contain novel class objects. In the k-shot fine-tuning stage, a small number (k) of bounding boxes from each class are used to refine the model. This approach is particularly useful in medical image analysis, where it may be difficult to collect data, with some diseases being extremely rare.

Meta-learning, which focuses on "learning to learn", is crucial for FSOD, where models are trained on tasks from dataset subsets to rapidly adapt to new tasks. This fine-tuning approach was previously considered less effective until the two-stage fine-tuning approach TFA [39] challenged this view. TFA, built on Faster R-CNN, was initially trained on base classes and fine-tuned only the box predictor for all classes, improving accuracy by replacing the fully connected classifier with a cosine similarity-based classifier. The classification accuracy for novel classes was improved through FSCE [40] by using contrastive learning to separate novel instances from base classes. The contrastive proposals encoding loss were added to the Faster R-CNN loss, enhancing accuracy. In Meta-DETR [41], the first image-level few-shot detector, generalization was improved by incorporating the correlational aggregation module to capture inter-class correlations and reduce misclassification.

While these methods have shown success in domains like Pascal VOC and MS COCO, their performance on CXR datasets remains underexplored. Current FSOD techniques often struggle with inter-class variability and can suffer from misclassification, particularly in complex medical datasets where diseases may have overlapping visual features. Thus, developing a specialized FSOD approach for CXRs could significantly enhance model adaptability and reliability in detecting rare diseases.

2.4. Deep Learning-Based Object Detection for CXR Images

In recent years, deep learning-based object detection has been applied to CXR images for identifying foreign objects [42] and localizing abnormalities [43,44] in assisting the diagnosis of various diseases. Advanced architectures such as YOLO, RetinaNet, Mask R-CNN, and Faster R-CNN have been adapted for CXR analysis, achieving high accuracy and fast localization [45]. Notably, in a direct comparison of performance, the YOLOX model surpassed radiologists [44]. Large datasets with ground-truth bounding boxes, such as VinDr-CXR (open dataset of 18,000 CXRs with 28 abnormalities) [46] and CXR-AL14 (dataset available upon request for 165,988 CXRs with 14 abnormalities) [44], have been created to enhance model training.

To improve nodule detection performance, Behrendt et al. [47] evaluated strategies such as transfer learning using pre-trained weights from the VinDr-CXR and COCO datasets, as well as training from scratch. They addressed class imbalance by augmenting training data with generated nodules in healthy CXRs and compared this to oversampling the less frequent class (CXRs with nodules). After testing various state-of-the-art object detection algorithms, they developed a systematic approach that incorporated the most effective techniques, ultimately outperforming all competitors in the NODE21 competition's detection track [48].

3. Materials

3.1. Dataset

We used a dataset containing 2123 CXR images, featuring both normal cases and 18 types of diseases. The images were in the DICOM format. These images were retrospec-

tively collected from the archiving and communication system (PACS) at E-Da Hospital, covering patient CXRs from January 2008 to December 2018. Along with the images, the dataset included patient information such as gender, age, and diagnostic reports from radiologists. The Institutional Review Board of E-Da Hospital approved this study, and all patients provided written informed consent.

The 18 disease types were chosen by reviewing diagnostic reports and selecting those with reliable high-quality annotations. We also prioritized diseases with visual manifestations that could be effectively detected by object detection algorithms, making them suitable for automated analysis.

An experienced radiological physician (K.-Y.L.) identified and marked lesion regions on the image. The rectangular bounding box was carefully placed to closely surround the lesion, capturing its full extent while minimizing any inclusion of unaffected tissue. After this initial placement, another senior radiologist (N.-H.L.) reviewed and confirmed that bounding boxes were accurately sized and precisely positioned.

Images were excluded if they were of poor quality or had unclear diagnostic reports. We also excluded images of minors (patients under 18). After removing duplicates and missing data, we retained 1802 images, each representing a unique patient. The image sizes varied, with heights ranging from 1304 to 4280 pixels and widths from 1066 to 4280 pixels. The dataset employs multi-label classification, as a single patient can have multiple diseases. Considering the number of cases as well as the sizes and locations of bounding boxes, we grouped the 18 diseases into 12 categories based on medical guidance. Table 1 presents the number of images before and after merging the diseases, with the abbreviations of the 12 disease names provided for simplicity. Notably, the number of normal cases is much higher than the combined total of the 12 diseases, indicating there is a significant class imbalance in the dataset.

Table 1. Number of images in each disease category before and after merging.

Before		**After**		
Categories	**Count**	**New Categories**	**Abbr.**	**Count**
Normal	1212	Normal	Normal	1212
Aortic arch atherosclerotic plaque	28	Aortic sclerosis (calcification)	AorScl(Cal)	83
Aortic arch calcification	16			
Aortic atherosclerosis	25			
Aortic wall calcification	20			
Aortic curvature	65	Arterial curvature	ArtCur	93
Thoracic vertebral artery curvature	28			
Small pulmonary nodules	15	Small pulmonary nodules	SmaPulNod	15
Shadows of pulmonary nodules	8	Shadows of pulmonary nodules	ShaOfPulNod	8
Tuberculosis	6	Tuberculosis	tuberculosis	6
Pulmonary fibrosis	30	Pulmonary fibrosis	PulFib	30
Increased lung streak	89	Increased lung patterns	IncLunPat	225
Lung field infiltration	138			
Obvious hilar	55	Obvious hilar	ObvHil	55
Degenerative joint disease of the thoracic spine	75	Spinal lesions	SpiLes	170
Scoliosis	100			
Intercostal pleural thickening	52	Intercostal pleural thickening	IntPleThi	52
Cardiac hypertrophy	41	Cardiac hypertrophy	CarHyp	41
Heart pacemaker placement	9	Heart pacemaker placement	HeaPacPla	9

Each image in the 12 disease categories contains one or more bounding boxes. For example, annotations for ObvHil (obvious hilar) and PulFib (pulmonary fibrosis) are often paired, while multiple bounding boxes are typical for SmaPulNod (small pulmonary nodules). Figure 1 is a bar chart illustrating the total number of images and bounding boxes for each disease category, and Figure 2 shows sample X-ray images with their corresponding bounding boxes across the 12 categories.

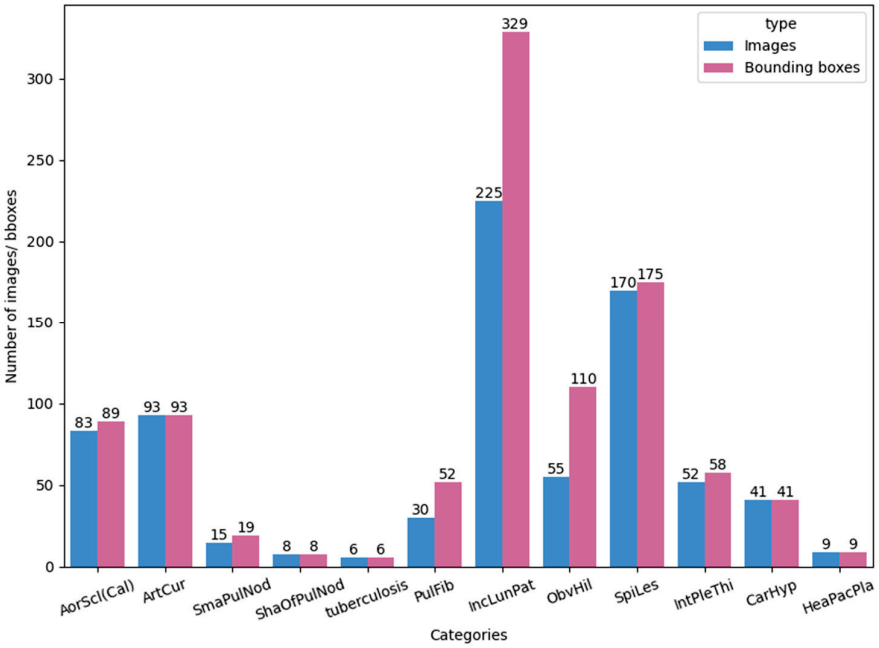


Figure 1. The number of images and bounding boxes for each of the 12 disease categories.

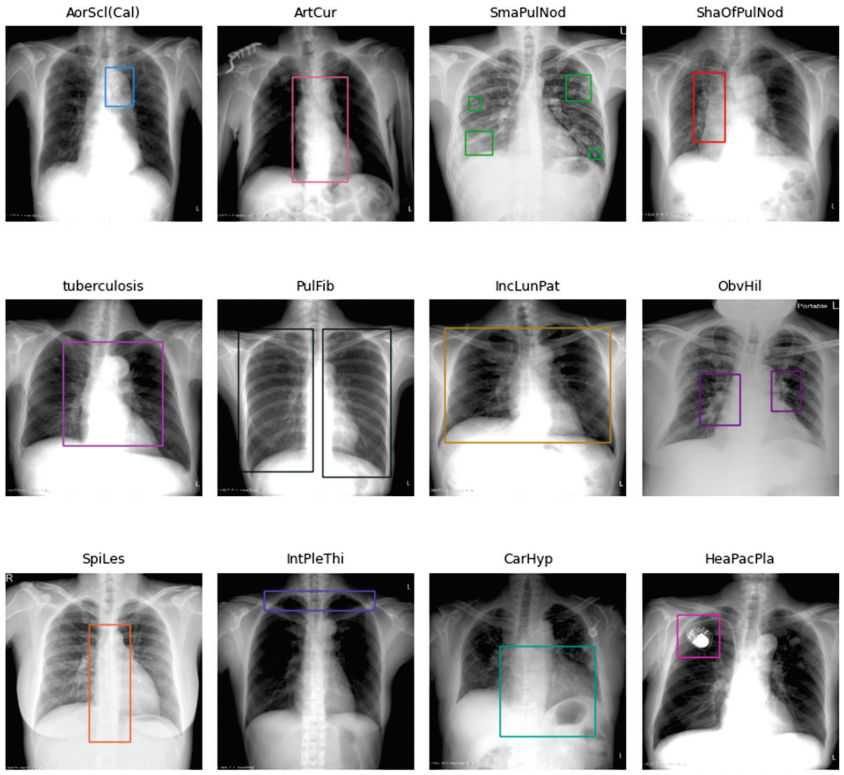


Figure 2. Examples of X-ray images and their corresponding bounding boxes across the 12 disease categories.

3.2. Data Preprocessing

We processed the CXR images using header data embedded in the DICOM files. If the DICOM file indicated a logarithmic relationship between pixel values and X-ray beam intensity, we applied "intensity log transformation". In this process, each pixel value $x[i]$ is adjusted based on the visible range defined by the Window Center (WC) and Window Width (WW). The visible pixel range is between $iMin = \text{WC} - \frac{\text{WW}}{2}$ and $iMax = \text{WC} + \frac{\text{WW}}{2}$, while the number of bits for each pixel is defined by BitsStored. The steps for the intensity log transformation are depicted in Algorithm 1.

Algorithm 1 Pseudocode of the intensity log transformation

Input: x
  for $i = 0, \cdots, N-1$ **do**
    if $x[i] < iMin$, **then** $x[i] = iMin$
    if $x[i] > iMax$, **then** $x[i] = iMax$
  $z[i] = -\log\left(\frac{1+x[i]}{2^{\text{BitsStored}}}\right)$
  end for
Output: z

CXR images often contain elements, such as chest markers, that are irrelevant to disease detection. These markers often appear overexposed after logarithmic transformation, such as in the example of letter "L" (Figure 3a). To enhance the areas of interest, we adjusted image contrast using the "simplest color balance algorithm", in which saturation limits of v_{min} and v_{max} are set to improve contrast (Algorithm 2).

Algorithm 2 Pseudocode of the simplest color balance

Input: z
  for $i = 0, \cdots, N-1$ **do**
  $c[i] = \frac{z[i]-v_{min}}{v_{max}-v_{min}}$
    if $c[i] < 0$, **then** $c[i] = 0$
    if $c[i] > 1$, **then** $c[i] = 1$
  end for
Output: c

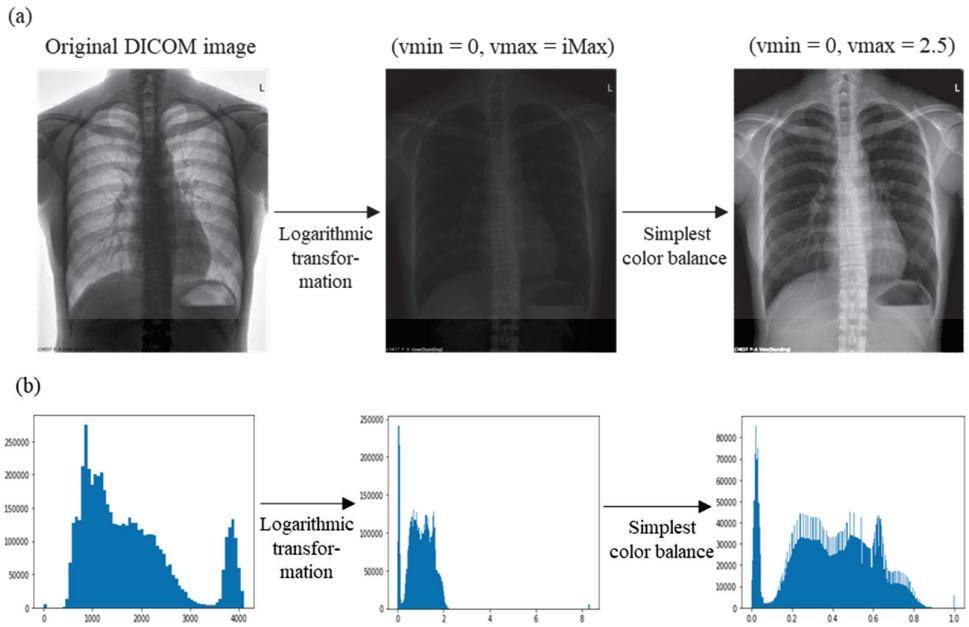


Figure 3. (**a**) X-ray images before and after data preprocessing and (**b**) their corresponding intensity histograms.

In this study, we set $v_{min} = 0$ and $v_{max} = 2.5$. Figure 3 shows the progression from the original DICOM image, through intensity log transformation, to the final contrast-adjusted image using the simplest color balance algorithm. Intensity histograms for each step are also shown.

3.3. Experimental Data Setups

To assess the impact of having a large proportion of normal images in our dataset, a common issue in clinical practice, we created three datasets for our object detection models. First, the entire dataset (1802 samples), labeled Dataset A, was divided into training, validation, and test sets with approximate proportions of 63.4%, 16.5%, and 20.1%, respectively, while ensuring similar disease distributions across all subsets. Next, we created Dataset B by removing two-thirds of the normal images from the training set and Dataset C by removing all normal images from the training set. Table 2 shows the number of images in each dataset.

Table 2. Number of images in the three datasets used for object detection.

Categories [1]	Training [2]	Validation	Test	Total [2]
Normal	779/178/0	189	244	1212/611/433
AorScl(Cal)	52	15	16	83
ArtCur	57	18	18	93
SmaPulNod	9	3	3	15
ShaOfPulNod	6	1	1	8
tuberculosis	3	1	2	6
PulFib	20	6	4	30
IncLunPat	130	42	53	225
ObvHil	34	12	9	55
SpiLes	107	28	35	170
IntPleThi	35	8	9	52
CarHyp	23	9	9	41
HeaPacPla	6	1	2	9
Unique images	1143/542/364	297	362	1802/1201/1023

[1] Categories in bold represent novel classes. [2] Datasets A, B, and C have the same number of images for each disease category but differ regarding the number of images for the normal category in the training set. For simplicity, the notation A/B/C is used to represent the number of images in Datasets A, B, and C, respectively.

For the FSOD models, we performed an extra step, dividing the disease categories into base and novel classes for base training and *k*-shot fine-tuning. We used the same three datasets created earlier, designating categories with fewer images—SmaPulNod, ShaOfPulNod (shadows of pulmonary nodules), tuberculosis, PulFib, and HeaPacPla (heart pacemaker placement)—as novel classes. The remaining seven categories were treated as base classes. During *k*-shot fine-tuning, we set *k* to 1, 2, 3, 5, or 10, meaning that we randomly selected up to 10 images for annotation per novel category. When there were fewer than 10 images in a category, we used all of the available images. An image selected for one category was not reused for another.

4. Methods

In this study, we applied object detection and FSOD methods to identify disease types and lesion areas in CXR images. We designed four analytical schemes to determine the most effective approach. These schemes involved using either object detection or the FSOD models, with or without a preliminary binary classification step to determine the presence of disease in the image. For binary classification, object detection, and FSOD tasks, we

selected two, five, and three models, respectively, as shown in Table 3. To understand the impact of model architecture on performance for both object detection and FSOD tasks, we chose models from two primary categories: CNN-based and Transformer-based models. Additionally, we calculated the specificity of normal images in the test set for each scheme to assess whether the models could maintain a low misdiagnosis rate while excelling at disease detection.

Table 3. Models used in this study.

		Architecture	
		CNN-Based	**Transformer-Based**
Task	Binary classification	EfficientNet-B3 [35], DenseNet121 [16]	
	Object detection	RetinaNet [22], YOLOX [37], Dynamic R-CNN [38]	DETR [26], Deformable DETR [27]
	Few-shot object detection	TFA [39], FSCE [40]	Meta-DETR [41]

4.1. Scheme 1: Object Detection

In Scheme 1, we trained object detection models using three datasets: A, B, and C. After training, we tested these models on the test set and calculated key evaluation metrics, including average precision (AP) and mean average precision (mAP), for detecting diseases in the 12 disease categories. The overall process is depicted in Figure 4.

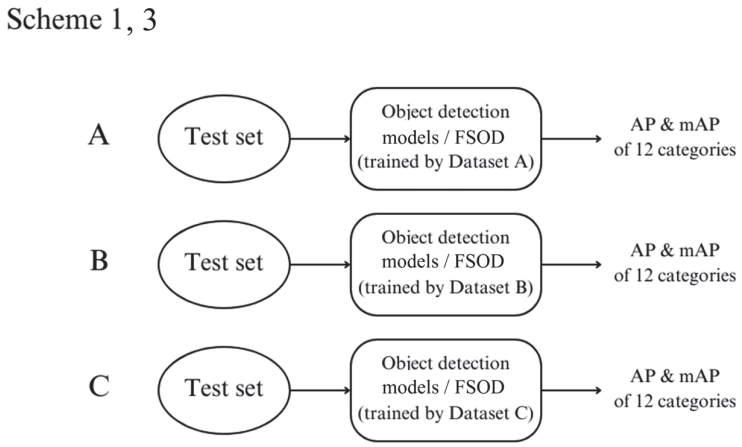


Figure 4. Flowchart for Schemes 1 and 3, outlining the steps involved in object detection or few-shot object detection (FSOD).

4.2. Scheme 2: Binary Classification + Object Detection

In this scheme, a binary classification step is introduced before object detection. Two classification models, Classification Models A and B, were trained on Datasets A and B, respectively, to determine whether a patient had any disease. Classification Model A uses the EfficientNet-B3 architecture, while Classification Model B employs DenseNet121.

During testing, images from the test set were first classified by the binary models. Images classified as positive (indicating the presence of disease) were passed onto the object detection models trained in Scheme 1. Those classified as negative (indicating no disease) were labeled as normal and were not subjected to further object detection.

Since Scheme 1 involves training object detection models on three datasets and this schedule includes classification models on two datasets; there are six possible outcomes for each test image: A + A, A + B, B + A, B + B, C + A, and C + B. For instance, in the A + B case, the image was first classified by the classification model trained on Dataset B and, if positive, analyzed by the object detection model trained on Dataset A. Figure 5 illustrates this process.

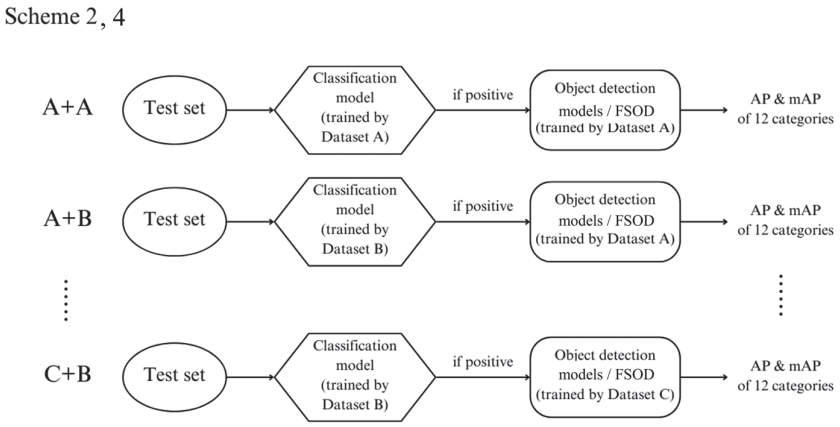


Figure 5. Flowchart for Schemes 2 and 4, illustrating the process for binary classification followed by object detection or few-shot object detection (FSOD).

4.3. Scheme 3: Few-Shot Object Detection

Scheme 3 mirrors Scheme 1 but is focused on the FSOD models. These models were trained on three datasets (A, B, and C), and AP and mAP were calculated for detecting the 12 diseases using the test set.

4.4. Scheme 4: Binary Classification + Few-Shot Object Detection

This scheme is similar to Scheme 2 except that after binary classification, the FSOD models from Scheme 3 are used for further detection. Like Scheme 2, this method generates six possible outcomes: A + A, A + B, B + A, B + B, C + A, and C + B.

4.5. Evaluation Metrics

To assess the performance of the binary classifiers, we used standard metrics: accuracy, precision, recall, and F1-score. For object detection and image segmentation, we used the intersection over union (IoU) metric, which measures the overlap between a predicted bounding box (Pred) and the ground-truth box (GT). The IoU is calculated as follows:

$$\text{IoU} = \frac{|\text{GT} \cap \text{Pred}|}{|\text{GT} \cup \text{Pred}|}, \quad 0 \leq \text{IoU} \leq 1$$

Here, $|\text{GT} \cap \text{Pred}|$ represents the overlapping pixels between the predicted and ground truth boxes, and $|\text{GT} \cup \text{Pred}|$ is the total number of pixels in both boxes. An IoU of 0 indicates no overlap, while an IoU of 1 indicates a perfect match. We set a threshold of 0.5 for this study, meaning that predictions with IoU values above this threshold are considered correct.

In object detection, the mean average precision (mAP) is a key metric for evaluating model performance. It combines precision and recall by calculating the average precision (AP) for each class. For M object classes, the AP for the mth class is calculated as follows:

$$\text{AP}_m = \int_0^1 \text{PR}_m(r) dr$$

where $\text{PR}_m(r)$ is the precision–recall curve for the mth class. To compute precision and recall, predicted boxes are ranked based on confidence scores. If the IoU between a predicted and ground–truth box exceeds the threshold, it is considered a true positive; otherwise, it is a false positive. After calculating precision and recall for all predictions, the precision–recall

curve is plotted, and the area under the curve is calculated for each class. The mAP is then calculated as the average of APs across all classes:

$$\text{mAP} = \frac{1}{M}\sum_{m=1}^{M} \text{AP}_m$$

The mAP score ranges from 0 to 1, with values closer to 1 indicating better model performance in detecting and localizing objects.

5. Results

We present the results from experiments conducted on three custom-designed datasets using four different analysis approaches. Two binary classification models were trained on a P100 GPU, while the object detection and FSOD models were trained using NVIDIA GeForce RTX 3080 and GTX 1080 Ti GPUs.

5.1. Binary Classification

EfficientNet-B3 was trained on Dataset A, and DenseNet121 on Dataset B. These models were used in Schemes 2 and 4 to predict whether an image contained at least one instance of disease. Pretrained ImageNet weights were used, with only the fully connected layer retrained. The hyperparameters are listed in Supplementary Table S1, and images were normalized using the ImageNet mean and standard deviation.

Table 4 presents the performance of these models on the test set. Due to class imbalance, we used both the accuracy and F1-score for evaluation. EfficientNet-B3, trained on Dataset A, outperformed DenseNet121, trained on Dataset B, across all metrics except precision.

Table 4. Performance comparison of the two binary classification models.

	Accuracy	**F1-Score**	**Precision**	**Recall**
EfficientNet-B3 on Dataset A	88.12%	85.85%	84.41%	88.05%
DenseNet121 on Dataset B	86.74%	85.34%	86.44%	84.56%

5.2. Comparison of Analytic Schemes

The training hyperparameters for object detection and FSOD models are provided in Supplementary Tables S2 and S3.

5.2.1. Results on mAP

Figure 6 shows the mAP performance of various object detection models for Schemes 1 to 4 across multiple datasets. Here, AP was calculated using IoU greater than 0.5 (mAP@0.5).

Some key patterns are observed. 1. Overall Performance: Processing test images through a binary classification model before object detection yielded better results. 2. Top Performers: FSCE 10-shot consistently achieved top mAP values, particularly on the C + A dataset, where it peaked at 0.343. YOLOX also performed well, peaking at 0.300 on the B + A dataset, although it did not maintain top performance across all datasets. Dynamic R-CNN and RetinaNet also performed competitively, with notable peaks around 0.26 and 0.261, respectively. 3. Low Performers: The three Transformer-based models consistently achieved very low mAP values across datasets, with DETR and Meta-DETR maintaining a flat trend near zero. 4. Few-Shot Trends: Models trained with a higher number of shots generally performed better, with the TFA and FSCE 10-shot achieving higher mAP than TFA and FSCE 1-shot across datasets. However, the mAP gains from increasing the shot number were not always linear and varied across datasets. 5. Dataset Influence: Training classification models on Dataset B appeared to be more challenging, as evidenced by the lower performance of models when trained on the x + B rather than the x + A dataset.

Performance generally peaked on the C + A dataset, where more models achieved their highest mAP values. 6. Model Stability: Some models, such as Dynamic R-CNN and RetinaNet, exhibited greater stability with relatively smaller fluctuations in mAP across datasets. On the other hand, YOLOX and FSCE had more variability, suggesting that their performance may be more sensitive to dataset characteristics. In summary, Figure 6 shows that processing test images with a binary classification model before object detection generally improved results, with FSCE 10-shot achieving the highest mAP values, especially on the C + A dataset. Models like YOLOX, Dynamic R-CNN, and RetinaNet performed well but showed varying stability across datasets, while Transformer-based models (DETR, Meta-DETR) consistently had low mAP values.

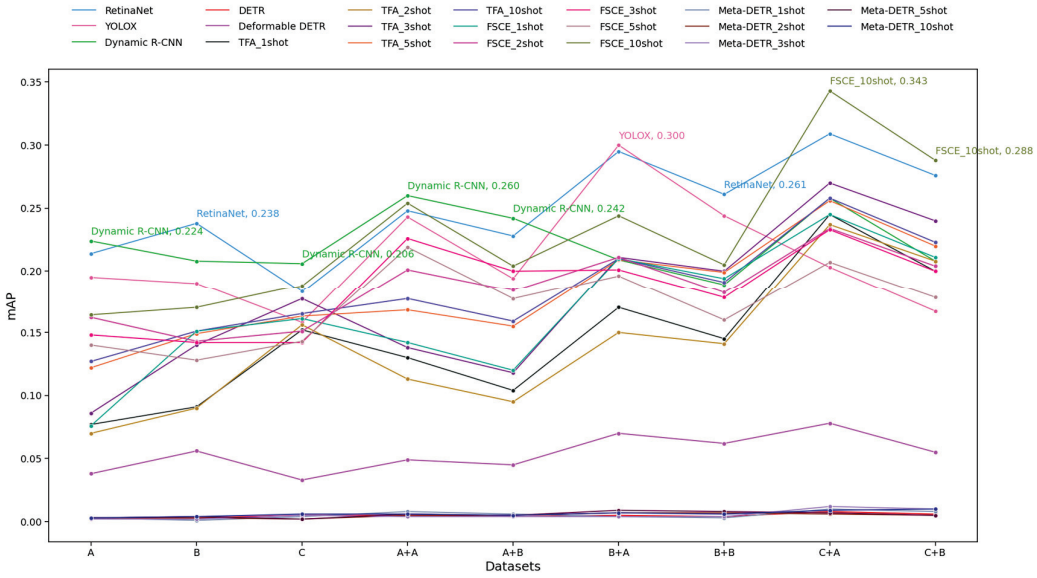


Figure 6. Mean average precision (mAP) results for Schemes 1 to 4. The y-axis represents mAP values, while the x-axis shows the datasets used for model training. "A" refers to training the object detection model on Dataset A (Schemes 1 and 3), while "A + B" indicates first using the classification model trained on Dataset B and then the object detection model trained on Dataset A (Schemes 2 and 4). The same applies to other labels. For the object detection models, each line is the performance of the indicated model on the test set. For the FSOD models, the five lines are the performance for each few-shot fine-tuning scenario (1-shot, 2-shot, 3-shot, 5-shot, 10-shot). The mAP of the best-performing model is shown for each dataset labeled on the *x*-axis.

5.2.2. Results for Base and Novel mAP

In FSOD, the categories were divided into base and novel classes. During base training, only base class bounding boxes were used, with novel classes reserved for few-shot fine-tuning. It was expected that the FSOD models would perform better on novel classes, on account of their fewer samples, while the traditional object detection models would excel on base classes.

Figures 7 and 8 show the mAP for base and novel classes of various object detection models across a series of datasets. First, we discuss base classes. 1. Top Performers: YOLOX achieved the highest base mAP value of 0.339 on the A + A dataset. It consistently ranked among the top-performing models across multiple datasets. RetinaNet also performed strongly, with a peak of 0.335 and high base mAP values across several datasets. It exhibited more consistent performance with smaller fluctuations than YOLOX. Dynamic R-CNN also performed relatively well, although its base mAP values were slightly lower and showed some variability compared to RetinaNet. 2. Low Performers: The three Transformer-based models consistently had very low base mAP values across datasets, with DETR and Meta-DETR remaining almost flat near zero. 3. Few-Shot-Based Model Performance: Higher-shot FSOD models (such as FSCE 10-shot) tended to have better base mAP values

than their lower-shot counterparts but were generally outperformed by traditional object detection models such as YOLOX and RetinaNet on most datasets. 4. Model Stability: RetinaNet and Dynamic R-CNN showed trends of more stable performance, with fewer abrupt changes in base mAP across datasets. YOLOX, while generally performing well, had some larger fluctuations in base mAP, indicating that it may be more sensitive to changes in dataset characteristics. Overall, Figure 7 highlights that YOLOX and RetinaNet performed well and were relatively robust across datasets, while DETR and Meta-DETR consistently underperformed. Models trained with a higher number of shots (e.g., FSCE 10-shot) generally achieved better mAP than lower-shot variants, though not at the level of the top models like YOLOX and RetinaNet.

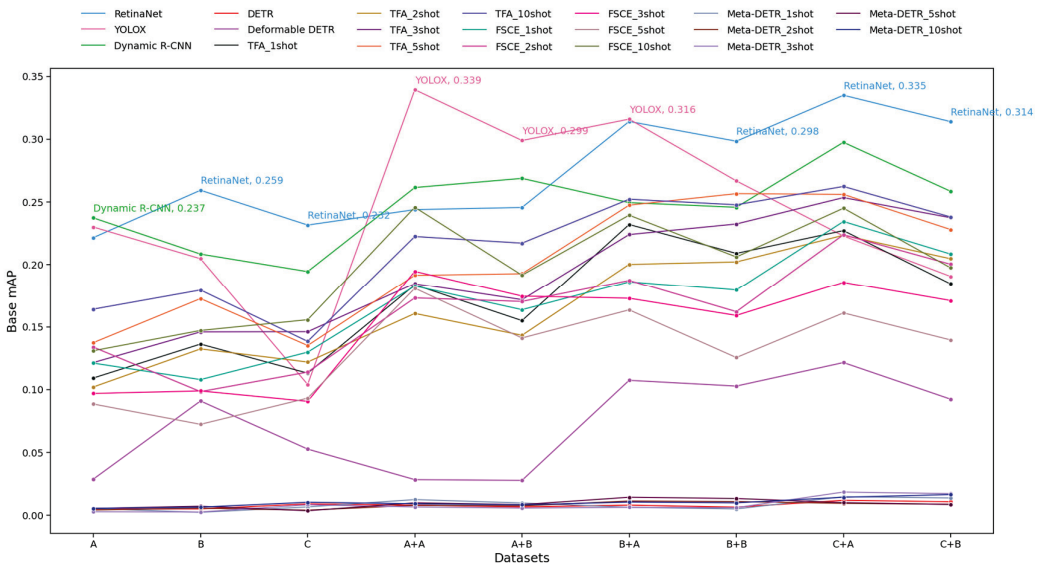


Figure 7. Base mean average precision (base mAP) results for Schemes 1 to 4. The y-axis represents base mAP values, while the x-axis shows the datasets used for model training. "A" refers to training the object detection model on Dataset A (Schemes 1 and 3), while "A + B" indicates first using the classification model trained on Dataset B and then the object detection model trained on Dataset A (Schemes 2 and 4). The same applies to other labels. For the object detection models, each line is the performance of the indicated model on the test set. For the FSOD models, the five lines are the performance for each few-shot fine-tuning scenario (1-shot, 2-shot, 3-shot, 5-shot, 10-shot). The base mAP of the best-performing model is shown for each dataset labeled on the *x*-axis.

Second, we discuss novel classes. (1). Top Performers: FSCE 10-shot was the best-performing model, reaching a peak novel mAP of 0.515 on the C + A dataset and another high of 0.414 on the C + B dataset. This suggests that FSCE 10-shot was particularly effective at handling novel data. Other FSCE variants and YOLOX also performed well, achieving top novel mAP values across multiple datasets. (2). Low Performers: DETR and Meta-DETR continued to exhibit low performance, with novel mAP values near zero across most datasets. The traditional object detection models, such as Dynamic R-CNN and RetinaNet, showed relatively low novel mAP values, suggesting they may not generalize as well to novel classes. (3). Few-Shot Trends: Models trained with a higher number of shots (e.g., FSCE 5-shot and FSCE 10-shot) generally performed better on novel classes than their lower-shot counterparts. The increase in mAP with higher shot numbers suggests that these models benefited from having additional samples to learn novel object detection. (4). Dataset Influence: Training classification models on Dataset B appeared to be more challenging, given the lower performance of models trained on the x + B than on the x + A dataset, and there was a general peak in performance on the C + A dataset, with more models achieving their highest mAP values. In summary, Figure 8 highlights that FSCE (particularly 10-shot) excelled in novel detection tasks, and YOLOX also performed well.

Higher-shot models generally performed better in detecting novel objects, while lower-shot and non-few-shot models struggled, particularly on challenging datasets.

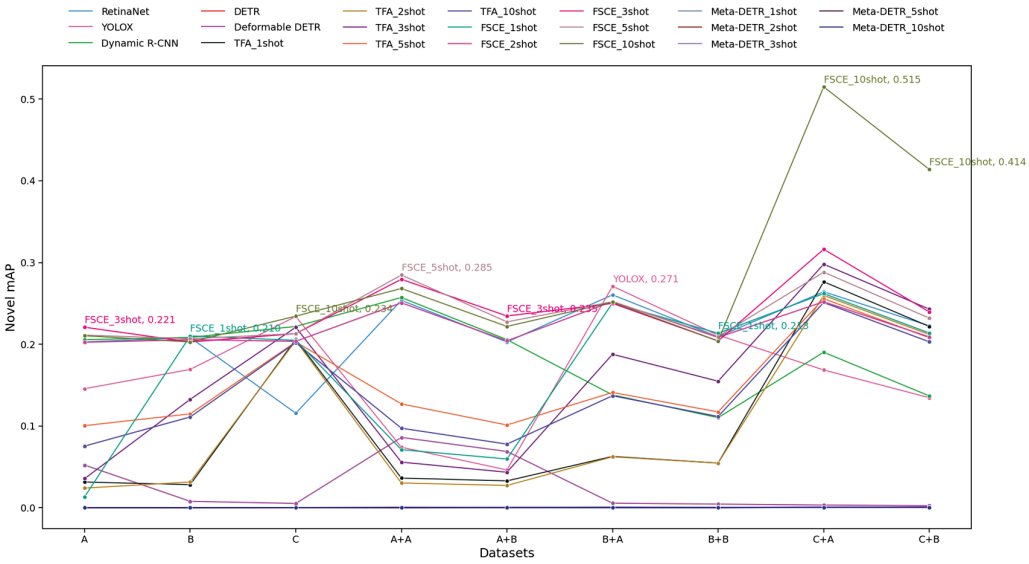


Figure 8. Novel mean average precision (novel mAP) results for Schemes 1 to 4. The y-axis represents novel mAP values, while the x-axis shows the datasets used for model training. "A" refers to training the object detection model on Dataset A (Schemes 1 and 3), while "A + B" indicates first using the classification model trained on Dataset B and then the object detection model trained on Dataset A (Schemes 2 and 4). The same applies to other labels. For the object detection models, each line is the performance of the indicated model on the test set. For the FSOD models, the five lines are the performance for each few-shot fine-tuning scenario (1-shot, 2-shot, 3-shot, 5-shot, 10-shot). The novel mAP of the best-performing model is shown for each dataset labeled on the *x*-axis.

5.2.3. Disease-Wise AP Results

We evaluated model performance for each disease category, as shown in Supplementary Figures S1–S4. For the FSOD models, only the 10-shot fine-tuning results are presented. In Schemes 1 and 2, DETR performed poorly, detecting a few lesions in the categories ArtCur (arterial curvature) and SpiLes (spinal lesions). Similarly, Meta-DETR, a Transformer-based model, underperformed in Schemes 3 and 4, showing only limited lesion detection in the IncLunPat (increased lung patterns) category. By contrast, the CNN-based models performed well in most categories, with object detection models excelling in the categories IncLunPat, SpiLes, CarHyp (cardiac hypertrophy), and HeaPacPla. The FSOD models performed better when the test images were first processed through a classification model. The best prediction results were observed for HeaPacPla, with many models demonstrating strong performance. For the lower-performing category ShaOfPulNod, its AP can be boosted to as high as 1 using FSCE with the combinations C + A or C + B.

We also investigated whether the number of shots used in fine-tuning affects novel class performance. Supplementary Figure S5 shows the results of FSCE, the best-performing FSOD model, for five novel classes: SmaPulNod, ShaOfPulNod, tuberculosis, PulFib, and HeaPacPla. For SmaPulNod and ShaOfPulNod, models were trained on Dataset C with 10-shot fine-tuning outperforming the others. For PulFib, better performance was achieved using 1-shot and 5-shot fine-tuning, while for tuberculosis, Classification Model A misclassified images as normal, preventing AP calculation for the combinations A + A, B + A, and C + A. HeaPacPla achieved perfect predictions with 2- or higher-shot fine-tuning.

5.2.4. Accuracy of Normal Images

To avoid misclassifying normal images as diseased, we calculated the specificity (accuracy of normal images) across the four schemes using confidence score thresholds of 0.3

and 0.5. If a predicted bounding box on a normal image exceeded the threshold, the image was considered misclassified. Figure 9 shows the specificities across all methods. Models that passed test images through a classification model before object detection achieved significantly higher specificity. The three Transformer-based models, despite their lower mAP performance indicated earlier, showed near-perfect specificity. Transformer-based models often excel in capturing the global context due to their self-attention mechanism, which enhances their ability to differentiate between object and non-object regions and thereby more accurately classify normal images, contributing to their higher specificity. However, they struggle with the localization and accurate detection of fine-grained objects and may require more extensive training data or context diversity to achieve high precision and recall, which explains the discrepancy between their strong specificity and weaker mAP. In contrast, models like Dynamic R-CNN, TFA, and FSCE (all based on the faster R-CNN framework), which had higher mAP, tended to show less satisfactory specificity. This may be due to the design and training objectives of the faster R-CNN architecture, where their region proposal networks excel at generating candidate regions to contain objects, but are more prone to misclassify background or non-object regions as objects when confident regions are identified. Thus, these CNN-based models are tuned to prioritize sensitivity in finding objects, potentially sacrificing specificity. The specificity of the two FSOD models, TFA and FSCE, decreased as the number of shots used in fine-tuning increased. YOLOX stood out by achieving almost perfect specificity at the 0.5 threshold, excelling in both mAP and the accuracy of normal image detection.

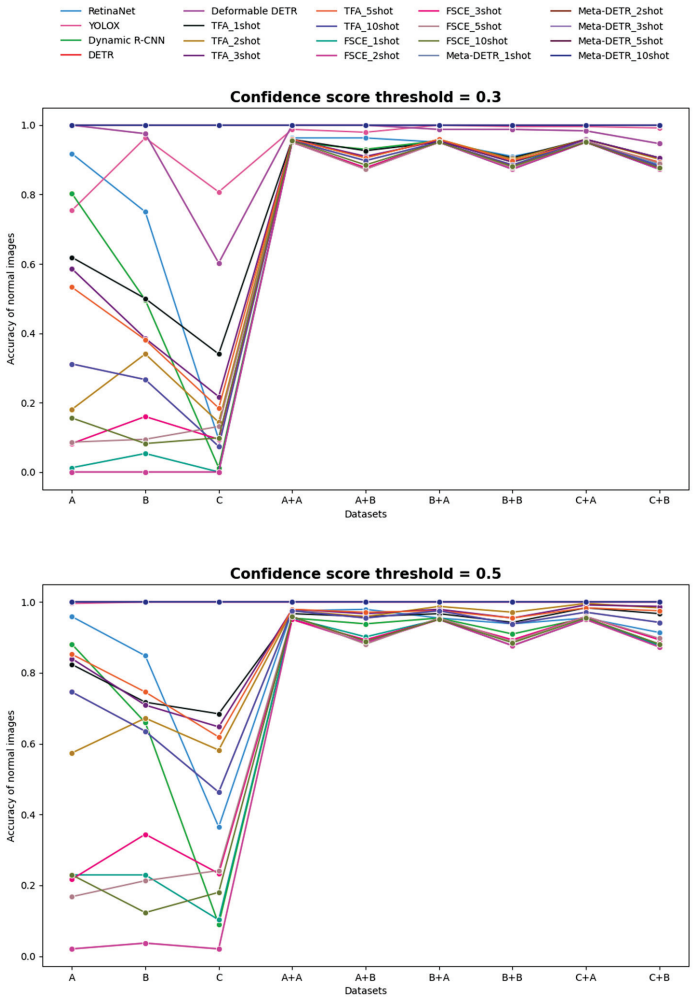


Figure 9. Accuracies of normal images for Schemes 1 to 4. The top graph shows the results using a confidence score threshold of 0.3, while the bottom graph displays the results for a threshold of 0.5.

6. Discussion

In this study, we developed deep learning-based object detection strategies with two main goals: first, to address the class imbalance in the dataset for more accurate predictions, and second, to reduce the false positive rate while maintaining high accuracy. To achieve these objectives, we established several datasets and experimental approaches. The results showed that the CNN-based models consistently outperformed the Transformer-based models, and the proportion of background images in the training sets had a significant effect on the inference capabilities of these models. When comparing the four proposed analytic schemes, we found that Schemes 2 and 4, which first applied a classification model, outperformed Schemes 1 and 3, which relied solely on object detection or the FSOD models. In particular, the best results were obtained using the approach where test images were first processed by Classification Model A and then by FSCE trained on Dataset C with 10-shot fine-tuning.

Despite attempts to balance the data by adjusting the proportion of background images in the training datasets, class imbalance remained an issue. This led to the use of FSOD models in Schemes 3 and 4, which were expected to handle class imbalance better based on their architecture. The experimental results confirmed that the FSOD models outperformed the object detection models for novel classes, while the reverse was true for base classes. Class-wise AP analysis showed that different k-shot fine-tuning settings affected categories in varying ways; more shots did not always result in better performance.

To achieve our second goal, we also calculated the specificity of the test data (i.e., accuracy for normal image detection) across all four schemes. The results indicated that the accuracy of normal images could first be improved by using a binary classification model. Overall, models that excelled in terms of mAP tended to have lower accuracy for normal images, and vice versa. YOLOX was the only model that performed well in terms of both mAP and normal image accuracy.

Studies have shown that CNN architectures are particularly effective in detecting and localizing abnormalities in CXR images [45]. In our study, we also found that CNN-based models, particularly YOLOX and FSCE with 10-shot fine-tuning, achieved the highest mAP scores in disease detection. The limitations posed by small imbalanced annotated datasets in developing deep learning models for localization have been highlighted in previous research and addressed by using transfer learning and augmentation techniques [47], in combination with large datasets containing ground-truth bounding boxes [44]. Similarly, we observed that FSOD techniques, like FSCE, significantly enhanced the accuracy in detecting diseases with limited samples. The use of Transformer-based object detection models in medical image analysis is less common. In our study, we extended the literature by showing that, unlike CNNs, Transformer-based models such as DETR and Meta-DETR, which excel in general object detection tasks on diverse datasets such as COCO, exhibited lower performance in CXR disease detection. While most studies have focused on a single detection model for specific tasks, Behrendt et al. [47] distinguished themselves by evaluating transfer learning, nodule augmentation, and various detection algorithms in building a robust nodule detection system. Our study further contributes by systematically comparing CNN- and Transformer-based object detection algorithms along with FSOD techniques to identify and fine-tune the most suitable deep learning models for various disease detection tasks in CXR images.

The proposed methods for disease detection and localization in CXRs show significant promise, yet there are limitations that affect their robustness and scalability in clinical settings. First, the models relied on a single-institution dataset, which may lead to overfitting and reduce their generalizability to diverse clinical environments with varying imaging protocols and patient populations. Although FSOD techniques were employed to address class imbalance, accurate detection remained a challenge for certain categories of rare diseases with very few examples. Transformer-based models like DETR and Meta-DETR also exhibited limitations in detecting small abnormalities in CXR images, and their high computational demands further limit their feasibility in real-time or resource-constrained

settings. Additionally, the reliance on precise bounding box annotations introduces potential subjectivity, impacting localization accuracy. Future research could focus on enhancing model generalizability by incorporating multi-institutional datasets and reducing class imbalance through data augmentation. Exploring lightweight model architectures or hybrid approaches that integrate CNNs with Transformers could allow for optimizing performance while reducing computational requirements. Improved annotation techniques, such as weak- or self-supervised learning [34] or semi-automated labeling [49,50] may also be used to enhance model training quality and overall detection accuracy, paving the way for more robust and clinically viable AI-based diagnostic tools.

7. Conclusions

This study explored the application of deep learning-based object detection models for disease detection and localization in CXR images. By employing CNN and Transformer-based architectures, as well as FSOD techniques, we developed and evaluated approaches for accurately detecting 12 thoracic diseases across multiple analytic schemes. Our results indicate that the CNN-based models, particularly YOLOX and FSCE (10-shot), consistently achieved high mAP scores, underscoring their robustness and adaptability to clinical settings. In comparison, Transformer-based models such as DETR and Meta-DETR exhibited limitations in small object localization, which may stem from both the model architecture and dataset constraints. Our approach highlights the potential of binary classification as a preliminary step to reduce false positives in object detection, leading to improved disease detection accuracy and specificity for normal images. Furthermore, incorporating FSOD enhanced the capability of our model to handle rare diseases with minimal training samples, suggesting that few-shot learning can be a valuable addition to resource-constrained medical imaging tasks.

While promising results were achieved, certain limitations, including potential overfitting to dataset-specific features, high computational demands, and class imbalances, highlight avenues for future research. These limitations could be addressed through approaches such as cross-institutional validation, lightweight model design, and data augmentation to further enhance model precision and clinical applicability. Ultimately, this work contributes valuable insights to the development of robust automated diagnostic tools, paving the way toward more accurate and efficient CXR disease detection in real-world healthcare settings.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics14232636/s1, Table S1. Hyperparameter settings for the binary classification models. Table S2. Hyperparameter settings for the object detection models. Table S3. Hyperparameter settings for the FSOD models. Figure S1. Disease-wise average precision (AP) values for Scheme 1. Figure S2. Disease-wise average precision (AP) values for Scheme 2. Figure S3. Disease-wise average precision (AP) values for Scheme 3. Figure S4. Disease-wise average precision (AP) values for Scheme 4. Figure S5. Novel class APs of FSCE for different shot settings.

Author Contributions: Conceptualization, G.-H.H. and T.-B.C.; Data curation, T.-B.C., N.-H.L., K.-Y.L. and K.-H.L.; Formal analysis, Y.-C.C., Y.-C.H. and G.-H.H.; Funding acquisition, G.-H.H.; Investigation, T.-B.C., N.-H.L., K.-Y.L. and K.-H.L.; Methodology, G.-H.H., Y.-C.C. and Y.-C.H.; Project administration, G.-H.H.; Resources, T.-B.C.; Software, Y.-C.C. and Y.-C.H.; Supervision, G.-H.H.; Writing—original draft, Y.-C.C., Y.-C.H. and G.-H.H.; Writing—review and editing, G.-H.H., T.-B.C., N.-H.L., K.-Y.L. and K.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by grants from the Ministry of Science and Technology, Taiwan (MOST 111-2118-M-A49-003-MY2), and the National Science and Technology Council, Taiwan (NSTC 113-2118-M-A49-006).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the E-Da Hospital, Kaohsiung, Taiwan (protocol number: EMRP-108-115 and approval date: 20 September 2019).

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

Data Availability Statement: The data used and analyzed in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Author T.-B.C. was employed by the companies Infinity Co., Ltd. and Der Lih Fuh Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ali, S.; Hussain, A.; Bhattacharjee, S.; Athar, A.; Abdullah; Kim, H.-C. Detection of COVID-19 in X-ray images using densely connected squeeze convolutional neural network (DCSCNN): Focusing on interpretability and explainability of the black box model. *Sensors* **2022**, *22*, 9983. [CrossRef] [PubMed]
2. Singh, T.; Mishra, S.; Kalra, R.; Satakshi; Kumar, M.; Kim, T. COVID-19 severity detection using chest X-ray segmentation and deep learning. *Sci. Rep.* **2024**, *14*, 19846. [CrossRef] [PubMed]
3. Garstka, J.; Strzelecki, M. Pneumonia detection in X-ray chest images based on convolutional neural networks and data augmentation methods. In Proceedings of the Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 23–25 September 2020; pp. 18–23.
4. Rana, S.; Hosen, M.J.; Tonni, T.J.; Rony, M.A.H.; Fatema, K.; Hasan, M.Z.; Rahman, M.T.; Khan, R.T.; Jan, T.; Whaiduzzaman, M. DeepChestGNN: A comprehensive framework for enhanced lung disease identification through advanced graphical deep features. *Sensors* **2024**, *24*, 2830. [CrossRef] [PubMed]
5. Javed, R.; Abbas, T.; Khan, A.H.; Daud, A.; Bukhari, A.; Alharbey, R. Deep learning for lungs cancer detection: A review. *Artif. Intell. Rev.* **2024**, *57*, 197. [CrossRef]
6. Xu, Z.; Wang, T.; Skidmore, A.K.; Lamprey, R. A review of deep learning techniques for detecting animals in aerial and satellite images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103732. [CrossRef]
7. Li, Y.; Wang, H.; Dang, L.M.; Nguyen, T.N.; Han, D.; Lee, A.; Jang, I. A deep learning-based hybrid framework for object detection and recognition in autonomous driving. *IEEE Access* **2020**, *8*, 194228–194239. [CrossRef]
8. Zheng, X.; Zheng, S.; Kong, Y.; Chen, J. Recent advances in surface defect inspection of industrial products using deep learning techniques. *Int. J. Adv. Manuf. Technol.* **2021**, *113*, 35–58. [CrossRef]
9. Chandan, G.; Jain, A.; Jain, H.; Mohana. Real time object detection and tracking using deep learning and OpenCV. In Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 11–12 July 2018; pp. 1305–1308.
10. Sethi, S.; Kathuria, M.; Kaushik, T. Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread. *J. Biomed. Inform.* **2021**, *120*, 103848. [CrossRef]
11. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for largescale image recognition. *arXiv* **2014**, arXiv:1409.1556.
14. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385.
16. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. *arXiv* **2016**, arXiv:1608.06993.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, realtime object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
19. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. *arXiv* **2016**, arXiv:1612.08242.
20. Redmon, J.; Farhadi, A. YOLOV3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision-ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9905, pp. 21–37.
22. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
23. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
24. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.

25. 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.* **2017**, *39*, 1137–1149. [CrossRef]
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Computer Vision-ECCV 2020, 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I. Springer: Cham, Switzerland, 2020; pp. 213–229.
27. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
28. Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L.-J.; Li, F.-F. Thoracic disease identification and localization with limited supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8290–8299.
29. Nguyen, N.H.; Nguyen, H.Q.; Nguyen, N.T.; Nguyen, T.V.; Pham, H.H.; Nguyen, T.N.-M. Deployment and validation of an AI system for detecting abnormal chest radiographs in clinical settings. *Front. Digit. Health* **2022**, *4*, 890759. [CrossRef]
30. Alaskar, H.; Hussain, A.; Almaslukh, B.; Vaiyapuri, T.; Sbai, Z.; Dubey, A.K. Deep learning approaches for automatic localization in medical images. *Comput. Intell. Neurosci.* **2022**, *2022*, 6347307. [CrossRef] [PubMed]
31. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common objects in context. *arXiv* **2014**, arXiv:1405.0312.
32. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2419–2428.
33. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta R-CNN: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9577–9586.
34. Huang, G.; Laradji, I.; Vazquez, D.; Lacoste-Julien, S.; Rodriguez, P. A survey of self-supervised and few-shot object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4071–4089. [CrossRef] [PubMed]
35. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
36. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
37. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.
38. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards high quality object detection via dynamic training. *arXiv* **2020**, arXiv:2004.06002.
39. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv* **2020**, arXiv:2003.06957.
40. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. FSCE: Few-shot object detection via contrastive proposal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7348–7358.
41. Zhang, G.; Luo, Z.; Cui, K.; Lu, S.; Xing, E.P. Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12832–12843. [CrossRef]
42. Santosh, K.C.; Dhar, M.K.; Rajbhandari, R.; Neupane, A. Deep neural network for foreign object detection in chest X-rays. In Proceedings of the IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; pp. 538–541.
43. Kim, Y.G.; Lee, S.M.; Lee, K.H.; Jang, R.; Seo, J.B.; Kim, N. Optimal matrix size of chest radiographs for computer-aided detection on lung nodule or mass with deep learning. *Eur. Radiol.* **2020**, *30*, 4943–4951. [CrossRef]
44. Wang, S.; Wang, G.; Xia, Y.; Wu, Q.; Fan, X.; Chen, X.; He, M.; Xiao, J.; Yang, L.; Liu, Y.; et al. A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest X-ray. *Nat. Commun.* **2024**, *15*, 1347.
45. Çallı, E.; Sogancioglu, E.; van Ginneken, B.; van Leeuwen, K.G.; Murphy, K. Deep learning for chest X-ray analysis: A survey. *Med. Image Anal.* **2021**, *72*, 102125. [CrossRef]
46. Nguyen, H.Q.; Lam, K.; Le, L.T.; Pham, H.H.; Tran, D.Q.; Nguyen, D.B.; Le, D.D.; Pham, C.M.; Tong, H.T.T.; Dinh, D.H.; et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci. Data* **2022**, *9*, 429. [CrossRef] [PubMed]
47. Behrendt, F.; Bengs, M.; Bhattacharya, D.; Krüger, J.; Roland Opfer, R.; Alexander Schlaefer, A. A systematic approach to deep learning-based nodule detection in chest radiographs. *Sci. Rep.* **2023**, *13*, 10120. [CrossRef] [PubMed]
48. NODE21 Competition. Available online: https://node21.grand-challenge.org/ (accessed on 12 November 2024).
49. Wu, J.; Gur, Y.; Karargyris, A.; Syed, A.B.; Boyko, O.; Moradi, M. Automatic bounding box annotation of chest X-ray data for localization of abnormalities. In Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 799–803.
50. Zhao, J. CrossEAI: Using explainable AI to generate better bounding boxes for chest X-ray images. *arXiv* **2023**, arXiv:2310.19835.

diagnostics

MDPI

Article

Enhancing Retina Images by Lowpass Filtering Using Binomial Filter

Mofleh Hannuf AlRowaily [1], Hamzah Arof [1,*], Imanurfatiehah Ibrahim [1], Haniza Yazid [2] and Wan Amirul Mahyiddin [1]

[1] Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia; mofleh2@hotmail.com (M.H.A.); imanurfatiehah@gmail.com (I.I.); wanamirul@um.edu.my (W.A.M.)

[2] Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis, Ulu Pauh Campus, Arau 02600, Malaysia; hanizayazid@unimap.edu.my

* Correspondence: ahamzah@um.edu.my

Abstract: This study presents a method to enhance the contrast and luminosity of fundus images with boundary reflection. In this work, 100 retina images taken from online databases are utilized to test the performance of the proposed method. First, the red, green and blue channels are read and stored in separate arrays. Then, the area of the eye also called the region of interest (ROI) is located by thresholding. Next, the ratios of R to G and B to G at every pixel in the ROI are calculated and stored along with copies of the R, G and B channels. Then, the RGB channels are subjected to average filtering using a 3×3 mask to smoothen the RGB values of pixels, especially along the border of the ROI. In the background brightness estimation stage, the ROI of the three channels is filtered by binomial filters (BFs). This step creates a background brightness (BB) surface of the eye region by levelling the foreground objects like blood vessels, fundi, optic discs and blood spots, thus allowing the estimation of the background illumination. In the next stage, using the BB, the luminosity of the ROI is equalized so that all pixels will have the same background brightness. This is followed by a contrast adjustment of the ROI using CLAHE. Afterward, details of the adjusted green channel are enhanced using information from the adjusted red and blue channels. In the color correction stage, the intensities of pixels in the red and blue channels are adjusted according to their original ratios to the green channel before the three channels are reunited. The resulting color image resembles the original one in color distribution and tone but shows marked improvement in luminosity and contrast. The effectiveness of the approach is tested on the test images and enhancement is noticeable visually and quantitatively in greyscale and color. On average, this method manages to increase the contrast and luminosity of the images. The proposed method was implemented using MATLAB R2021b on an AMD 5900HS processor and the average execution time was less than 10 s. The performance of the filter is compared to those of two other filters and it shows better results. This technique can be a useful tool for ophthalmologists who perform diagnoses on the eyes of diabetic patients.

Keywords: retina images; luminosity; contrast; boundary reflection; binomial filter (BF)

Citation: AlRowaily, M.H.; Arof, H.; Ibrahim, I.; Yazid, H.; Mahyiddin, W.A. Enhancing Retina Images by Lowpass Filtering Using Binomial Filter. *Diagnostics* **2024**, *14*, 1688. https://doi.org/10.3390/diagnostics14151688

Academic Editor: Antonio Ferreras

Received: 9 July 2024
Revised: 26 July 2024
Accepted: 29 July 2024
Published: 5 August 2024

1. Introduction

Retina images significantly contribute to the diagnosis of various eye-related conditions including diabetic retinopathy [1], macula edema [2], neovascularization [3] and glaucoma [4]. They are not only used for diagnosis but are also instrumental in the monitoring of ocular diseases as changes and progression are observed using these images. Ophthalmologists look for peculiar structures or changes in retina images to identify present or impending problems. Although ophthalmologists can analyze retina images effectively, it is a time-consuming process and when there are many patients, inaccurate diagnosis may occur due to lapses, fatigue or carelessness. Thus, computer-assisted diagnosis of retina images is an indispensable tool that can expedite the process and increase

its accuracy [5–10]. Retina image quality is important in safeguarding the efficiency of manual or computer-assisted diagnosis. Retina images that have varying illumination, low contrast and haziness may cause incorrect interpretation by ophthalmologists and affect the performance of automated analysis systems [5,6].

Images affected by contrast and luminosity unevenness can be corrected by many enhancement approaches proposed by researchers [11,12]. A comprehensive review of all image enhancement techniques is not available, but compilations of well-known enhancement methods in specific areas have been presented by researchers from time to time. Schuch et al. focused on reviewing image enhancement of fingerprint images. Various approaches were reviewed and classified into several categories or models [13]. A review by Saba et al. [14] concentrates on image enhancements of knee joint images where the techniques are divided into frequency and spatial classes. A survey of general enhancement methods by Singh and Mittal also classifies the approaches into frequency and spatial groups [15]. Methods that belong to the spatial domain manipulate the pixels in images by transforming or altering them directly. Many techniques belong to this group inclusive of magnitude and log transforms, alpha rooting and histogram-based operations. Frequency-related techniques perform discrete transformations on entire images using Fourier, cosine, wavelet or other bases. Then, the transformed images are processed and reverted to the original image space [16].

In practice, the image enhancement methods are chosen according to the issue to solve, the image model used, the software platform, available toolboxes and the database. Of late, there have been reported works on using neural networks with deep learning for noise removal and image enhancement [15–17]. There are also attempts to use metaheuristic optimizations to fine-tune the values of parameters used in established approaches [18,19]. The proposed work can be considered a spatial method where a BF is employed to improve retina images in the RGB domain. Information from all the three channels of the color images is used to enhance the region of interest (ROI).

Several image enhancement methods for color or grayscale images have been implemented on retinal images. When processing color images in RGB space, many researchers utilize only the green channel [20,21]. Foracchia et al. developed a statistical technique to normalize the background of retina images based on some statistical parameters [22]. Coa et al. implemented a method like the retinex algorithm to enhance retina images. There are four primary procedures comprising padding, contrast improvement, grayscale adjustment and refinement. The background padding is implemented to avoid over-enhancing the retinal boundary. Then all low-frequency components are filtered in the root domain. The visual color of the image changes significantly once the contrast is adjusted. Then, to regain the initial image color, every channel is recalibrated. The contrast is then further enhanced using a refinement step [23]. In earlier works, only the green channel is used to locate abnormalities that are present retinal images [24]. But lately, retina images are processed in color.

Zhao et al. processed the B, G and R color channels of retina images to enhance them. They found that the luminosity and color of the images were correlated [25]. The G, R and B channels should be changed in the same ratio to preserve color integrity. Rao et al. transformed fundus images into HSV and then L*a*b* space, to modify the corresponding V and L components using gamma and CLAHE, respectively [26]. On the other hand, Mitra et al. converted the images into HIS space before equalizing the histogram of the intensity I. The suggested max–min color correction strategy is used to acquire the final image with suitable color [27].

Bala et al. used an adaptive histogram equalization with curvelet and claimed that the method generated better results [28]. Again, some researchers transformed retina images into L*a*b space to enhance the contrast of the L component using CLAHE. They scaled the output range by adopting Hubbard's retinal image brightness standard [29,30]. Qureshi et al. used a different space called CIECAM02 to obtain its brightness component J. Then, the J value was assigned using non-linear mapping based on threshold. They asserted that

the results were better than histogram-based approaches, but in some cases, a greenish color shade was introduced [31].

In 2021, Alwazzan et al. utilized the green channel of color images to remove noise from them using the Wiener filter. CLAHE was also needed to improve their contrast and then reunite them with the unaltered blue and red components [32]. Coa et al. presented a new approach by using an intensity transfer strategy. Initially, the image will undergo contrast stretching. Then, an efficient intensity transfer approach is presented that can deliver the necessary illumination for a single channel. This procedure requires a guided image, which could be the original image or another image [33]. Kumar et al. converted fundus into HSV and L*a*b spaces and performed histogram equalization on their intensity components, V and L, respectively. Then, the contrasted results are recombined according to the given weights. The color scheme remained because other components were not touched [34].

Based on the review, there are lots of methods that have been proposed. Retina images can be enhanced either in grayscale or color. Most of the researchers utilized the green channel since the green channel contains more useful information when compared to red and blue channels. Meanwhile, color images also show significant improvement after the enhancement process. Several color spaces have been used to isolate the intensity or luminance component so that traditional enhancement techniques can be applied to it without affecting the color integrity. Sometimes, retina images contain boundary reflection besides the usual contrast and luminosity variability. The bright or hazy reflection occludes the area underneath. In this work, a method to alleviate the problem is presented. The method relies on estimating the background brightness of an image and the subsequent compensation of its non-uniformity. Both the contrast and luminosity of grayscale and RGB color images are enhanced to improve their appearance and statistical parameters.

The rest of this paper is composed of the following sections. Section 2 briefly reviews the materials and methods of this study. This is followed by the experimental results, and a discussion of the work is provided in Section 4.

2. Materials and Methods

In the acquisition stage, the fundus images are usually unevenly illuminated causing contrast and luminosity to vary throughout the ROI. Sometimes they contain boundary reflection as shown in Figure 1. Boundary reflection can obscure abnormalities and other foreground objects like exudates and blood vessels under its locality.

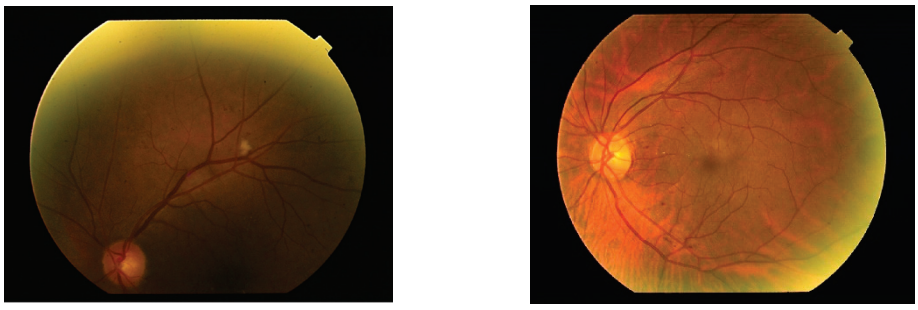


Figure 1. Two samples of retina images with boundary reflection.

In this work, a method that equalizes the underlying luminosity and enhances the contrast of retina images with boundary reflection is proposed. The main idea is to estimate the background brightness of each pixel in the eye region (ROI) using a binomial filter (BF). This step creates a background brightness (BB) surface for the ROI. Based on the BB, luminosity correction is performed by equalizing the background luminosity of all pixels in the ROI to 128 so that they experience the same brightness. This is followed by contrast adjustment using CLAHE. The approach is implemented in 4 stages involving pre-filtering, background brightness estimation, pixel intensity adjustment, and color correction.

Figure 2 shows the flow of stages in the process and the algorithm of the proposed method is presented in Algorithm 1.

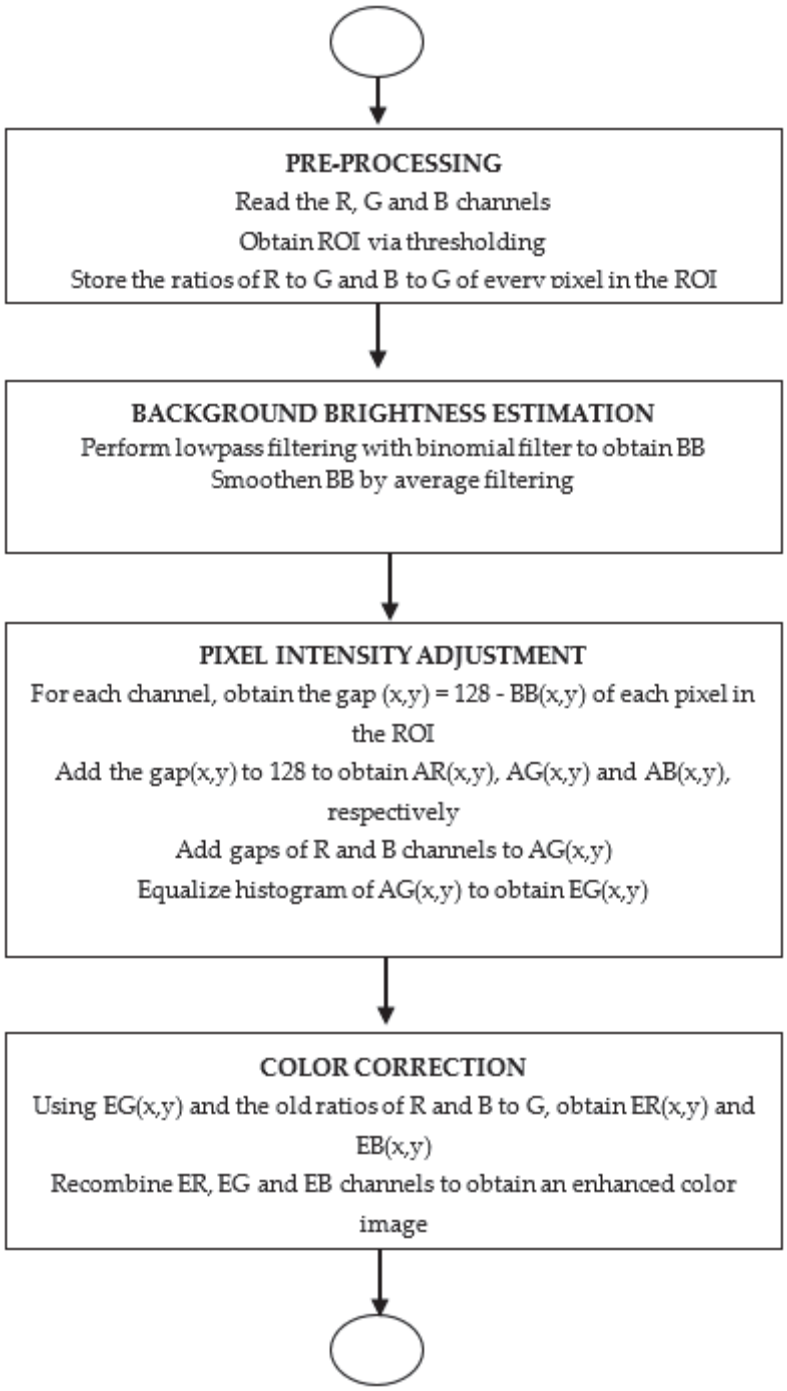


Figure 2. Flow of stages in the enhancement process.

First, the red, green and blue (RGB) channels of the retina image are read and stored. Before further processing, the ratios of R to G and B to G at each pixel location in the ROI are computed and stored. These ratios are needed for color correction later. Then, the ROI and its border are localized by thresholding. Once the ROI is available, it can be made symmetrical to the vertical center line and the area outside of the ROI can be set to zero. Often, in an image that is affected by boundary reflection, the reflection center is located along its border. Then, the RGB channels are subjected to average filtering using a 3×3 mask to smoothen the RGB values of pixels, especially along the border of the ROI.

In the background brightness estimation, the ROI of the three channels of an input image is processed one by one. For every pixel in the ROI, a $w \times w$ neighborhood centered at the pixel is established. There are w^2 pixels in the neighborhood (N), including the center pixel. If the center pixel is located near or on the border of the ROI, some of the pixels in the neighborhood (N) will fall outside of the ROI. The pixels in N that lie outside of the ROI are not convolved with the filter coefficients and thus their multiplications are dropped. This step creates the smooth background brightness (BB) surface given by Equation (1).

$$BB(x,y) = \frac{1}{D} \sum_{n=-w}^{w} \sum_{m=-w}^{w} i(x-m,\ y-n)b(m,n) \tag{1}$$

where D is the sum of all binomial coefficients in N that are involved in the convolution

$$D = \sum_{n=-w}^{w} \sum_{m=-w}^{w} b(m,n) \tag{2}$$

such that

$$\sum_{m,n\ \epsilon\ N} \frac{b(m,n)}{D} = 1 \tag{3}$$

Finally, the result is smoothened by average filtering with a 3×3 average filter. The output is a slow varying surface of the background brightness $BB(x,y)$. The brightness surfaces of the red, green and blue channels are designated as $BBR(x,y)$, $BBG(x,y)$ and $BBB(x,y)$, respectively.

The background brightness (BB) approximates the underlying luminosity for each pixel in the ROI. It is a smooth surface that is free of foreground objects. Note that for each pixel, its blue, green and red values, $B(x,y)$, $G(x,y)$ and $R(x,y)$, can be higher or lower than its respective background values provided by the $BBB(x,y)$, $BBG(x,y)$ or $BBR(x,y)$. The difference between the value of a channel and its background brightness is called the channel gap. The adjustment strategy is to level the underlying luminosity of each pixel to the same value of 128 as follows.

Algorithm 1: Algorithm of the proposed method

Read input image
Separate R, G and B channels
Calculate ratios of R and B to G
Threshold the channels to obtain the ROI
Obtain $BB(x,y)$ of each channel by convolving the ROI with BF
for each pixel in the ROI of each channel
convolve with 41×41 BF
endSmoothen the $BB(x,y)$ by 3×3 averaging filter
Obtain the adjusted surface for each channel
for every pixel
$AR(x,y) = 128 + R(x,y) - BBR(x,y)$
$AG(x,y) = 128 + G(x,y) - BBG(x,y)$
$AB(x,y) = 128 + B(x,y) - BBB(x,y)$
end
Further improve the green channel
$AGF(x,y) = 128 + R(x,y) + G(x,y) + B(x,y) - BBR(x,y) - BBG(x,y) - BBB(x,y)$
Equalize $AGF(x,y)$ by CLAHE to obtain $EG(x,y)$
Obtain the new R, G and B using $EF(x,y)$ and the stored ratios of R and B to G
$Rnew(x,y) = R/G \times EG(x,y)$
$Bnew(x,y) = B/G \times EG(x,y)$
$Gnew(x,y) = EG(x,y)$
end procedure

For every pixel (x,y) of the ROI, the background brightness is subtracted from the pixel value to obtain the gap. For instance, for the red channel $R(x,y)$ and its background brightness $BBR(x,y)$, the gap is $R(x,y) - BBR(x,y)$. It can be zero, positive or negative. The adjusted red, $AR(x,y)$, is formed by adding the gap to 128. Then the same adjustment is made to the green and blue channels such that

$$AR(x,y) = 128 + R(x,y) - BBR(x,y) \tag{4}$$

$$AG(x,y) = 128 + G(x,y) - BBG(x,y) \tag{5}$$

$$AB(x,y) = 128 + B(x,y) - BBB(x,y) \tag{6}$$

The last three equations show that at each location (x,y), the gap is added to the base level of 128. In other words, for each pixel, the gap is maintained although the background brightness is shifted to 128. Throughout the ROI of the adjusted images, the background luminosity appears to be uniformly distributed. After adjusting the background brightness of the ROI to 128, the channels seem to have lost some contrast due to the brighter background.

The green channel is key to the color correction step of the image in the post processing stage. As such, it should contain the most information since the other two channels will be adjusted based on their ratios to the green channel. From observation, the red channel is more volatile than the green or blue, as most foreground objects contain high values of red. The blue channel is the most stable and it captures the background very well. The green channel usually fluctuates in tandem with the red channel but with less variation. The gaps of the red and blue channels to their respective background brightness surfaces can be added to the green channel to improve its contrast and stability. So, Equation (5) becomes

$$AGF(x,y) = 128 + R(x,y) + G(x,y) + B(x,y) - BBR(x,y) - BBG(x,y) - BBB(x,y) \tag{7}$$

where $AGF(x,y)$ stands for the adjusted green channel with full information. For retina images, this step helps improve the appearance and content of the green channel. Figure 3 shows the green channel $G(x,y)$ of an image, its brightness surface $BBG(x,y)$ and its fully adjusted form $AGF(x,y)$. The size of the BF used is 41×41 and its coefficients are scaled so that the highest coefficient is 100 and the lowest is bottom-limited at 1. The 41×41 size is chosen because it produces decent results and is not too slow to implement.

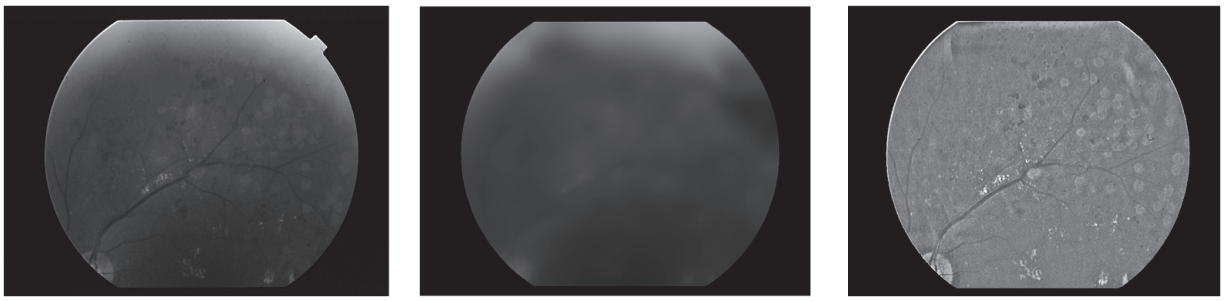


Figure 3. The green channel of an image, its background brightness and fully adjusted form.

Then, the contrast of the $AGF(x,y)$ is further improved by applying contrast-limited adaptive histogram equalization (CLAHE). The operation is implemented on the entire ROI. A constraint is imposed on maintaining the number of grey levels in the image. Due to the constraint, the histogram of the ROI is mainly stretched to improve the contrast of the background and the foreground of the image. The frequency of any value is limited to twice the value of the uniform distribution to restrict the contrast. Another constraint imposed is the gap between consecutive pixel values in the histogram. Each gap can be

expanded but should not be reduced to avoid the loss of details. Due to these constraints, the improvement caused by the adaptive equalization to $AGF(x,y)$ is not significant. The outcome of the equalization step is $EG(x,y)$. The red $AR(x,y)$ and blue $AB(x,y)$ can be equalized using CLAHE too.

If the equalized channels are reunited directly, the resulting color image will have peculiar color tones that do not correspond well to the original image. Therefore, the red $AR(x,y)$ and blue $AB(x,y)$ can be disregarded because they are deemed unsuitable for recombination. The new red and blue channels are the stored ratios of B and R over G multiplied by the equalized green channel $EG(x,y)$. As for the new green channel, it is simply $EG(x,y)$ itself since it is the anchor in this color correction. Thus, the new R, B and G channels are given by

$$R_{new}(x,y) = R/G \times EG(x,y) \tag{8}$$

$$B_{new}(x,y) = B/G \times EG(x,y) \tag{9}$$

$$G_{new}(x,y) = EG(x,y) \tag{10}$$

Now, $R_{new}(x,y)$, $B_{new}(x,y)$ and $G_{new}(x,y)$ can be recombined to form a corrected color image.

3. Results

The databases which consist of 100 retina images were obtained from the online databases of e-OPHTHA, DIARETDB and EyePACS. About half of the images were selected because they contain strong boundary reflections caused by over exposure. The rest were chosen randomly. Most of the images were resized to make them uniform at approximately 500 × 700. The 1D binomial filter was constructed by convolving vector [1 1] to itself consecutively several times to obtain the desired length (Burt P. (1981)). The way to construct a filter with length 2w + 1 was to perform a cascaded convolution of the [1 1] vector to itself 2w times. After the cascaded convolution, the coefficients were divided by the sum of all coefficients in the filter. The 2D binomial filter was obtained from the vector outer product of a 1D binomial filter to itself. If the 1D binomial filter B(x) is regarded as a column vector with length 2w + 1, taking the cross product of B(x) to its transpose B(x)' produces a 2D binomial filter B(m,n) whose size is (2w + 1) × (2w + 1).

It seemed that the peaky binomial filter was the ideal choice to capture the background brightness of retina images because it was high at the center. Convolving an image with the filter resulted in a $BS(x,y)$ that followed the fluctuation of the data closely but did not capture the underlying trend of the luminosity variation very well. As the size increased, its peakiness became more acute. Hence, the BF coefficients were scaled so that the maximum coefficient is 100 and the minimum is lower-limited at 1. The binomial filter was flattened further by multiplying it with α, where $0.1 < \alpha \leq 1$. Figure 4 shows the samples in Figure 1 that have been processed by a BF whose size is 41 × 41 multiplied with ab α of 1, 0.5, 0.2, 0.1 and 0.01. Take note that when $\alpha = 1$, the BF is unflattened, and at $\alpha = 0.01$, all coefficients of the BF become one.

It is seen that the output quality improves as α decreases. The 41 × 41 BF with $\alpha = 0.1$ or 0.01 generates good results for all three samples. The filter with $\alpha = 0.1$ was chosen for further testing as it was not too flat. Thus, this filter was tested on eight samples and changes in contrast and luminosity were calculated. Its performance was compared to those of median and Gaussian filters. All of the filters had the same size of 41 × 41 and they were executed in the same framework so that the only factor that influenced the results was the filter used. The median filter was implemented using a histogram of pixel values. The Gaussian filter was derived from the standard formula with $\sigma = 5$ so that it was not too peaky and then scaled by 10 so that the maximum coefficient value was 10 and the minimum value was limited to 1. Figure 5 shows the results produced by the three filters.

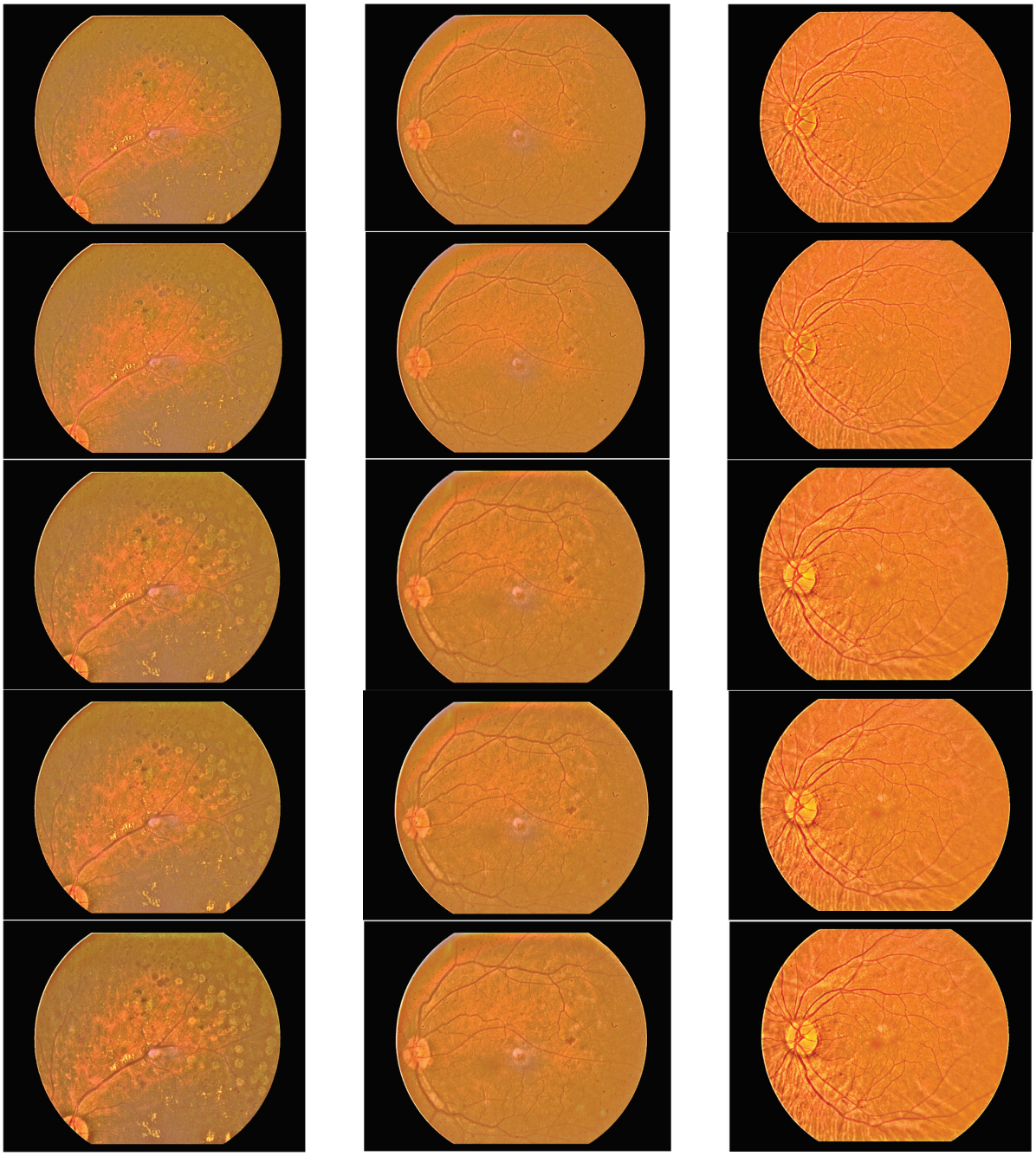


Figure 4. Three retina images processed by BF with α of 1, 0.5, 0.2, 0.1 and 0.01, row wise from top to bottom.

The results look similar as the filters were implemented in the same framework. All images show enhancement in luminosity and contrast throughout the ROI, even in the areas that are partially occluded by reflection. The improvement can be verified visually and quantitatively. The luminosity of the images was obtained by converting them from RGB into L*a*b space, where L stands for luminosity. The average luminosity gain of an image was calculated using equation 11 to estimate the improvement introduced by the proposed method.

$$Luminosity\ Gain = \frac{Avg\ Filtered\ Luminance - Avg\ Unfiltered\ Luminance}{Avg\ Unfiltered\ Luminance} \times 100\% \quad (11)$$

The contrast was estimated using the metric of Matkovic et al. (2005). The local contrasts were the averages of absolute differences of one pixel and its eight nearest neighbors at different resolutions. The global contrast factor (GCF) for the ROI was the aggregate of all local contrasts multiplied by defined weightings. In the experiments, three local contrasts were calculated at three resolutions, and they were multiplied by three weightings of 0.12, 0.142 and 0.154, respectively. These weightings were suggested by the authors themselves. The GCF, or contrast, was calculated before and after filtering. The contrast gain for the images was given by Equation (12) below.

$$Contrast\ Gain = \frac{Avg\ Filtered\ Contrast - Avg\ Unfiltered\ Contrast}{Avg\ Unfiltered\ Contrast} \times 100\% \quad (12)$$

The contrast and luminosity gains for the eight samples and their averages are presented in Table 1. The figures show that on average, contrast and luminosity gains for the samples are nearly identical for the three filters. Tests carried out on the remaining 92 images revealed similar outcomes with nearly the same figures. Comparatively, the BF performed slightly better than the other two filters. Therefore, it is fair to conclude that the contrast and luminosity of the images have been enhanced by the framework.

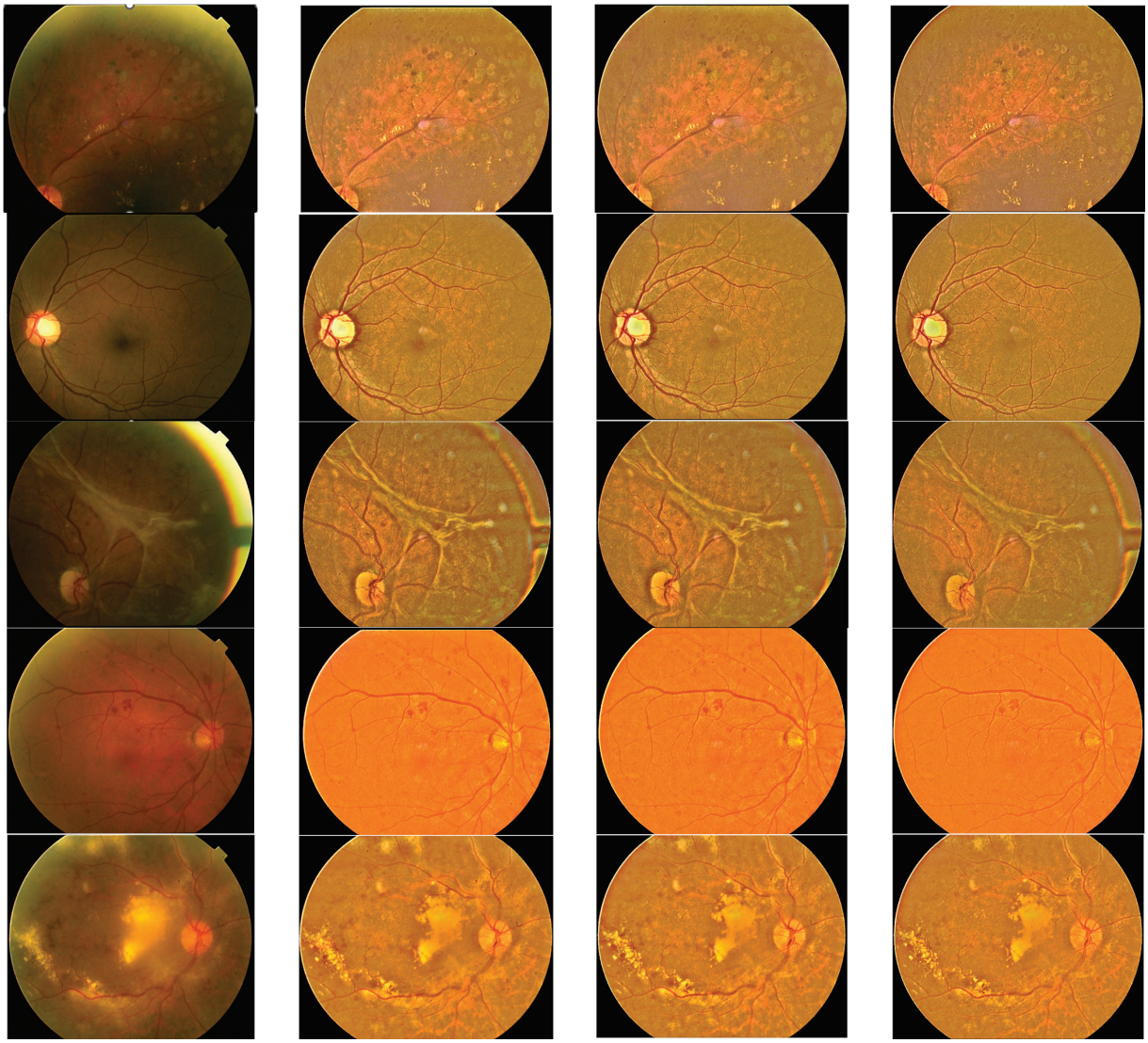
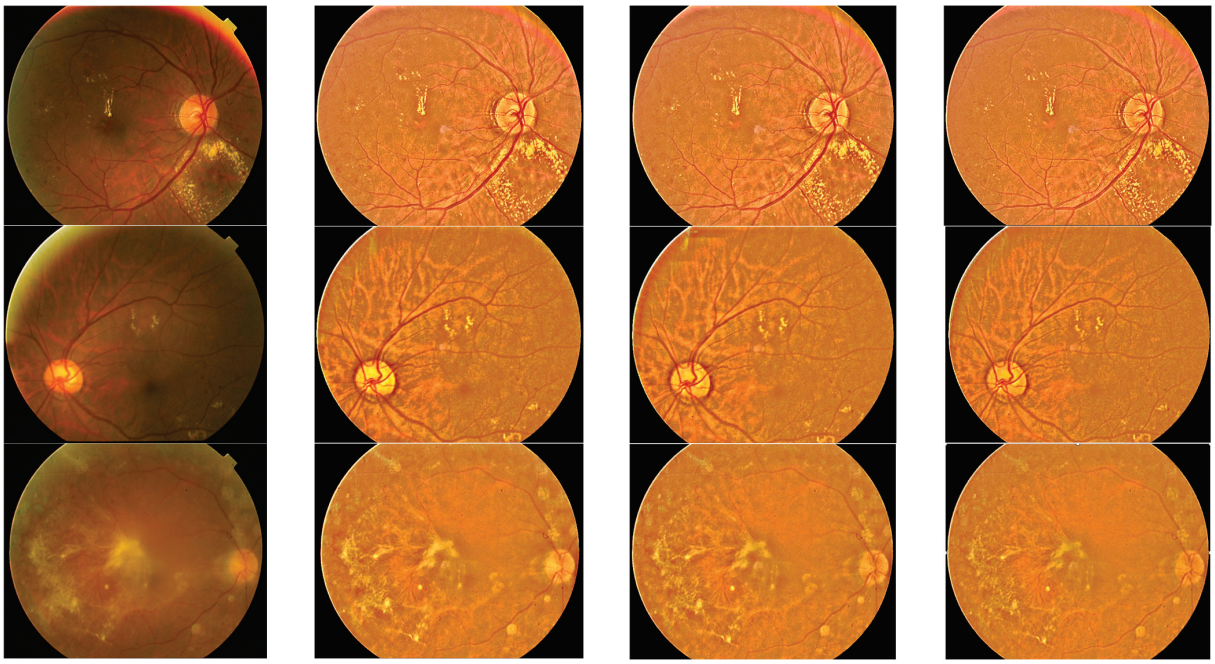


Figure 5. *Cont.*



(**a**) (**b**) (**c**) (**d**)

Figure 5. The original, binomial-filtered, median-filtered and Gaussian-filtered samples, column-wise from left to right. (**a**) Original, (**b**) binomial, (**c**) median and (**d**) Gaussian.

Table 1. The luminance (L) and contrast (C) of the samples before and after filtering.

Sample	Before Filtering		Binomial Filtered		Median Filtered		Gaussian Filtered	
	L	C	L	C	L	C	L	C
1	33.0	0.37	57.6	1.05	57.4	1.05	57.5	1.04
2	32.3	0.48	56.8	1.40	56.6	1.27	56.7	1.35
3	30.5	0.38	57.3	1.21	57.7	1.08	56.9	1.18
4	37.15	0.39	64.1	1.13	64.3	1.12	64.2	1.09
5	43.0	0.50	59.43	1.21	60.0	1.23	59.0	1.20
6	35.39	0.62	60.7	1.69	60.5	1.58	60.3	1.66
7	33.48	0.38	60.2	1.11	60.22	1.03	60.4	1.05
8	42.57	0.42	59.7	1.22	59.6	1.10	59.6	1.13
Average	35.92	0.44	59.44	1.25	59.3	1.18	59.3	1.21
Gain	-	-	65.4%	184.6%	65.3%	168.2%	65.2%	175%

4. Discussion and Conclusions

It is observed that the test images in Figure 5 contain various levels of contrast and illumination variations. Four of them appear to have boundary reflections of various degrees. After processing, marked improvements in contrast and luminosity are observed in all images. The method manages to remove the boundary reflection and expose the surface with some foreground objects. The second and last samples suffer from a dim appearance, but their corrected versions are brighter. The third image has an intense boundary reflection that splits into two parts. The bottom part is small enough that it might be mistaken as a foreground object. In its corrected form, the boundary reflection is removed but traces around its edge remain. This is because there exists a strong transition of red and yellow around its edge which looks like rainbow stripes. The method considers them as foreground objects and decides to keep them. The same observation applies to sample 6, where a red stripe remains near its top border.

The performance of the method was further tested on the remaining 92 test images of the database and the average gains in contrast and luminosity for the binomial filter were

above 170% and 60%, respectively. The median and Gaussian filters achieved slightly less gain for contrast and about the same gain for luminosity. After recombination, the resulting color images resembled the inputs in color tones and distribution, but they showed marked improvement in contrast and luminosity. The framework was executed on MATLAB R2021b powered by an AMD 5900HS processor. The average execution times for the three filters were less than 10 s.

The method works best when the boundary reflection varies smoothly. If the reflection has a steep transition at its border, it is regarded as a foreground object that needs to be preserved. On the contrary, if a large exudate happens to be located at the boundary of the ROI and if its border has a smooth transition, it can be misconstrued as boundary reflection. In this case, it might be eliminated.

In summary, compared to the original images, marked improvements in contrast and luminosity are observed in all processed images. In the boundary areas that are previously shaded by the reflection, the method manages to expose the surface underneath. Overall, the processed images look better than their raw counterparts. The luminosity of the filtered images looks uniform, and the original color shades of the ROI are well preserved. All the foreground objects, including exudates, blood spots, optic discs and blood vessels, are clearly contrasted.

Author Contributions: This study was conceptualized by M.H.A. and H.A.; H.A. developed the methodology and M.H.A. was responsible for the formal analysis. The investigation was performed by M.H.A., with I.I. additionally handling the validation of the results. The initial draft of the manuscript was written by M.H.A. and H.A., while H.Y. and W.A.M. were responsible for reviewing and editing subsequent versions. 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: Publicly available datasets were analyzed in this study. This data can be found here: [https://www.adcis.net/en/third-party/e-ophtha/, accessed on 18 March 2024], [https://www.kaggle.com/datasets/nguyenhung1903/diaretdb1-standard-diabetic-retinopathy-database, accessed on 18 March 2024], [https://www.kaggle.com/datasets/deathtrooper/glaucoma-dataset-eyepacs-airogs-light-v2/data, accessed on 18 March 2024].

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

References

1. Sebastian, A.; Elharrouss, O.; Al-Maadeed, S.; Almaadeed, N. A survey on diabetic retinopathy lesion detection and segmentation. *Applied Sciences.* **2023**, *13*, 5111. [CrossRef]
2. Mathews, M.R.; Anzar, S.M. A comprehensive review on automated systems for severity grading of diabetic retinopathy and macular edema. *Int. J. Imaging Syst. Technol.* **2021**, *31*, 2093–2122. [CrossRef]
3. Tang, M.C.S.; Teoh, S.S.; Ibrahim, H.; Embong, Z. Neovascularization detection and localization in fundus images using deep learning. *Sensors* **2021**, *21*, 5327. [CrossRef] [PubMed]
4. Sarhan, A.; Rokne, J.; Alhajj, R. Glaucoma detection using image processing techniques: A literature review. *Comput. Med. Imaging Graph.* **2019**, *78*, 101657. [CrossRef] [PubMed]
5. Xiao, D.; Bhuiyan, A.; Frost, S.; Vignarajan, J.; Tay-Kearney, M.L.; Kanagasingam, Y. Major automatic diabetic retinopathy screening systems and related core algorithms: A review. *Mach. Vis. Appl.* **2019**, *30*, 423–446. [CrossRef]
6. Vives-Boix, V.; Ruiz-Fernández, D. Diabetic retinopathy detection through convolutional neural networks with synaptic metaplasticity. *Comput. Methods Programs Biomed.* **2021**, *206*, 106094. [CrossRef] [PubMed]
7. Tavakoli, M.; Jazani, S.; Nazar, M. Automated detection of microaneurysms in color fundus images using deep learning with different preprocessing approaches. In *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*; SPIE: Bellingham, WA, USA, 2020; Volume 11318, pp. 110–120.
8. Kang, Y.; Fang, Y.; Lai, X. Automatic detection of diabetic retinopathy with statistical method and Bayesian classifier. *J. Med. Imaging Health Inform.* **2020**, *10*, 1225–1233. [CrossRef]
9. Das, S.; Saha, S.K. Diabetic retinopathy detection and classification using CNN tuned by genetic algorithm. *Multimed. Tools Appl.* **2022**, *81*, 8007–8020. [CrossRef]

10. Chudzik, P.; Majumdar, S.; Calivá, F.; Al-Diri, B.; Hunter, A. Microaneurysm detection using fully convolutional neural networks. *Comput. Methods Programs Biomed.* **2018**, *158*, 185–192. [CrossRef] [PubMed]
11. Palanisamy, G.; Ponnusamy, P.; Gopi, V.P. An improved luminosity and contrast enhancement framework for feature preservation in color fundus images. *Signal Image Video Process.* **2019**, *13*, 719–726. [CrossRef]
12. Gupta, B.; Tiwari, M. Color retinal image enhancement using luminosity and quantile-based contrast enhancement. *Multidimens. Syst. Signal Process.* **2019**, *30*, 1829–1837. [CrossRef]
13. Schuch, P.; Schulz, S.; Busch, C. Survey on the impact of fingerprint image enhancement. *IET Biom.* **2018**, *7*, 102–115. [CrossRef]
14. Saba, T.; Rehman, A.; Mehmood, Z.; Kolivand, H.; Sharif, M. Image enhancement and segmentation techniques for detection of knee joint diseases: A survey. *Curr. Med. Imaging* **2018**, *14*, 704–715. [CrossRef]
15. Singh, G.; Mittal, A. Various image enhancement techniques—A critical review. *Int. J. Innov. Sci. Res.* **2014**, *10*, 267–274.
16. Qi, Y.; Yang, Z.; Sun, W.; Lou, M.; Lian, J.; Zhao, W.; Deng, X.; Ma, Y. A comprehensive overview of image enhancement techniques. *Arch. Comput. Methods Eng.* **2021**, *29*, 583–607. [CrossRef]
17. Soundrapandiyan, R.; Satapathy, S.C.; PVSSR, C.M.; Nhu, N.G. A comprehensive survey on image enhancement techniques with special emphasis on infrared images. *Multimed. Tools Appl.* **2021**, *81*, 9045–9077. [CrossRef]
18. Vijayalakshmi, D.; Nath, M.K.; Acharya, O.P. A comprehensive survey on image contrast enhancement techniques in spatial domain. *Sens. Imaging* **2020**, *21*, 1–40. [CrossRef]
19. Sahu, S.; Singh, A.K.; Ghrera, S.P.; Elhoseny, M. An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE. *Opt. Laser Technol.* **2019**, *110*, 87–98.
20. Mazlan, N.; Yazid, H.; Arof, H.; Mohd Isa, H. Automated microaneurysms detection and classification using multilevel thresholding and multilayer perceptron. *Journal of Medical and Biological Engineering.* **2020**, *40*, 292–306.
21. Cao, L.; Li, H.; Zhang, Y. Retinal image enhancement using low-pass filtering and α-rooting. *Signal Process.* **2020**, *170*, 107445. [CrossRef]
22. Foracchia, M.; Grisan, E.; Ruggeri, A. Luminosity and Contrast Normalization in Retinal Images. *Med. Image Anal.* **2005**, *9*, 179–190. [CrossRef] [PubMed]
23. Yang, L.; Yan, S.; Xie, Y. Detection of microaneurysms and hemorrhages based on improved Hessian matrix. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 883–894. [CrossRef] [PubMed]
24. Mayya, V.; Kamath, S.; Kulkarni, U. Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A Comprehensive review. *Comput. Methods Programs Biomed. Update* **2021**, *1*, 100013. [CrossRef]
25. Zhou, M.; Jin, K.; Wang, S.; Ye, J.; Qian, D. Color retinal image enhancement based on luminosity and contrast adjustment. *IEEE Trans. Biomed. Eng.* **2017**, *65*, 521–527. [CrossRef] [PubMed]
26. Rao, K.; Bansal, M.; Kaur, G. A hybrid method for improving the luminosity and contrast of color retinal images using the JND model and multiple layers of CLAHE. *Signal Image Video Process.* **2022**, *17*, 207–217. [CrossRef]
27. Mitra, A.; Roy, S.; Roy, S.; Setua, S.K. Enhancement and restoration of non-uniform illuminated fundus image of retina obtained through thin layer of cataract. *Comput. Methods Programs Biomed.* **2018**, *156*, 169–178. [CrossRef] [PubMed]
28. Anilet Bala, A.; Aruna Priya, P.; Maik, V. Retinal image enhancement using adaptive histogram equalization tuned with nonsimilar grouping curvelet. *Int. J. Imaging Syst. Technol.* **2021**, *31*, 1050–1064. [CrossRef]
29. Dissopa, J.; Kansomkeat, S.; Intajag, S. Enhance Contrast and Balance Color of Retinal Image. *Symmetry* **2021**, *13*, 2089. [CrossRef]
30. Vonghirandecha, P.; Karnjanadecha, M.; Intajag, S. Contrast and color balance enhancement for non-uniform illumination retinal images. *Teh. Glas.* **2019**, *13*, 291–296. [CrossRef]
31. Qureshi, I.; Ma, J.; Shaheed, K. A hybrid proposed fundus image enhancement framework for diabetic retinopathy. *Algorithms* **2019**, *12*, 14. [CrossRef]
32. Alwazzan, M.J.; Ismael, M.A.; Ahmed, A.N. A hybrid algorithm to enhance colour retinal fundus images using a Wiener filter and CLAHE. *J. Digit. Imaging* **2021**, *34*, 750–759. [CrossRef] [PubMed]
33. Cao, L.; Li, H. Enhancement of blurry retinal image based on non-uniform contrast stretching and intensity transfer. *Med. Biol. Eng. Comput.* **2020**, *58*, 483–496. [CrossRef] [PubMed]
34. Kumar, R.; Bhandari, A.K. Luminosity and contrast enhancement of retinal vessel images using the weighted average histogram. *Biomed. Signal Process. Control* **2022**, *71*, 103089. [CrossRef]

diagnostics

MDPI

Communication

Clinical Trial Validation of Automated Segmentation and Scoring of Pulmonary Cysts in Thoracic CT Scans

Aneesha Baral [1], Simone Lee [1], Farah Hussaini [1], Brianna Matthew [1], Alfredo Lebron [1], Muyang Wang [1], Li-Yueh Hsu [2], Joel Moss [3] and Han Wen [1,*]

[1] Laboratory of Imaging Physics, Biochemistry and Biophysics Center, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA; aneesha.baral@nih.gov (A.B.); simone.lee@nih.gov (S.L.); farah.hussaini@nih.gov (F.H.); brianna.matthew@nih.gov (B.M.); alfredolebron08@gmail.com (A.L.); muyang.wang2@nih.gov (M.W.)
[2] Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, MD 20892, USA; lyhsu@nih.gov
[3] Pulmonary Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA; mossj@nhlbi.nih.gov
* Correspondence: wenh@nhlbi.nih.gov

Abstract: In cystic lung diseases such as lymphangioleiomyomatosis (LAM), a CT-based cyst score that measures the percentage of the lung volume occupied by cysts is a common index of the cyst burden in the lungs. Although the current semi-automatic measurement of the cyst score is well established, it is susceptible to human operator variabilities. We recently developed a fully automatic method incorporating adaptive features in place of manual adjustments. In this clinical study, the automatic method is validated against the standard method in several aspects. These include the agreement between the cyst scores of the two methods, the agreement of each method with independent tests of pulmonary function, and the temporal consistency of the measurements in the consecutive visits of the same patients. We found that the automatic method agreed with the standard method as well as the agreement between two trained operators running the same standard method; both methods obtained the same level of correlation with laboratory pulmonary function tests; the automated method had better temporal consistency than the standard method ($p < 0.0001$). The study indicates that the automatic method could replace the standard method and provide better consistency in assessing the extent of cystic changes in the lungs of patients.

Keywords: lymphangioleimyomatosis; cystic lung disease; CT score; cyst burden; cyst score; low-attenuation volume; emphysema

Citation: Baral, A.; Lee, S.; Hussaini, F.; Matthew, B.; Lebron, A.; Wang, M.; Hsu, L.-Y.; Moss, J.; Wen, H. Clinical Trial Validation of Automated Segmentation and Scoring of Pulmonary Cysts in Thoracic CT Scans. *Diagnostics* **2024**, *14*, 1529. https://doi.org/10.3390/diagnostics14141529

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 13 June 2024
Revised: 7 July 2024
Accepted: 10 July 2024
Published: 15 July 2024

1. Introduction

Lymphangioleiomyomatosis (LAM) is a rare, progressive lung disease that affects 3.4–7.8 women per million [1]. It is caused by mutations in the tumor-suppressing tuberous sclerosis complex (TSC), leading to the constant activation of the mechanistic target of rapamycin (mTOR) biochemical pathway [2,3]. This results in the proliferation of smooth muscle cells that can lead to the formation of air-filled cysts in the lungs due to airway obstruction, airway narrowing, and air trapping [3]. There are two types of LAM: TSC-LAM and sporadic LAM. TSC-LAM is a hereditary form of the disease that results from germline mutations in TSC genes [2]. Sporadic LAM, the less common form, occurs due to somatic mutations primarily in TSC 2 and predominantly affects premenopausal women [4]. The clinical features of LAM disease include recurrent pneumothorax, chylous effusions, and shortness of breath caused by airflow obstruction and hyperinflation [5]. Most patients with LAM have declines in pulmonary function tests (PFTs) including a decline in airflow due to an increase in airway resistance and poor gas exchange via a reduction in diffusion

capacity [3]. Although LAM primarily affects the lungs, it can also have multiorgan manifestations such as epilepsy, brain lesions, and renal angiomyolipoma [6].

Given that patients in the US have a median transplant-free survival of 29 years, regular monitoring is an important part of the treatment and management of the disease [7,8]. Traditionally, pulmonary function tests have been used to monitor lung function with an emphasis on metrics such as forced expiratory volume (FEV_1), the Tiffeneau–Pinelli index of forced expiratory volume over forced vital capacity (FEV_1/FVC), and diffusion capacity for carbon monoxide (DL_{CO}) [9,10]. A more recent method to help in the diagnosis and prognosis of this rare disease has been the use of high-resolution computed tomography (CT). CT scans offer the direct visualization of cystic changes and are essential for generating the cyst score, a quantitative measure of the percentage of lung volume occupied by cysts (Figure 1). Studies have demonstrated the close association between cyst scores and declines in pulmonary function, offering clinicians an additional tool for managing the disease in patients with LAM [7,11–13].

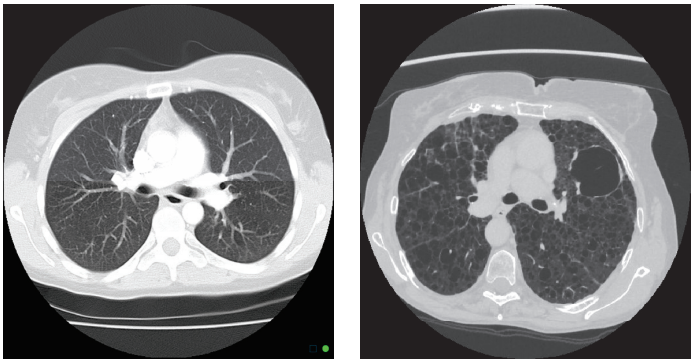


Figure 1. Comparing the thoracic CT scan of healthy lungs (**left**) and lungs with LAM (**right**). In the image on the right, round, air-filled cysts of a range of sizes appear as dark voids in the lung tissue.

Currently, the gold-standard methods to segment the cysts and calculate the cyst score are the FDA-approved semi-automatic software (e.g., Canon Medical Systems USA, Inc., Tustin, CA, USA). However, semi-automatic methods have inherent limitations. They rely on trained operators to make manual adjustments based on visual impressions to correct for instrumental and patient-to-patient variabilities. As a result, they are susceptible to inter- and intra-operator variability, leading to potential inconsistencies in the cyst score [14]. These affect the evaluation of the rate of the change or progression of the cyst burden in the lungs over time. Additionally, the requirement for operators to undergo training to maintain uniform visual standards limits the accessibility of the procedure.

A solution to this problem is fully automatic cyst segmentation. Previously, Schmithorst and co-authors have described a method based on the frequency histogram of the CT values of the whole lung to automatically calculate the percentage volume occupied by cysts [15]. However, the method does not provide the identification and segmentation of the cysts in the images. More recently, we introduced a fully automatic method for generating cyst segmentation and cyst scores based on the CT attenuation values of the air space surrounding the body, the large airways, and the local parenchyma [14]. These calculations remove the need for the subjective visual adjustment of the attenuation threshold by the trained operator, thus removing the human influence in the process. The technical details of the method are described in a previous study [14].

In this clinical study, we assessed the automatic method in a cohort of patients with LAM by comparing it with the gold-standard semi-automatic method and correlating it with pulmonary function tests. Specifically, we evaluated the agreement between the measurements of the automatic and semi-automatic methods, the agreement of each method with the laboratory tests of pulmonary function, and the consistency of the measurements in consecutive visits by the same patients.

2. Materials and Methods

2.1. Study Population

We performed a retrospective study on 208 CT scans from 152 female patients with LAM at the Clinical Center of the National Institutes of Health (NIH), Bethesda, MD, USA during the period from June 2018 to May 2024. The patients were enrolled in the clinical research protocol "Role of Genetic Factors in the Pathogenesis of Lung Disease" (clinicaltrials.gov, NCT00001532), which was approved by the National Heart, Lung, and Blood Institute, the National Institutes of Health Institutional Review Board (IRB # 96-H-0100). The diagnosis of LAM was confirmed by the American Thoracic Society/Japanese Pulmonary Society criteria [10]. Ten patients had a history of tuberous sclerosis complex (TSC). The age range of the study population was from 25.9 to 76.0 years with a median age of 50.9 years. Some of the patients are followed at our hospital on a regular basis and have undergone CT scans at annual or bi-annual intervals as part of their regular visits.

2.2. Thoracic CT Scan Protocol

All the patients received inspiratory volumetric high-resolution helical chest CT scans with a nominal breath-hold period of 10 s and a nominal dose of 0.16 rem. The reconstruction slice thickness was 1 to 2 mm and the slice center-to-center spacing was 1 mm. The in-plane pixel resolution was between 0.6 and 0.8 mm. Depending on the scanner platform, a soft-tissue type of convolution kernel was employed in the reconstruction.

2.3. Laboratory Pulmonary Function Tests

Pulmonary function tests (PFTs) were performed in a clinical pulmonary physiology laboratory of the hospital during the same visit as the CT exams. The PFT values that were utilized for this study were the forced expiratory volume in the first second, expressed in the percentage of predicted values (FEV_1_pp); FEV_1 normalized to the forced vital capacity expressed in the percentage of predicted values (FEV_1/FVC_pp); and the diffusion capacity of the lungs for carbon monoxide, adjusted for hemoglobin and expressed in percentage of predicted values (DL_{CO}_adj_pp) [16,17].

2.4. Standard Semi-Automatic Cyst Score Procedure

The FDA-approved software for cyst segmentation and cyst score calculation is part of the CT scanner platform (Canon Aquilion ONE, Canon Medical Systems USA, Tustin, CA, USA). Depending on the version of the scanner software and hardware, the chest CT series with the appropriate reconstruction setting is loaded into the software. The software makes an initial segmentation of the cystic areas based on a fixed global threshold of the CT attenuation value of −940 Hounsfield Units. A trained operator then inspects the segmentation and adjusts the threshold either upward or downward by the appropriate amount until a visually satisfactory segmentation is achieved. After segmentation, the software automatically sums up the total volume of the cysts and calculates the percentage ratio of total cyst volume/total lung volume as the cyst score. In this study, two trained operators used the FDA-approved software to generate standard scores for inter-operator comparison, as explained in Section 2.6.1 below.

2.5. Automatic Cyst Score Procedure

The automatic procedure was applied to the same CT image series as the standard semi-automatic procedure. The software uses established algorithms to automatically isolate the lung volumes and the large airways [18,19] before applying information from several sources, including the surrounding air background, the airways, and the local parenchyma to calculate the CT attenuation thresholds for cystic areas on a location-specific basis. The thresholds were then applied to automatically segment the cystic areas, and the software then calculated the percentage volume fraction of the lungs occupied by the cysts as the cyst score. A detailed description of the software pipeline is given by Lee and co-authors [14].

2.6. Statistical Analysis

2.6.1. Comparing Automatic and Standard Cyst Scores

We assessed the inter-method difference between the standard semi-automatic procedure and the new automatic procedure, using as a reference the intra-method variability of the semi-automatic procedure itself. The differences between cyst scores generated by the semi-automated and automated methods were evaluated in 100 CT scans from 100 patients. These scans were accompanied by concurrent pulmonary function tests in the same visits. The differences between the two scores were analyzed in two ways: through the Bland–Altman analysis, which examines the mean and variance of the difference [20], and by assessing the mean and variance of the absolute difference. We then considered whether the differences between the automatic and standard scores could be accounted for by the human variability within the standard cyst score itself. For this purpose, we obtained the standard cyst scores from two trained operators scoring the same scans independently. These data were available for a separate set of 26 CT scans from a group of 19 patients. Consequently, two sets of differences were obtained from two pairs of cyst scores: one set between the automatic and standard cyst scores, the other between the standard cyst scores from the two operators. We compared the means and variances of the two differences with Welch's t-test and the two-sample F-test, respectively. Since the first pair of scores covered a wider range of cyst scores than the second pair, it was truncated down to the same range for the comparison to be valid.

2.6.2. Comparing the Correlation of the Cyst Scores to Pulmonary Function Tests

We obtained the correlations of the automatic and standard cyst scores with the three pulmonary function tests in the 100 CT scans described above. We then compared the two cyst scores in terms of their correlation with the PFTs using the test of the difference between two dependent correlations with one variable in common [21].

2.6.3. Comparing the Consistency of the Cyst Scores from the Two Methods

The inconsistency of the cyst scores over time, whether due to operator inconsistencies, instrumental factors, or patient factors, will introduce stochastic fluctuations in the cyst scores resulting in a greater variance of the rate of the change in the cyst scores. With this reasoning, we studied the cyst score rate of change from the most recent two consecutive visits in a group of 41 patients. The median interval between the two visits was 15.0 months. We compared the automatic and the standard semi-automatic methods in terms of the average and the variance of the cyst score rate of change.

3. Results

An example of pulmonary cyst segmentation by the standard semi-automatic method and by the automatic method is illustrated in Figure 2. The results of the statistical comparisons between the two methods are described in the sub-sections below.

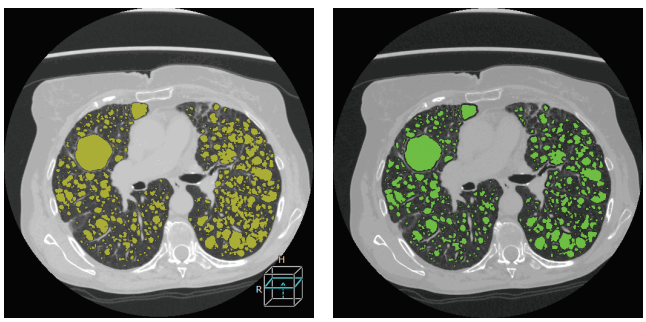
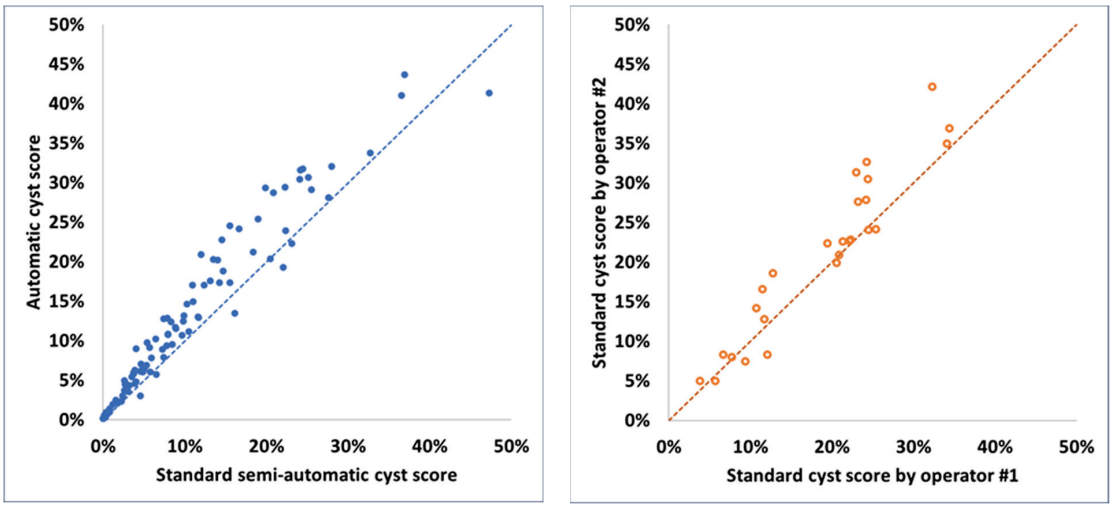


Figure 2. An example of the standard semi-automatic cyst segmentation (**left**) and fully automatic cyst segmentation (**right**) of an axial section of the chest CT scan of a patient with LAM. The cystic areas are highlighted in chartreuse color in the image on the left, and green color in the image on the right.

3.1. Comparing Automatic and Standard Cyst Scores

The standard semi-automatic cyst scores and the matching automatic cyst scores are represented in the scatter plot in Figure 3a. The standard scores on the same CT scans by two different trained operators are represented in Figure 3b. Both pairs of scores lay generally near the identity lines. The second pair covered a smaller range of cyst scores between 4.3% and 37.3%. In this range, the Bland–Altman analyses of the differences in the pairs of scores are summarized in Figure 4. The average difference between the automatic and the standard cyst scores was (mean ± std) (3.28 ± 2.72)%; the average difference between the two standard scores from the two operators was (2.29 ± 3.40)%. The two sets of differences did not differ from each other significantly either in their average values ($p = 0.19$) or in their variances ($p = 0.078$).



(**a**) (**b**)

Figure 3. Scatter plots comparing the two pairs of cyst scores. (**a**) The automatic cyst scores versus the standard semi-automatic cyst scores of 100 chest CT scans from 100 patients. The dashed blue line is the identity line. (**b**) The standard semi-automatic cyst scores generated by trained operator #1 versus trained operator #2 for the same 26 chest CT scans from 19 patients. The dashed orange line is the identity line.

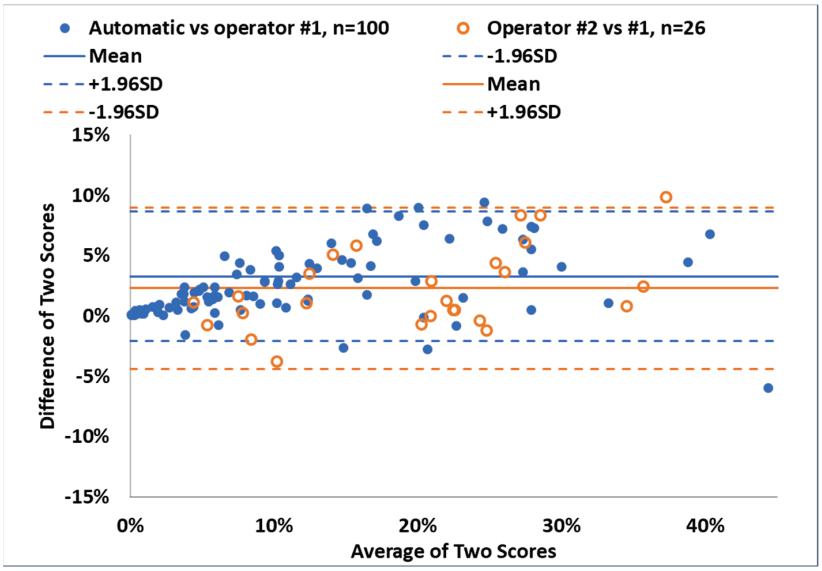


Figure 4. Bland–Altman plots of difference versus average value for the two pairs of cyst scores. The blue symbols and lines are the analysis of the difference between the standard semi-automatic cyst

scores and the automatic cyst scores, both of 100 CT scans from 100 patients. The blue dots represent the difference versus the average for individual scans. The solid horizontal blue line represents the average of the differences, and the horizontal dashed blue lines represent the 95% confidence interval of the difference. Similarly, the orange symbols and lines are the analysis of the difference between a pair of standard semi-automatic cyst scores generated by operator #1 and operator #2, of 26 CT scans from 19 patients. The orange circles represent the difference versus the average for individual scans. The solid horizontal orange line represents the average of the differences, and the horizontal dashed orange lines represent the 95% confidence interval of the difference.

In terms of the absolute values of the difference between the automatic scores and the standard scores, the average of the absolute differences was (3.48 ± 2.46)%; it was (2.95 ± 2.81)% for the difference between the two standard scores from the two operators. These are summarized in Figure 5. Again, the two sets of absolute differences did not differ significantly from each other either in their average values (p = 0.40) or in their variances (p = 0.19).

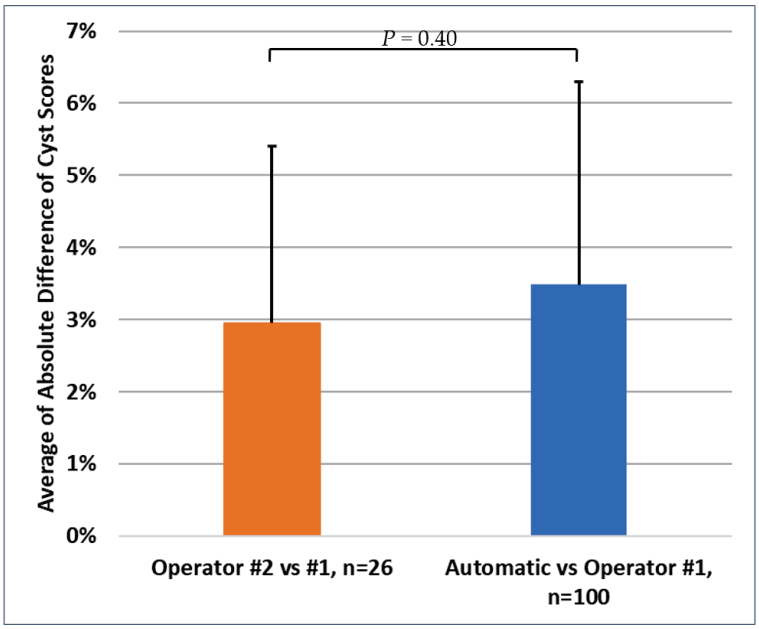


Figure 5. Average of the absolute value of the difference in the cyst scores generated by operator #1 vs. operator #2 (brown bar), and the average of the absolute value of the difference in the cyst scores generated by the automatic method vs. operator #1 (blue bar). The error bars represent the standard deviation. The two average values were statistically comparable (p = 0.40 for the comparison of the mean values and p = 0.19 for the comparison of the variances).

3.2. Comparing the Correlation of the Cyst Scores to Pulmonary Function Tests

All the cyst scores were negatively correlated with the pulmonary function test results, meaning that higher cyst scores were associated with lower values of pulmonary function indices. The absolute values of Pearson's correlation are summarized in Figure 6. The automatic cyst scores had a slightly stronger correlation with all three categories of PFTs compared to the standard semi-automatic scores, but the difference was not statistically significant. The correlation values were −0.765 vs. −0.760 for FEV_1_pp with p = 0.73 for the comparison, −0.772 vs. −0.757 for FEV_1/FVC_pp with p = 0.36 for the comparison, and −0.715 vs. −0.681 for DL_{CO}_adj_pp with p = 0.062 for the comparison.

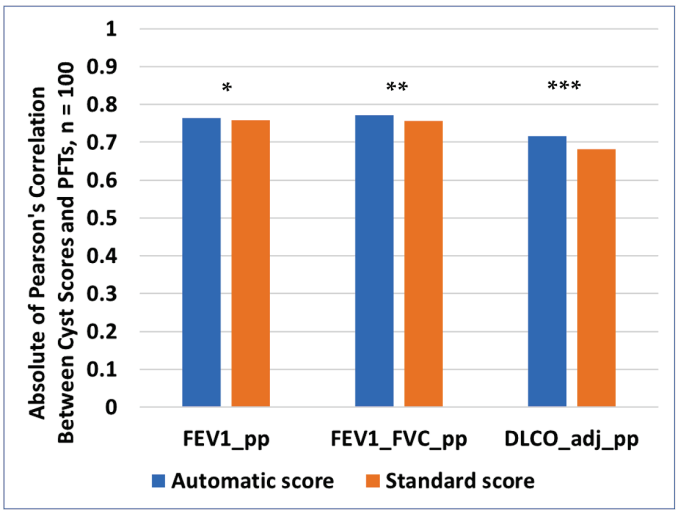


Figure 6. Absolute value of correlation between pulmonary function tests and the two types of cyst scores. The blue bars represent the cyst scores generated by the automatic method, and the orange bars represent the cyst scores generated by a trained operator using the standard semi-automatic method. The two sets of correlations were statistically comparable (*: $p = 0.73$; **: $p = 0.36$; ***: $p = 0.062$).

3.3. Comparing the Consistency of the Cyst Scores from the Two Methods

The consistency of the cyst scores was evaluated through the rate of change in the cyst scores over two consecutive visits. The results are summarized in Figure 7. The average rate of cyst score changes was (0.25 ± 1.12)%/year by the automatic method, and (0.68 ± 2.19)%/year by the standard semi-automatic method. The average value of the rate of change was not statistically different between the two methods with $p = 0.20$. The variance of the rate of change from the automatic method was significantly smaller than the one from the standard method (standard deviation of 1.12%/year vs. 2.19%/year, $p < 0.0001$).

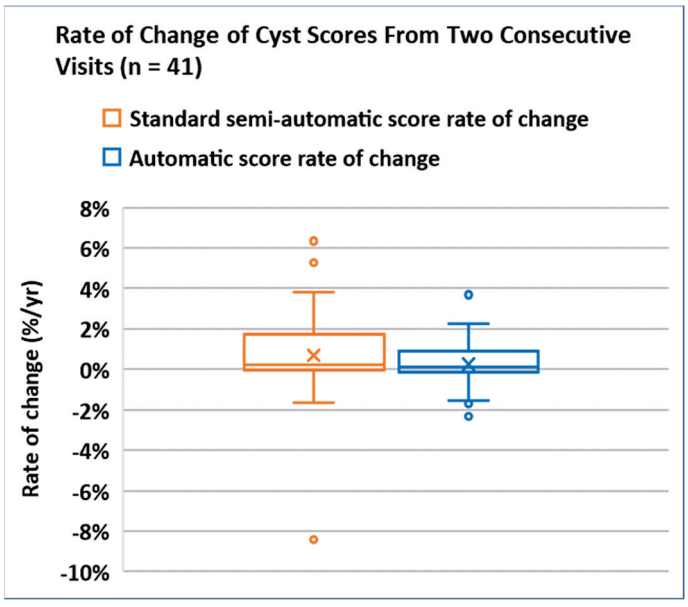


Figure 7. Box-and-whisker plot of the rate of change in the cyst scores obtained from two consecutive patient visits. The average rate of change for the standard semi-automatic (orange) and automatic (blue) methods were statistically comparable ($p = 0.20$). However, the spread of the rate of change in the automatic method was significantly smaller than that of the semi-automatic method (a standard deviation of 1.12%/year vs. 2.19%/year, $p < 0.0001$).

4. Discussion

In cystic lung diseases such as lymphangioleiomyomatosis (LAM), a CT-based quantitative score of the extent of cystic changes in the lungs is often an integral part of the management of the disease which aids the evaluation of the stage of the condition and the effect of treatment [11–13]. Current FDA-approved software to segment the cysts in the CT images and generate the score is semi-automatic, where a trained operator visually adjusts the segmentation until it is deemed optimal. Although this procedure routinely provides clinically useful information, it is affected by inherent human subjective factors in the form of inter- and intra-operator inconsistencies [14], particularly when assessing changes over time. In response, we developed a fully automatic cyst segmentation and scoring method. In this study, the automatic method is validated against the standard semi-automatic method and against independent pulmonary function tests from the clinical physiology laboratory.

In the direct comparison between the automatic and the standard semi-automatic cyst scores, the difference between the standard cyst scores generated by the two operators independently was used as a reference. This reference provided a measure of the variability within the semi-automatic score itself. The results showed that the discrepancy between the automatic method and the standard method was as large as the discrepancy between the two trained operators running the same standard method. Therefore, the automatic method agreed with the standard method to the level allowed by the operator-related variability in the standard method. By the same reasoning, it is plausible that the difference between the automatic and semi-automatic scores could be accounted for by the human variability within the semi-automatic scores.

When it comes to the correlation between the cyst scores and pulmonary function as measured by the physiology laboratory tests, the automatic scores had slightly stronger associations with the PFT values compared to the standard semi-automatic scores, although the differences were not statistically significant ($p \geq 0.062$).

In terms of the consistency of the cyst scores over time, we reasoned that inconsistency would lead to random fluctuations in the scores which would broaden the spread of the rate of change in the scores. Experimentally, the rate of change in the standard semi-automatic scores had about twice the spread of the one from the automatic scores ($p < 0.0001$). The evidence supports the notion that the automatic method improved the consistency of the cyst scores. This is expected, since the semi-automatic procedure is influenced by operator judgment, which may vary slightly between different operators and may fluctuate over time in the same operator. Therefore, by removing the human factor in the process, the level of consistency should improve when all the other factors are equal.

In the general context of the automatic segmentation of radiologic images in lung disease, machine learning-based approaches have been the focus of many efforts recently [22]. In the literature, we are aware of one attempt at developing a machine learning-based method to segment cysts in the chest CT scans of patients with LAM [23] which came from our institute. Although preliminary results showed promise, it is not yet fully functional. In contrast, the image-based adaptive method described in this study proved to be sufficiently robust to yield an operational software that is being used routinely in patients with LAM under a clinical research protocol at our hospital. The several experimental results in combination provide evidence that the automatic cyst segmentation and scoring serves as an equally accurate method as the FDA-approved standard semi-automated segmentation and scoring method. The results also suggest that the automatic method improved the consistency of the scores over time, which is an important factor when it comes to assessing the progression of the disease and the effect of treatment. A limitation of this study is the relatively small number of CT scans available to assess the variability within the standard method itself (26 scans from 19 patients). This point may be addressed in the future expansions of the study. Overall, it may be concluded that the automatic cyst score is a reliable replacement for the current gold-standard semi-automatic cyst score in the evaluation of patients with LAM disease.

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

Funding: This research was funded by the Division of Intramural Research, National Heart, Lung and Blood Institute, NIH Internal Research Program, the National Institutes of Health, USA, grant number HL006141.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the National Heart, Lung, and Blood Institute, the National Institutes of Health, USA (protocol code 96-H-0100, approved annually starting in 1996).

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

Data Availability Statement: Please contact the corresponding author with reasonable requests for clinical study data.

Acknowledgments: We thank Dumitru Mazilu for obtaining some of the CT image series from archive servers.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Harknett, E.C.; Chang, W.Y.C.; Byrnes, S.; Johnson, J.; Lazor, R.; Cohen, M.M.; Gray, B.; Geiling, S.; Telford, H.; Tattersfield, A.E.; et al. Use of Variability in National and Regional Data to Estimate the Prevalence of Lymphangioleiomyomatosis. *QJM* **2011**, *104*, 971–979. [CrossRef] [PubMed]
2. McCarthy, C.; Gupta, N.; Johnson, S.R.; Yu, J.J.; McCormack, F.X. Lymphangioleiomyomatosis: Pathogenesis, Clinical Features, Diagnosis, and Management. *Lancet Respir. Med.* **2021**, *9*, 1313–1327. [CrossRef] [PubMed]
3. Hohman, D.W.; Noghrehkar, D.; Ratnayake, S. Lymphangioleiomyomatosis: A Review. *Eur. J. Intern. Med.* **2008**, *19*, 319–324. [CrossRef] [PubMed]
4. Sato, T.; Seyama, K.; Fujii, H.; Maruyama, H.; Setoguchi, Y.; Iwakami, S.; Fukuchi, Y.; Hino, O. Mutation Analysis of the TSC1 and TSC2 Genes in Japanese Patients with Pulmonary Lymphangioleiomyomatosis. *J. Hum. Genet.* **2002**, *47*, 20–28. [CrossRef] [PubMed]
5. Gupta, N.; Vassallo, R.; Wikenheiser-Brokamp, K.A.; McCormack, F.X. Diffuse Cystic Lung Disease. Part I. *Am. J. Respir. Crit. Care Med.* **2015**, *191*, 1354–1366. [CrossRef] [PubMed]
6. Khaddour, K.; Sankari, A.; Shayuk, M. Lymphangioleiomyomatosis. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2024.
7. Cohen, M.M.; Pollock-BarZiv, S.; Johnson, S.R. Emerging Clinical Picture of Lymphangioleiomyomatosis. *Thorax* **2005**, *60*, 875–879. [CrossRef] [PubMed]
8. Oprescu, N.; McCormack, F.X.; Byrnes, S.; Kinder, B.W. Clinical Predictors of Mortality and Cause of Death in Lymphangioleiomyomatosis: A Population-Based Registry. *Lung* **2013**, *191*, 35–42. [CrossRef] [PubMed]
9. Johnson, S.R.; Cordier, J.F.; Lazor, R.; Cottin, V.; Costabel, U.; Harari, S.; Reynaud-Gaubert, M.; Boehler, A.; Brauner, M.; Popper, H.; et al. European Respiratory Society Guidelines for the Diagnosis and Management of Lymphangioleiomyomatosis. *Eur. Respir. J.* **2010**, *35*, 14–26. [CrossRef] [PubMed]
10. McCormack, F.X.; Gupta, N.; Finlay, G.R.; Young, L.R.; Taveira-DaSilva, A.M.; Glasgow, C.G.; Steagall, W.K.; Johnson, S.R.; Sahn, S.A.; Ryu, J.H.; et al. Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guidelines: Lymphangioleiomyomatosis Diagnosis and Management. *Am. J. Respir. Crit. Care Med.* **2016**, *194*, 748–761. [CrossRef] [PubMed]
11. Aberle, D.R.; Hansell, D.M.; Brown, K.; Tashkin, D.P. Lymphangiomyomatosis—CT, Chest Radiographic, and Functional Correlations. *Radiology* **1990**, *176*, 381–387. [CrossRef] [PubMed]
12. Argula, R.G.; Kokosi, M.; Lo, P.; Kim, H.J.; Ravenel, J.G.; Meyer, C.; Goldin, J.; Lee, H.-S.; Strange, C.; McCormack, F.X.; et al. A Novel Quantitative Computed Tomographic Analysis Suggests How Sirolimus Stabilizes Progressive Air Trapping in Lymphangioleiomyomatosis. *Ann. Am. Thorac. Soc.* **2016**, *13*, 342–349. [CrossRef] [PubMed]
13. Matthew, B.P.; Hasani, A.M.; Chen, Y.-C.; Pirooznia, M.; Stylianou, M.; Rollison, S.F.; Machado, T.R.; Quade, N.M.; Jones, A.M.; Julien-Williams, P.; et al. Ultra-Small Lung Cysts Impair Diffusion Without Obstructing Air Flow in Lymphangioleiomyomatosis. *Chest* **2021**, *160*, 199–208. [CrossRef] [PubMed]

14. Lee, S.; Lebron, A.; Matthew, B.; Moss, J.; Wen, H. Automated Segmentation and Measurements of Pulmonary Cysts in Lymphangioleiomyomatosis across Multiple CT Scanner Platforms over a Period of Two Decades. *Bioengineering* **2023**, *10*, 1255. [CrossRef] [PubMed]
15. Schmithorst, V.J.; Altes, T.A.; Young, L.R.; Franz, D.N.; Bissler, J.J.; McCormack, F.X.; Dardzinski, B.J.; Brody, A.S. Automated Algorithm for Quantifying the Extent of Cystic Change on Volumetric Chest CT: Initial Results in Lymphangioleiomyomatosis. *Am. J. Roentgenol.* **2009**, *192*, 1037–1044. [CrossRef] [PubMed]
16. Quanjer, P.H.; Stanojevic, S.; Cole, T.J.; Baur, X.; Hall, G.L.; Culver, B.H.; Enright, P.L.; Hankinson, J.L.; Ip, M.S.M.; Zheng, J.; et al. Multi-Ethnic Reference Values for Spirometry for the 3–95-Yr Age Range: The Global Lung Function 2012 Equations. *Eur. Respir. J.* **2012**, *40*, 1324–1343. [CrossRef] [PubMed]
17. Stanojevic, S.; Graham, B.L.; Cooper, B.G.; Thompson, B.R.; Carter, K.W.; Francis, R.W.; Hall, G.L. Official ERS Technical Standards: Global Lung Function Initiative Reference Values for the Carbon Monoxide Transfer Factor for Caucasians. *Eur. Respir. J.* **2017**, *50*, 1700010. [CrossRef] [PubMed]
18. Hedlund, L.W.; Anderson, R.F.; Goulding, P.L.; Beck, J.W.; Effmann, E.L.; Putman, C.E. Two Methods for Isolating the Lung Area of a CT Scan for Density Information. *Radiology* **1982**, *144*, 353–357. [CrossRef] [PubMed]
19. King, G.G.; Müller, N.L.; Whittall, K.P.; Xiang, Q.-S.; Paré, P.D. An Analysis Algorithm for Measuring Airway Lumen and Wall Areas from High-Resolution Computed Tomographic Data. *Am. J. Respir. Crit. Care Med.* **2000**, *161*, 574–580. [CrossRef] [PubMed]
20. Bland, J.M.; Altman, D.G. Statistical-Methods for Assessing Agreement between 2 Methods of Clinical Measurement. *Lancet* **1986**, *1*, 307–310. [CrossRef] [PubMed]
21. Lee, I.A.; Preacher, K.J. Calculation for the Test of the Difference between Two Dependent Correlations with One Variable in Common [Computer Software]. Available online: https://quantpsy.org/corrtest/corrtest2.htm (accessed on 6 June 2024).
22. Kufel, J.; Bielówka, M.; Rojek, M.; Mitręga, A.; Lewandowski, P.; Cebula, M.; Krawczyk, D.; Bielówka, M.; Kondoł, D.; Bargieł-Łączek, K.; et al. Multi-Label Classification of Chest X-Ray Abnormalities Using Transfer Learning Techniques. *J. Pers. Med.* **2023**, *13*, 1426. [CrossRef] [PubMed]
23. Zhang, L.; Gopalakrishnan, V.; Lu, L.; Summers, R.M.; Moss, J.; Yao, J. Self-Learning to Detect and Segment Cysts in Lung CT Images without Manual Annotation. *arXiv* **2018**, arXiv:1801.08486v1.

diagnostics

MDPI

Article

Convolutional Neural Network Model for Intestinal Metaplasia Recognition in Gastric Corpus Using Endoscopic Image Patches

Irene Ligato [1], Giorgio De Magistris [2], Emanuele Dilaghi [1], Giulio Cozza [1], Andrea Ciardiello [3], Francesco Panzuto [1], Stefano Giagu [3], Bruno Annibale [1], Christian Napoli [2] and Gianluca Esposito [1,*]

[1] Department of Medical-Surgical Sciences and Translational Medicine, Sant'Andrea Hospital, Sapienza University of Rome, 00185 Roma, Italy; irene.ligato@uniroma1.it (I.L.); emanuele.dilaghi@uniroma1.it (E.D.); giulio.cozza@uniroma1.it (G.C.); francesco.panzuto@uniroma1.it (F.P.); bruno.annibale@uniroma1.it (B.A.)
[2] Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy; giorgio.demagistris@uniroma1.it (G.D.M.); cnapoli@diag.uniroma1.it (C.N.)
[3] Department of Physics, Sapienza University of Rome, P.le A. Moro 5, 00185 Rome, Italy; andrea.ciardiello@gmail.com (A.C.); stefano.giagu@uniroma1.it (S.G.)
* Correspondence: gianluca.esposito@uniroma1.it

Abstract: Gastric cancer (GC) is a significant healthcare concern, and the identification of high-risk patients is crucial. Indeed, gastric precancerous conditions present significant diagnostic challenges, particularly early intestinal metaplasia (IM) detection. This study developed a deep learning system to assist in IM detection using image patches from gastric corpus examined using virtual chromoendoscopy in a Western country. Utilizing a retrospective dataset of endoscopic images from Sant'Andrea University Hospital of Rome, collected between January 2020 and December 2023, the system extracted 200 × 200 pixel patches, classifying them with a voting scheme. The specificity and sensitivity on the patch test set were 76% and 72%, respectively. The optimization of a learnable voting scheme on a validation set achieved a specificity of 70% and sensitivity of 100% for entire images. Despite data limitations and the absence of pre-trained models, the system shows promising results for preliminary screening in gastric precancerous condition diagnostics, providing an explainable and robust Artificial Intelligence approach.

Keywords: gastric intestinal metaplasia; virtual chromoendoscopy; BLI; CNN; imaging diagnostics; ResNet50; classification; segmentation

Citation: Ligato, I.; De Magistris, G.; Dilaghi, E.; Cozza, G.; Ciardiello, A.; Panzuto, F.; Giagu, S.; Annibale, B.; Napoli, C.; Esposito, G. Convolutional Neural Network Model for Intestinal Metaplasia Recognition in Gastric Corpus Using Endoscopic Image Patches. *Diagnostics* **2024**, *14*, 1376. https://doi.org/10.3390/diagnostics14131376

Academic Editors: Hiam Alquran and Wan Azani Mustafa

Received: 27 May 2024
Revised: 23 June 2024
Accepted: 26 June 2024
Published: 28 June 2024

1. Introduction

Gastric cancer (GC) is the fifth most common neoplasia and the fourth most common cause of death from neoplastic pathology worldwide [1], and its unfavorable prognosis is primarily attributed to late diagnosis at an advanced stage of cancer [2]. However, in low-incidence countries (i.e., European countries), GC screening is not recommended, and the recognition of patients at risk for the development of GC is fundamental to avoiding unnecessary gastroscopies and reducing the burden of care. High-risk patients are those with extensive precancerous gastric conditions or with additional risk factors (i.e., autoimmune gastritis, a family history of GC), and there is evidence that endoscopic surveillance in these patients is beneficial [3]. Intestinal metaplasia (IM) is a gastric precancerous condition characterized by the replacement of the original gastric glands with intestinal epithelium [4]. Gastric IM diagnosis is histological. Therefore, biopsies are necessary during gastroscopy to stage the precancerous conditions and to define the *Helicobacter pylori* status. For reporting the histological evaluation of gastritis and gastric precancerous conditions, the consensus of the updated Sydney system, based on five biopsies of the stomach (two of the antrum and one of the incisura sent in the same vial and two of the corpus sent in another vial), was developed [5]. Therefore, a histological classification system like the Operative Link on Gastric Intestinal Metaplasia (OLGIM) was proposed to stage IM [6]. The OLGIM score

includes stages from 0 to IV. According to these studies, patients with advanced stages of OLGIM (such as stages III and IV) have an increased risk of developing GC [6]. Another important risk factor for developing GC is the presence of IM or AG in the context of autoimmune atrophic gastritis (AAG) that extends to the corpus mucosa while sparing the antrum [7]. The topographic extension of IM is important to define, considering that the risk of developing cancerous lesions is closely linked to the extent of IM across the gastric mucosa. In this context, gastric IM can be divided into diffuse or focal, and for this reason, it is necessary to use diagnostic tools for the support and recognition of gastric precancerous conditions during endoscopic examination that allow target biopsies to be performed and patients to be staged correctly. During White Light (WL) endoscopy, random gastric biopsies may fail to detect IM. In recent years, several studies have demonstrated that electronic chromoendoscopy, namely Narrow-Band Imaging (NBI) and Blue-Light Imaging (BLI), has increased accuracy in the diagnosis of precancerous conditions and cancerous lesions [8]. Electronic chromoendoscopy improved the visualization of vascular and mucosal patterns by employing narrower bands of blue and green filters instead of conventional red–green–blue filters [9]. Chromoendoscopically, gastric IM is characterized by some markers, such as the light blue crest, the marginal turbid band, the white opaque substance, and the groove type [10]. The ability to perform target biopsies with electronic chromoendoscopy increases the chance of obtaining a diagnosis of IM and correct staging. In a previous study, NBI showed an accuracy of 85%-90% for the diagnosis of IM and dysplasia [8]. These results were then confirmed in a European multicenter study in which an NBI classification scheme was proposed and validated to stage IM based only on electronic chromoendoscopy appearance (EGGIM—endoscopic grading of gastric intestinal metaplasia). An EGGIM score > 4 has been shown to be associated with advanced stages of OLGIM (III, IV) [11]. Another type of electronic chromoendoscopy (BLI) showed the same results as NBI for gastric IM diagnosis and staging [12]. However, electronic chromoendoscopy is not widespread, and not every endoscopist can use it. Over the past few decades, significant attention has been directed towards the advancement of computer-assisted systems applicable in endoscopy [13]. Computer vision is a crucial field of research in Artificial Intelligence (AI) regarding images and visual data, and convolutional neural networks (CNNs) have completely changed the field of computer vision. CNNs first appeared in 1980 [14]. A CNN is a type of neural network architecture that learns a hierarchy of features directly from a large dataset for a specific task [15]. Several existing image classification algorithms, such as ResNet50 [16], SE-ResNet50, VGG 16, VGG 19 [17], and Inception V3 [18], are used to resolve complex problems related to image classification and segmentation benchmarks. In the task of image classification, the ResNet-50 model appears to be the most well-known and effective architecture in the medical field [19].

However, the development of a CNN for the task of classifying endoscopic images must be based on several factors [20]:

The dataset being retrospective or prospective;
The dataset being based on images or videos;
The dataset being derived from a single center (internal set) or multiple centers (external set);
The use of endoscopic images exclusively in conventional WL or exclusively with electronic chromoendoscopy (BLI, NBI, LCI), or both types of lights combined;
Types of deep learning algorithms (which can include image classification algorithms, object detection algorithms, semantic segmentation algorithms, or a combination of these algorithms).

The task of the CNN can be focused exclusively on a single task (i.e., the recognition of a single precancerous condition, such as intestinal metaplasia) or address multiple tasks (i.e., the simultaneous recognition of IM, AG, and *H. pylori* gastritis).

Different studies have applied AI models for the detection of upper GI lesions, and specifically, GC, especially in Eastern countries, and a recent meta-analysis showed promising results in the application of AI for GC [21–23]. Another recent meta-analysis demon-

strated good results on the accuracy of AI and its application for early GC diagnosis [24], a condition with a missing lesion rate of up to 20% in Western countries [25].

On this topic, most studies derive from Asian countries, where there is a high prevalence of GC. Conversely, regarding gastric precancerous conditions, only a few studies have focused on the use of AI for AG and IM, and they are exclusively Asian studies, except for a German study on AG [26]. Regarding gastric precancerous conditions, there are three relevant meta-analyses: one about the use of AI for gastric precancerous conditions (based on nine studies) and one about *H. pylori* infection (based on four studies), which have demonstrated a pooled diagnosis accuracy of 90% and 80%, respectively [27]. Another meta-analysis focused on the diagnostic value of AI-assisted endoscopy for chronic AG (based on eight studies) reported a diagnostic accuracy of 98% [28].

Lastly, a meta-analysis was conducted on the diagnostic accuracy of AI exclusively for gastric IM based on 12 eastern studies, demonstrating a pooled sensitivity of 94% and a specificity of 93%. A clinically significant finding of this meta-analysis is the comparison between AI and endoscopists, revealing that AI exhibited higher sensitivity (95% vs. 79%) [29].

The results showed that AI performed excellently in diagnosing GIM, AG, and *H. pylori* infection.

Nonetheless, all the meta-analyses described are based on few studies. Furthermore, there are no Western studies in this field, and developing a CNN system would be useful in supporting endoscopists to recognize and stage IM in a Western country.

For these reasons, this pilot study aimed to develop and assess the precision and recall of a CNN system for the recognition of IM in images of a gastric corpus obtained during gastroscopies with BLI evaluation in a Western country.

2. Methods

2.1. Participants

We retrospectively collected a dataset of endoscopic images from prospective gastroscopies performed in a single center, Sant'Andrea Hospital, Sapienza University of Rome, from January 2020 to December 2023. Images were obtained during routine weekly endoscopic sessions of outpatients undergoing follow-up for atrophic gastritis corpus-restricted, multifocal, and extensive atrophic gastritis, celiac disease, and anemia.

The exclusion criteria were individuals below 18 years of age, incomplete or unavailable clinical data, insufficient gastric biopsies, contraindications for biopsies, patients who had undergone partial or total gastrectomy, and instances involving neoplastic gastric lesions. Ethical approval and informed consent to collect endoscopic images were obtained from all patients.

2.2. Endoscopic and Histological Procedures

Gastroscopies were performed using Fuji scopes with the use of pharyngeal anesthesia (xylocaine spray puffs) and/or conscious or deep sedation (midazolam or propofol). Gastroscopies were performed using High-Resolution (HR) White Light (WL) endoscopy, followed by a BLI assessment. For the aim of this study, multiple images were taken during the BLI assessment of the regions of the antrum and the corpus. If IM was endoscopically observed with the use of BLI, multiple images of the area of interest were taken, and targeted biopsies were performed. During the BLI assessment, if no IM was detected, multiple general images of gastric mucosa were taken, and biopsies according to the updated Sydney system protocol were performed [5]: two biopsies of the antrum, one of the incisura angularis, and two of the corpus. These biopsies were then placed in separate vials for histopathological evaluation, the gold standard for the diagnosis of gastric IM. Additional biopsies were obtained when other abnormalities were identified and subsequently submitted for histopathological examination.

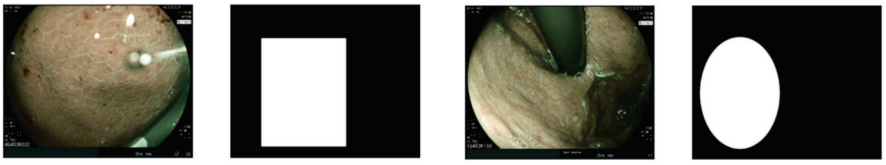
2.3. Image Dataset

For each patient included in the study, a data collection form was administered to collect demographic, endoscopic, and histological data. The images were acquired at a single center during dedicated sessions by two experienced endoscopists, thereby minimizing the risk of selection bias. All images of these gastroscopies were collected in a JPEG format, and after image acquisition, a resident endoscopist discarded WL images, esophagus images, duodenum images, and poor-quality images (i.e., out-of-focus or low-resolution images). The criteria for selecting images were the suitability of images to elucidate the presence or absence of IM; adequate graphic quality; accordance between endoscopic and histological evaluation; and guarantee of the anonymity of the images. Images with histological evaluation compatible with pseudopyloric metaplasia were excluded from the dataset.

Subsequently, an expert endoscopist categorized all the remaining gastric corpus and antrum BLI images into IM-positive and IM-negative images. For this study, only gastric body images that were acquired using BLI assessment were used for the CNN model development. Finally, an expert endoscopist of gastric precancerous conditions performed annotations of regions of interest that contained informative features using the Image J software (Image 1.53 t).

2.4. Development of AI Models

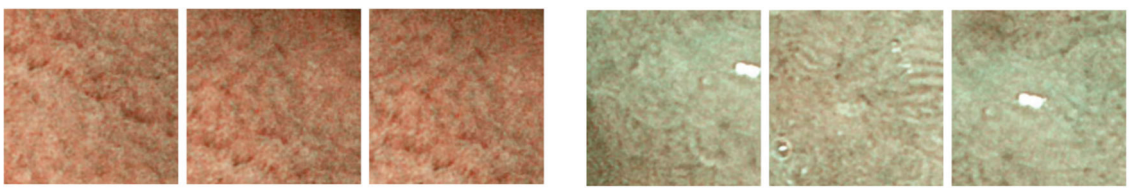
The primary challenges were data availability and annotation quality. Due to the absence of pixel-wise segmentation masks, it was not feasible to tackle the problem as a segmentation task. Instead, we collected annotations on regions of interest, highlighted as rectangular or oval regions in the images, as shown in Figure 1.



(**a**) GIM-Negative region (**b**) GIM-Positive region

Figure 1. Samples from the dataset showing BLI images with corresponding annotations.

The annotations, labeled as IM-positive or IM-negative, highlight regions with discriminative features (Figure 1). Our approach involves extracting and classifying 20 random patches of 200 × 200 pixels from each annotated region (Figure 2). The data were split into 80% for training, 10% for validation and hyperparameter tuning (Section 4), and 10% for testing. Various CNN architectures were tested, including ResNet50, VGG 16, VGG 19, and Inception V3, with the best results achieved using ResNet50. The models were trained from scratch using the Adam optimizer [30] for 200 epochs, selecting the best model based on validation loss. All images were used for testing once only. We calculated the accuracy, recall, and precision of this model. Recall is a metric used to determine the frequency at which a machine learning model accurately identifies true positive cases among all the actual positive cases in the dataset. It is also known as sensitivity. On the other hand, precision assesses how frequently a machine learning model accurately predicts the positive class. Precision can also be called specificity.



(**a**) GIM-Negative patches (**b**) GIM-Positive patches

Figure 2. Examples of randomly sampled patches from GlM-negative and GlM-positive regions.

3. Results

Overall, 279 patients were included (62.7% female). The histological examination revealed the absence of gastric IM in 146 patients, while 133 patients tested positive for gastric IM. Among those with positive results, 81 cases exhibited IM exclusively in the corpus, with the antrum spared. Regarding the histological staging of IM, 146 patients were categorized as OLGIM 0, 50 as OLGIM I, 63 as OLGIM II, 14 as OLGIM III, and 6 as OLGIM IV. Chromoendoscopic staging of IM using the EGGIM score was performed in 227 patients, with 108 classified as EGGIM 0, 148 as EGGIM 1–4, and 22 as EGGIM >4. Twenty-two patients tested positive in the histological examination for *H. pylori* infection.

The final dataset included 1384 high-resolution BLI images, with 721 gastric corpus images classified as IM-negative and 663 as IM-positive. Each image had a resolution of 1279 × 1023 pixels. A total of 1384 images yielded 505 IM-positive and 771 IM-negative annotations. Finally, a total of 6103 IM-positive and 4466 IM-negative patches were obtained. Each image patch had a resolution of 200 × 200 pixels. We split the dataset into three phases: the training set (which contained 4861 positive patches and 3403 negative patches), the validation set (which contained 681 positive patches and 521 negative patches), and the test set (which contained 561 positive patches and 542 negative patches). Images were saved in JPEG format and are available upon reasonable request to the corresponding author.

The model's performance on the patch test set, with patches classified as either IM-positive or IM-negative, demonstrated an accuracy of 74%, a precision of 76%, and a recall of 72%, as illustrated in Table 1. To support the statistical significance of these results, we performed a significance test comparing the performance of ResNet-50 with other CNN architectures like DenseNet, EfficientNet, and XceptionNet. Using a paired *t*-test, we found that ResNet-50's performance was significantly better than the other architectures ($p < 0.05$).

Table 1. Performance metrics of the CNN model on the patch test set, with patches classified as either GIM-positive or GIM-negative.

Model	Accuracy	Precision	Recall
ResNet	74%	76%	72%

While the patch-level results revealed certain limitations for standalone diagnostic use, they formed the foundational layer for further enhancements. For the classification of entire images as either IM-positive or IM-negative, we developed a custom voting scheme to achieve this, as illustrated in Figure 3. The voting scheme in our methodology involves learning a decision threshold to classify entire images based on the classification of individual patches. First, each image is divided into 30 non-overlapping patches. Our convolutional neural network (CNN) model, specifically the ResNet-50 architecture, is trained to classify each patch as either positive or negative for intestinal metaplasia (IM). Then, we employ a linear search to determine an optimal threshold for the number of patches that need to be classified as positive for the entire image to be considered positive. This threshold is learned by maximizing the overall image classification accuracy on the validation set. For each image, if the number of patches classified as positive exceeds the learned threshold, the entire image is classified as positive. Conversely, if the number of positive patches is below the threshold, the image is classified as negative. During the validation phase, we iteratively adjusted the threshold and evaluated the performance to ensure that the chosen threshold maximizes the overall accuracy. This process helps in fine-tuning the model to strike a balance between sensitivity (recall) and specificity (precision). Table 2 details the configurations yielding the most effective results. One of the best configurations, with a decision threshold of 0.8 and a patch threshold of 13/30, showed an accuracy of 78%, a precision of 70%, and a recall of 100%. This voting scheme and threshold optimization enhance the robustness and practical applicability of our method, allowing us to achieve high accuracy in classifying entire endoscopic images based on patch-level predictions. Our method enhances explainability and interpretability over

traditional CNN classifiers by allowing practitioners to inspect which specific patches influenced the final classification decision. For instance, patches showing strong specular reflections might have been erroneously classified as IM-positive. Such errors were readily identifiable and could be corrected, thereby enabling a more robust and interpretable diagnostic process.

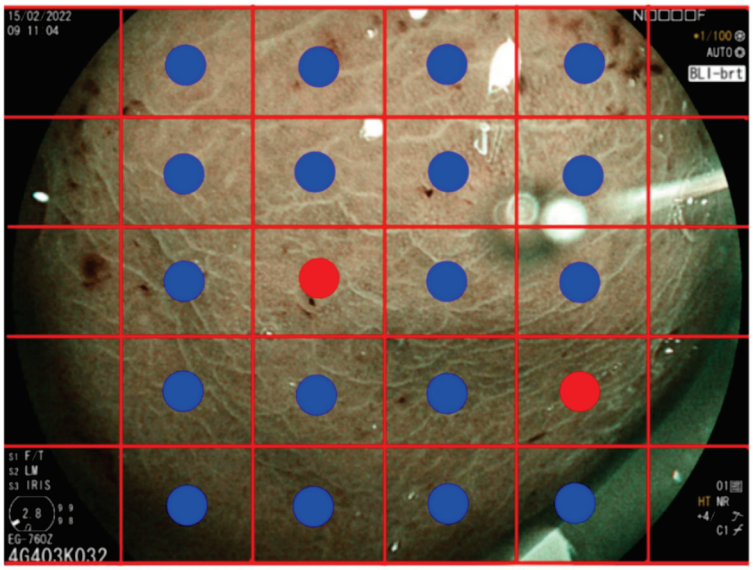


Figure 3. Illustration of the classification algorithm for entire images. The image is divided into 30 non-overlapping patches that are individually classified, followed by a voting scheme to determine the overall image classification.

Table 2. Test set metrics for entire image classification, showing the best configurations of decision and patch thresholds.

Decision Threshold	Patch Threshold	Test Accuracy	Test Precision	Test Recall
0.5	23/30	78%	75%	81%
0.8	13/30	78%	70%	100%
0.5	24/30	76%	68%	83%

4. Discussion

Most studies regarding AI in upper gastrointestinal management have focused on the detection of early gastric cancer [21–24], while only a few studies, exclusively from Asian countries, have investigated these learning systems for gastric precancerous conditions [27–29].

However, given the significant differences in diagnostic outcomes between countries with a high and low prevalence of GC, it is imperative to also investigate a specific CNN for the recognition of GC in countries with a low prevalence of GC. Furthermore, in the Western world, where there is a low prevalence of gastric cancer, recognizing patients with precancerous conditions who will subsequently undergo endoscopic follow-ups is crucial. For these reasons, our CNN could provide valuable support in identifying these conditions. Our system demonstrated high recall and moderate precision, making it particularly suitable for initial screening applications. However, in countries with a low prevalence of gastric cancer, a recognition system with greater precision would be more useful in terms of economic resources. In the initial stages of screening contexts, it is preferable to have high recall in order to have fewer false negatives. The decision to prioritize the identification of IM is based on its pivotal role in the gastric carcinogenic process, marking a critical point of no return. Additionally, in this study, we focused on IM, and distinguishing between

IM and pseudopyloric metaplasia is crucial because studies indicated that the latter is not associated with the development of GC [31]. Pseudopyloric metaplasia and other conditions, such as foveolar hyperplasia or exclusive gastric atrophy, could be disturbing factors for the correct functioning of the learning machine and should be investigated in future studies.

A relevant meta-analysis on the diagnostic accuracy of AI regarding gastric IM conducted subgroup analyses to examine factors influencing AI performance in recognizing gastric IM, including the number of images (>1500 or <1500), the study design (prospective or retrospective), the study center (multicenter or single center), the endoscopy type (WL only or others), and the algorithm type (classification algorithm or others). The analysis revealed that the type of algorithm is the most significant factor that significantly impacts precision (also known as specificity) [29]. By comparing our system with previous AI models for IM detection extracted from the meta-analysis already described [29], our study is the only one that uses the ResNet-50 algorithm in this context. We conducted extensive experimentation with convolutional neural network (CNN)-based architectures such as ResNet, DenseNet [32], EfficientNet [33], and XceptionNet [34], each of which consists of a CNN component for feature extraction and a linear classifier for making predictions. The aim was to determine which model performed best on our specific task, single patch classification, where only the regions of interest were considered, while irrelevant regions were ignored. Our comparative analysis focused on several key factors: 1. Performance: Through rigorous testing, we observed that ResNet consistently achieved the highest accuracy on our dataset. This superior performance was crucial in our decision, as accuracy directly impacts the reliability and effectiveness of the classification task. 2. Complexity: ResNet's architecture, characterized by its residual connections, allows for training deeper networks without the common issue of vanishing gradients. This balance between depth and ease of training made ResNet more suitable for our needs compared to more complex networks like DenseNet and EfficientNet, which, despite their advanced capabilities, did not provide accuracy improvements in our experiments. 3. Suitability for the task: The residual connections in ResNet facilitate better feature extraction from the patches by enabling the network to learn residual mappings, which are particularly effective for distinguishing between the classes in our dataset. This architectural advantage made ResNet more adept at handling the specific nuances of our classification task.

Regarding precision, which in our study was shown to be lower than recall (also called sensitivity), it is in line with the results of most of the previous studies analyzed, which all report values higher than 80% [29]. The diagnostic accuracy of AI in the studies analyzed in this meta-analysis averaged around 90% [29], while our study achieved approximately 80% diagnostic accuracy, although we reported very high recall values of approximately 100%. Comparing our system to earlier AI models for IM detection [29], we found that our system is the only one developed based exclusively on BLI assessment. Most of the previous studies used endoscopic images with electronic chromoendoscopy, such as NBI, BLI, and Linked Color Imaging (LCI), but only one study developed a system using WL, NBI, and BLI [35]. The other systems were developed either with NBI assessment alone or with a combination of WL and NBI assessment.

Another unique aspect, which can also be seen as a limitation, is that our system focuses exclusively on the gastric corpus. However, this is a preliminary analysis, and image collection also involved the antral region, which will be subsequently analyzed for the entire stomach.

To further improve our model, we conducted a detailed error analysis to understand common misclassifications and identify areas for enhancement. One common issue was that some patches containing strong specular reflections or image artifacts were erroneously classified as IM-positive. By implementing preprocessing steps to reduce or eliminate these reflections and artifacts, we could enhance the model's accuracy. Techniques like glare removal and advanced image normalization could be beneficial. Another challenge we encountered was with patches from out-of-focus or low-quality images, which led to

incorrect classifications. Developing a quality assessment module to filter out low-quality patches before classification could prevent such errors. Training a separate CNN to detect and exclude poor-quality patches might improve overall accuracy. Some patches contained regions that were difficult to classify, even for human experts. Enhancing the training dataset with more annotated examples of these ambiguous regions could help the model learn better representations. Rare patterns and edge cases not well represented in the training data were often misclassified. Expanding the dataset to include more examples of rare patterns and edge cases could help the model generalize better. Collaborating with multiple medical centers to gather a more diverse dataset would be advantageous.

Based on the error analysis, several potential improvements were identified to enhance our model's performance. Implementing advanced preprocessing techniques such as glare removal, noise reduction, and image normalization could improve patch quality and reduce misclassification due to artifacts. Incorporating a quality assessment module to automatically detect and exclude low-quality patches before classification could enhance overall accuracy. Collaborating with additional medical centers to gather a more diverse and representative dataset, including rare patterns and edge cases, could improve the model's ability to generalize to new and unseen data. Developing a multistage classification pipeline where an initial model filters out obvious negatives, followed by a more detailed analysis of potential positives, could reduce false positives and improve precision.

By conducting this study, we identified significant issues. Many of the images that were acquired during the endoscopic examination are of poor quality. Images with disturbing factors such as mucus, bubbles, or reduced visibility were discarded for CNN deep learning creation. To reduce the presence of mucus and bubbles in the gastric mucosa, there are studies that suggest premedication before the endoscopic examination [36]. This approach could enhance image quality and reduce the time spent washing/cleaning the gastric mucosa, allowing for more time to detect cancerous and precancerous conditions. Another important point is that the detection of gastric precancerous conditions during gastroscopy must be based on a high-quality gastroscopy, with specific performance measures for upper gastrointestinal endoscopy (exploration time, taking standardized photos, routine use of electronic chromoendoscopy, etc.) explained in the guidelines [37]. In this regard, it would be interesting in future research to explore the development of AI to enhance the quality control of upper gastrointestinal endoscopy [38]. Another problem is image storage spaces that allow for manipulation (such as the annotations of images), always considering that they are sensitive data obtained following authorization expressed in the informed consent of the endoscopic examination. Furthermore, our study had some limitations. Our dataset had a limited number of images, though there is no mandatory minimum number of images for accurate deep learning development. Previous studies used datasets ranging from 84 to 17,000 images to train deep learning models, showing no differences in diagnostic accuracy outcomes [29]. The development of deep learning was only carried out for the gastric corpus. The development of a CNN that investigates only one region of the stomach cannot be used in real practice in the general population but could be useful in identifying the percentage of patients at high risk of developing GC, such as patients with autoimmune gastritis [7]. To date, the gold standard for autoimmune gastritis diagnosis is the histopathological assessment by gastric biopsies during gastroscopy, although non-invasive serological screening before endoscopy may offer utility in some clinical contexts (e.g., serum biomarkers like parietal cells and intrinsic factor, autoantibodies, or serum gastrin and pepsinogens). According to the guidelines, patients with autoimmune gastritis with the presence of AG and/or IM that extends to the corpus mucosa while sparing the antrum require a surveillance gastroscopy every 3 years [3].

However, we will aim to also extend this methodology to the gastric antrum region, enhancing the utility and applicability of our diagnostic tool. Additionally, we will plan to improve the precision and accuracy of our system through semi-supervised pretraining strategies.

Using a dataset comprising thousands of gastric mucosa endoscopic images in WL for pre-training the system, increasing the number of images by including the antral region, and increasing the patient pool are all strategies that could enhance the performance of the CNN system. These strategies are particularly advantageous as they do not require a large, labeled dataset, thus overcoming one of the significant hurdles in medical image analysis. However, for data augmentation, we have already used techniques such as vertical/horizontal reflections, random rotations, and uniform scaling.

An innovative methodology in regenerative medicine to expand the dataset is the assessment of hyperspectral imaging and CycleGAN-simulated data, which have already shown promising results in the field of upper gastrointestinal endoscopy for the detection of early esophageal cancer [39]. Although all existing studies are in the early stages of development with only internal validation [40], this is an emerging field in medical AI. Research in this area is expected to grow exponentially in the coming years and should be considered for future studies. Our efforts will focus on refining these techniques to provide more reliable and robust diagnostic solutions.

The implementation of CNN models in routine endoscopic practice will enhance the quality of gastroscopy by improving the detection of patients at risk for the development of gastric cancer (i.e., those requiring endoscopic surveillance) as well as healthy patients. In this last subgroup of patients, AI assistance will potentially eliminate the need for biopsies on healthy mucosa, reducing the burden and healthcare costs involved in endoscopy. Additionally, more accurate detection of precancerous conditions will help reduce unnecessary follow-up gastroscopies.

By addressing these areas for improvement, we aim to enhance the accuracy and reliability of our deep learning system for detecting gastric IM.

In conclusion, in this pilot study, we have introduced a novel AI-driven approach for the diagnosis of gastric IM in a Western country and constructed a CNN system using endoscopic image patches from a monocentric hospital that was able to achieve high recall/sensibility and thus a low number of false negatives for the detection of IM in the gastric corpus.

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

Funding: This research was funded by [Sapienza University] grant number [RG123188B449C26A] and the APC was funded by [Sapienza University].

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Sapienza University (protocol code 176SA/2021 and date of approval 11/06/2021).

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

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

Acknowledgments: ED at the Department of Medical-Surgical Sciences and Translational Medicine, Sapienza University of Rome, Italy.

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

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN 146 estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Milano, A.F. 20-year comparative survival and mortality of cancer of the stomach by age, sex, race, stage, grade, cohort entry time-period, disease duration & selected ICD-O-3 oncologic phenotypes: A systematic review of 157,258 cases for diagnosis years 1973–2014:(SEER* Stat 8.3. 4). *J. Insur. Med.* **2019**, *48*, 5–23. [PubMed]
3. Pimentel-Nunes, P.; Libânio, D.; Marcos-Pinto, R.; Areia, M.; Leja, M.; Esposito, G.; Garrido, M.; Kikuste, I.; Megraud, F.; Matysiak-Budnik, T.; et al. Management of epithelial precancerous conditions and lesions in the stomach (maps II): European Society of gastrointestinal endoscopy (ESGE), European Helicobacter and microbiota Study Group (EHMSG), European Society of pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. *Endoscopy* **2019**, *51*, 365–388. [PubMed]
4. Rugge, M.; Correa, P.; Dixon, M.; Fiocca, R.; Hattori, T.; Lechago, J.; Leandro, G.; Price, A.; Sipponen, P.; Solcia, E.; et al. Gastric mucosal atrophy: Interobserver consistency using new criteria for classification and grading. *Aliment. Pharmacol. Ther.* **2002**, *16*, 1249–1259. [CrossRef]
5. Me, D. Classification and Grading of Gastritis. *The Updated Sydney System. Am. J. Surg. Pathol.* **1996**, *20*, 1161–1181.
6. Capelle, L.G.; de Vries, A.C.; Haringsma, J.; Ter Borg, F.; de Vries, R.A.; Bruno, M.J.; van Dekken, H.; Meijer, J.; van Grieken, N.C.; Kuipers, E.J. The staging of gastritis with the OLGA system by using intestinal metaplasia as an accurate alternative for atrophic gastritis. *Gastrointest. Endosc.* **2010**, *71*, 1150–1158. [CrossRef]
7. Lenti, M.V.; Rugge, M.; Lahner, E.; Miceli, E.; Toh, B.H.; Genta, R.M.; De Block, C.; Hershko, C.; Di Sabatino, A. Autoimmune gastritis. *Nat. Rev. Dis. Prim.* **2020**, *6*, 56. [CrossRef]
8. Pimentel-Nunes, P.; Libânio, D.; Lage, J.; Abrantes, D.; Coimbra, M.; Esposito, G.; Hormozdi, D.; Pepper, M.; Drasovean, S.; White, J.R.; et al. A multicenter prospective study of the real-time use of narrow-band imaging in the diagnosis of premalignant gastric 166 conditions and lesions. *Endoscopy* **2016**, *48*, 723–730.
9. ASGE Technology Committee; Song, L.M.; Adler, D.G.; Conway, J.D.; Diehl, D.L.; Farraye, F.A.; Kantsevoy, S.V.; Kwon, R.; Mamula, P.; Rodriguez, B.; et al. Narrow Band Imaging Multiband Imaging. *Gastrointest. Endosc.* **2008**, *67*, 581–589. [CrossRef]
10. Wei, N.; Mulmi Shrestha, S.; Shi, R.H. Markers of gastric intestinal metaplasia under digital chromoendoscopy: Systematic review and meta-analysis. *Eur. J. Gastroenterol. Hepatol.* **2021**, *33*, 470–478. [CrossRef]
11. Esposito, G.; Pimentel-Nunes, P.; Angeletti, S.; Castro, R.; Libânio, D.; Galli, G.; Lahner, E.; Di Giulio, E.; Annibale, B.; Dinis-Ribeiro, M. Endoscopic grading of gastric intestinal metaplasia (EGGIM): A multicenter validation study. *Endoscopy* **2019**, *51*, 515–521. [CrossRef] [PubMed]
12. Castro, R.; Rodriguez, M.; Libânio, D.; Esposito, G.; Pita, I.; Patita, M.; Santos, C.; Pimentel-Nunes, P.; Dinis-Ribeiro, M. Reliability and accuracy of blue light imaging for staging of intestinal metaplasia in the stomach. *Scand. J. Gastroenterol.* **2019**, *54*, 1301–1305. [CrossRef] [PubMed]
13. Pecere, S.; Milluzzo, S.M.; Esposito, G.; Dilaghi, E.; Telese, A.; Eusebi, L.H. Applications of Artificial Intelligence for the Diagnosis of Gastrointestinal Diseases. *Diagnostics* **2021**, *11*, 1575. [CrossRef] [PubMed]
14. Fukushima, K. Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **1980**, *36*, 193–202. [CrossRef] [PubMed]
15. Currie, G.; Hawk, K.E.; Rohren, E.; Vial, A.; Klein, R. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. *J. Med. Imaging Radiat. Sci.* **2019**, *50*, 477–487. [CrossRef] [PubMed]
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:cs.CV/1512.03385.
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:cs.CV/1409.1556.
18. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**, arXiv:cs.CV/1512.00567.
19. Mascarenhas, S.; Agarwal, M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In Proceedings of the 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 19–21 November 2021; IEEE: New York, NY, USA, 2021; Volume 1, pp. 96–99.
20. Du, W.; Rao, N.; Liu, D.; Jiang, H.; Luo, C.; Li, Z.; Gan, T.; Zeng, B. Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images. *IEEE Access* **2019**, *7*, 142053–142069. [CrossRef]
21. Lui, T.K.; Tsui, V.W.; Leung, W.K. Accuracy of artificial intelligence–assisted detection of upper GI lesions: A systematic review and meta-analysis. *Gastrointest. Endosc.* **2020**, *92*, 821–830. [CrossRef]
22. Mohan, B.P.; Khan, S.R.; Kassab, L.L.; Ponnada, S.; Dulai, P.S.; Kochhar, G.S. Accuracy of convolutional neural network-based artificial intelligence in diagnosis of gastrointestinal lesions based on endoscopic images: A systematic review and meta-analysis. *Endosc. Int. Open* **2020**, *8*, E1584–E1594. [CrossRef]
23. Arribas, J.; Antonelli, G.; Frazzoni, L.; Fuccio, L.; Ebigbo, A.; Van Der Sommen, F.; Ghatwary, N.; Palm, C.; Coimbra, M.; Renna, F.; et al. Standalone performance of artificial intelligence for upper GI neoplasia: A meta-analysis. *Gut* **2021**, *70*, 1458–1468. [CrossRef]

24. Jiang, K.; Jiang, X.; Pan, J.; Wen, Y.; Huang, Y.; Weng, S.; Lan, S.; Nie, K.; Zheng, Z.; Ji, S.; et al. Current Evidence and Future Perspective of Accuracy of Artificial Intelligence Application for Early Gastric Cancer Diagnosis With Endoscopy: A Systematic and Meta-Analysis. *Front. Med.* **2021**, *8*, 629080, Erratum in *Front. Med.* **2021**, *8*, 698483. [CrossRef] [PubMed]
25. Matsumoto, K.; Ueyama, H.; Yao, T.; Abe, D.; Oki, S.; Suzuki, N.; Ikeda, A.; Yatagai, N.; Akazawa, Y.; Komori, H.; et al. Diagnostic limitations of magnifying endoscopy with narrow-band imaging in early gastric cancer. *Endosc. Int. Open* **2020**, *8*, E1233–E1242. [CrossRef]
26. Guimarães, P.; Keller, A.; Fehlmann, T.; Lammert, F.; Casper, M. Deep-learning based detection of gastric precancerous conditions. *Gut* **2020**, *69*, 4–6. [CrossRef] [PubMed]
27. Dilaghi, E.; Lahner, E.; Annibale, B.; Esposito, G. Systematic review and meta-analysis: Artificial intelligence for the diagnosis of 184 gastric precancerous lesions and Helicobacter pylori infection. *Dig. Liver Dis.* **2022**, *54*, 1630–1638. [CrossRef] [PubMed]
28. Shi, Y.; Wei, N.; Wang, K.; Tao, T.; Yu, F.; Lv, B. Diagnostic value of artificial intelligence-assisted endoscopy for chronic atrophic gastritis: A systematic review and meta-analysis. *Front. Med.* **2023**, *10*, 1134980. [CrossRef]
29. Li, N.; Yang, J.; Li, X.; Shi, Y.; Wang, K. Accuracy of artificial intelligence-assisted endoscopy in the diagnosis of gastric intestinal metaplasia: A systematic review and meta-analysis. *PLoS ONE* **2024**, *19*, e0303421. [CrossRef]
30. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:cs.LG/1412.6980.
31. Dilaghi, E.; Baldaro, F.; Pilozzi, E.; Conti, L.; Palumbo, A.; Esposito, G.; Annibale, B.; Lahner, E. Pseudopyloric Metaplasia Is Not Associated With the Development of Gastric Cancer. *Am. J. Gastroenterol.* **2021**, *116*, 1859–1867. [CrossRef]
32. Zhang, K.; Guo, Y.; Wang, X.; Yuan, J.; Ding, Q. Multiple feature reweight densenet for image classification. *IEEE Access* **2019**, *7*, 9872–9880. [CrossRef]
33. Koonce, B.; Koonce, B. *EfficientNet. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Apress: New York, NY, USA, 2021; pp. 109–123.
34. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
35. Xu, M.; Zhou, W.; Wu, L.; Zhang, J.; Wang, J.; Mu, G.; Huang, X.; Li, Y.; Yuan, J.; Zeng, Z.; et al. Artificial intelligence in the diagnosis of gastric precancerous conditions by image-enhanced endoscopy: A multicenter, diagnostic study (with video). *Gastrointest. Endosc.* **2021**, *94*, 540–548.e4. [CrossRef] [PubMed]
36. Song, M.; Kwek, A.B.; Law, N.M.; Ong, J.P.L.; Tan, J.Y.-L.; Thurairajah, P.H.; Ang, D.S.W.; Ang, T.L. Efficacy of small-volume simethicone given at least 30 min before gastroscopy. *World J. Gastrointest. Pharmacol. Ther.* **2016**, *7*, 572–578. [CrossRef] [PubMed]
37. Bisschops, R.; Areia, M.; Coron, E.; Dobru, D.; Kaskas, B.; Kuvaev, R.; Pech, O.; Ragunath, K.; Weusten, B.; Familiari, P.; et al. Performance measures for upper gastrointestinal endoscopy: A European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. *Endoscopy* **2016**, *48*, 843–864. [CrossRef] [PubMed]
38. Song, Y.Q.; Mao, X.L.; Zhou, X.B.; He, S.Q.; Chen, Y.H.; Zhang, L.H.; Xu, S.-W.; Yan, L.-L.; Tang, S.-P.; Ye, L.-P.; et al. Use of artificial intelligence to improve the quality control of gastrointestinal endoscopy. *Front. Med.* **2021**, *8*, 709347. [CrossRef] [PubMed]
39. Yang, K.Y.; Mukundan, A.; Tsao, Y.M.; Shi, X.H.; Huang, C.W.; Wang, H.C. Assessment of hyperspectral imaging and CycleGAN-simulated narrowband techniques to detect early esophageal cancer. *Sci. Rep.* **2023**, *13*, 20502. [CrossRef]
40. Liao, W.C.; Mukundan, A.; Sadiaza, C.; Tsao, Y.M.; Huang, C.W.; Wang, H.C. Systematic meta-analysis of computer-aided detection to detect early esophageal cancer using hyperspectral imaging. *Biomed. Opt. Express* **2023**, *14*, 4383–4405. [CrossRef]

diagnostics

MDPI

Article

Semi-Supervised Facial Acne Segmentation Using Bidirectional Copy–Paste

Semin Kim, Huisu Yoon and Jongha Lee *

AI R&D Center, lululab, Dosan Dae-Ro 318, Seoul 06054, Republic of Korea; sm.kim@lulu-lab.com (S.K.); hs.yoon@lulu-lab.com (H.Y.)
* Correspondence: jongha.lee@lulu-lab.com

Abstract: Facial acne is a prevalent dermatological condition regularly observed in the general population. However, it is important to detect acne early as the condition can worsen if not treated. For this purpose, deep-learning-based methods have been proposed to automate detection, but acquiring acne training data is not easy. Therefore, this study proposes a novel deep learning model for facial acne segmentation utilizing a semi-supervised learning method known as bidirectional copy–paste, which synthesizes images by interchanging foreground and background parts between labeled and unlabeled images during the training phase. To overcome the lower performance observed in the labeled image training part compared to the previous methods, a new framework was devised to directly compute the training loss based on labeled images. The effectiveness of the proposed method was evaluated against previous semi-supervised learning methods using images cropped from facial images at acne sites. The proposed method achieved a Dice score of 0.5205 in experiments utilizing only 3% of labels, marking an improvement of 0.0151 to 0.0473 in Dice score over previous methods. The proposed semi-supervised learning approach for facial acne segmentation demonstrated an improvement in performance, offering a novel direction for future acne analysis.

Keywords: acne segmentation; semi-supervised learning; bidirectional copy–paste; deep learning; semantic segmentation

Citation: Kim, S.; Yoon, H.; Lee, J. Semi-Supervised Facial Acne Segmentation Using Bidirectional Copy–Paste. *Diagnostics* **2024**, *14*, 1040. https://doi.org/10.3390/diagnostics14101040

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 5 April 2024
Revised: 12 May 2024
Accepted: 14 May 2024
Published: 17 May 2024

1. Introduction

Facial skin disorders commonly occur among people, which is why various studies are being conducted to detect them [1–4]. Out of these, acne is a prevalent skin disorder that frequently occurs in the general population. Untreated acne has the potential to deteriorate or result in scarring. Therefore, multiple research investigations are currently being conducted to detect and classify acne based on facial images.

Figure 1 displays images of acne cropped from facial images, highlighting the variety in the color and shape of acne. Initially, acne detection primarily involved extracting features based on color or texture and utilizing classifiers. Budihi et al. [5] applied the region growing method based on pixel color similarity in facial skin images to select candidate areas for acne. Additionally, a self-organizing map was used to diagnose acne. Alamdari et al. [6] used techniques such as K-means clustering, texture analysis, and color-based segmentation to segment acne areas. Yadav et al. [7] identified candidate regions of acne presence on the face based on the hue–saturation–value (HSV) color space and then distinguished acne using classifiers trained with a support vector machine (SVM). However, these methods struggle with setting threshold values for classifying acne's color or texture. Moreover, defining a descriptor that reflects the diverse characteristics of acne accurately is challenging for humans. Recently, methods based on deep learning have been proposed to detect acne, overcoming the aforementioned issues. Rashataprucksa et al. [8] and Hyunh et al. [9] utilized object detection models such as the faster region-based convolutional neural network (Faster-RCNN) [10] and region-based fully convolutional networks

(R-FCN) [11] for acne detection. Min et al. [12] employed a dual encoder based on a convolutional neural network (CNN) and Transformer to detect face acne. Junayed et al. [13] also proposed a dual encoder based on CNN and Transformer, but they detected acne through a semantic segmentation approach. Kim et al. [14] enhanced the performance of acne segmentation by training on the positional information of acne in the final encoder.

Deep learning techniques fundamentally require labeled training data. However, labeling is time consuming and costly, and even securing original medical data, such as for acne, can be challenging. Recently, to partially address the difficulty of acquiring labeled data, much research has been conducted on semi-supervised learning, which utilizes both labeled and unlabeled data for training. Initially, various data augmentation techniques were applied to unlabeled data, employing consistency regularization [15]. Methods like Cutmix [16], which overlap parts of different images, have significantly aided in improving semi-supervised learning. Recently, a method that uses bidirectional copy–paste (BCP) alternately on the foreground and background between labeled and unlabeled data has been proposed for medical image segmentation [17]. This proposed method performs semi-supervised learning while maintaining a bidirectional relationship between the two sets of data. Additionally, to overcome the lack of training data, it was proposed to use generative models like StyleGAN2 [18,19] to create images for use in acne training [20,21].

Figure 1. Various acne samples. The shape and skin color of acne are considerably diverse.

We conducted semi-supervised learning of the acne segmentation model using BCP. We aimed to create an acne segmentation model based on semi-supervised learning using the BCP method. However, in BCP, training was conducted only with synthetic images created through copy–paste between labeled and unlabeled images. This approach made it challenging to fully reflect the characteristics of the labeled images. Although BCP performed well with computed tomography (CT) images, we discovered shortcomings in segmenting acne across various shapes and skin tones. To address this issue, we proposed adding a structure to the BCP framework that is directly trained on input labeled images. Thus, our proposed method aims to maintain the BCP structure while enabling semi-supervised learning and improving acne segmentation performance by learning from input labeled images. Additionally, in the ablation study and discussion, we conducted acne segmentation experiments using ACNE04 [22], which has similar skin tones, and analyzed issues related to this.

To verify the performance of our proposed method, we compared the acne segmentation performance with that of previous semi-supervised learning methods. For this purpose, we used images cropped primarily around acne from facial images, as illustrated

in Figure 1. In this paper's experiment, we compared the performance of acne segmentation with conventional semi-supervised learning methods by varying the proportion of labeled images during the training phase. The results showed that the proposed method achieved the highest Dice score and Jaccard index compared to previous semi-supervised methods. Notably, the superior performance over BCP demonstrates that training with both labeled images and synthetic images, rather than just synthetic images, is effective for acne segmentation applications. The main contributions of the proposed method are as follows:

- Fusion of labeled loss and synthetic loss: We propose a method that simultaneously calculates and fuses the labeled loss for training labeled images and the synthetic loss for training unlabeled images;
- Comparison of acne segmentation performance with semi-supervised learning methods: We compared the acne segmentation performance with previous semi-supervised learning methods based on our acne database and ACNE04. Additionally, through ablation studies, we compared the acne segmentation performance as the parameters of U-Net were increased.

The structure of this paper is as follows: Section 2 provides a brief introduction to acne segmentation research and semi-supervised learning. Section 3 describes the semi-supervised learning method proposed in this paper, and Section 4 presents the experimental results of the proposed method. Section 5 shows an ablation study, required when adding the training loss from input labeled images to the total loss, and Section 6 discusses the proposed method. Finally, Section 7 summarizes the conclusions of this paper.

2. Related Works

In this section, we briefly review previous acne detection methods based on deep learning and semi-supervised learning approaches to overcome the challenge of insufficient labeling.

2.1. Acne Detection

Rashataprucksa et al. [8] employed object detection models Faster-RCNN [10] and R-FCN [11] to detect acne in facial images, comparing these two models to evaluate their respective acne detection capabilities. Min et al. [12] utilized a dual encoder composed of a CNN and Transformer to extract features which were then processed through dynamic context enhancement and mask-aware multi-attention for final acne detection. Similarly, Junayed et al. [13] approached acne detection through semantic segmentation, employing a dual encoder setup with a CNN and Transformer. This method involved extracting both local and global information which was then integrated through a feature versatile block to the decoder. Kim et al. [14] segmented acne using a U-Net [23,24] structure and applied center point loss to train the last encoder on acne location information. Acne detection has primarily been conducted through either object detection or semantic segmentation. However, since the shape of acne can help differentiate the severity of the condition and assist in treatment [25], our paper aims to detect acne based on semantic segmentation.

2.2. Semi-Supervised Learning

Semi-supervised learning primarily applies various data augmentations to unlabeled images, focusing on consistency-based learning. FixMatch [26] extracts prediction probabilities by applying different levels of data augmentation to unlabeled data. Then, it ensures that the predictions from strongly augmented data maintain consistency based on the prediction probabilities of weakly augmented data. However, in FixMatch, only unlabeled data with predicted probabilities above a certain threshold were used for training, leading to the exclusion of many unlabeled data instances. To overcome this limitation, Full-Match [27] introduced adaptive negative learning to improve training performance. Furthermore, Wu et al. [28] applied pixel smoothness and inter-class separation in semi-supervised learning to address the blurring of pixels in edge or low-contrast areas. UniMatch [29] improved semi-supervised performance by applying stronger data augmentation in a dual structure.

Notably, for strong data augmentation, Cutmix [16] was applied, which involves inserting specific parts of an image into another image. Recently, a method similar to Cutmix, BCP [17], was proposed for medical image segmentation. This method pairs labeled and unlabeled data, overlapping specific parts of their images in a manner similar to Cutmix. Semi-supervised learning is then performed using a teacher and student network structure. Loss is calculated by comparing the labeled data region with the actual ground truth, while the loss for the unlabeled data region is determined using the pseudo ground truth from the teacher network. Given BCP's proven effectiveness in medical image segmentation, this paper aims to leverage BCP to improve facial acne segmentation performance.

3. Method

In this section, we provide a detailed explanation of the proposed method. First, we present the overall structure and explain the basic principles of BCP. Then, we show how the training loss is calculated in the proposed method.

3.1. Overall Structure

Figure 2 illustrates the overall structure proposed in this paper. It is fundamentally composed of a teacher network T and a student network S, similar to the BCP method [17]. Both T and S are constructed with U-Net [23], where the channel count of the first encoder is 16, which is the same model as BCP. Initially, T and S utilize the same pre-trained weights, Θ_p, which are trained through supervised learning using only labeled images. However, unlike previous approaches that trained Θ_p with synthetic images created by applying BCP among labeled images, this study employs labeled images. This is because using labeled images to generate Θ_p resulted in better performance than using BCP for acne segmentation. Algorithm 1 shows the overall pseudo code.

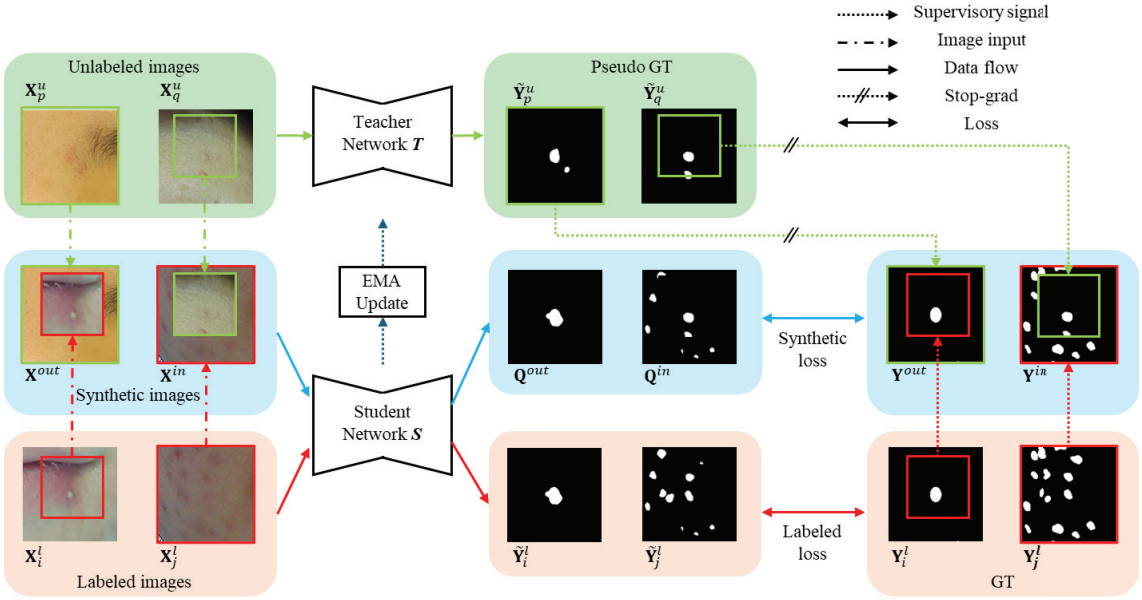


Figure 2. Overall structure of the proposed method. Synthetic images are generated using the bidirectional copy–paste method. The labeled image and synthetic image are each inferred for prediction values through the student network. Then, synthetic GT is generated using ground truth (GT) and pseudo GT. The training loss is calculated by sending a supervisory signal to the student network through each GT. Once the student network is trained, an EMA update is applied to the teacher network.

Algorithm 1 Training process of the proposed acne segmentation model.

Input: labeled images $\mathbf{X}^l$, labels $\mathbf{Y}^l$, unlabeled images $\mathbf{Y}^u$
Output: trained the student weights Θ_s

Step 1: Preparing
1.1 Setting α as a weight of unlabeled images
1.2 Setting γ as a weight of labeled loss
1.3 Setting λ as a weight of EMA update for Θ_t
1.4 Setting η as a learning rate
1.5 Initializing Θ_p
Step 2: Training the pre-trained weights Θ_p
2.1 Training and selecting the best Θ_p
* 2a. Computing labeled losses
* 3$\mathcal{L}_i^l = \mathcal{L}_{\text{seg}}(f(\mathbf{X}_i^l;\Theta_p),\mathbf{Y}_i^j), \quad \mathcal{L}_j^l = \mathcal{L}_{\text{seg}}(f(\mathbf{X}_j^l;\Theta_p),\mathbf{Y}_j^l)$
* 2b. Updating Θ_p
* 3 $\Theta_p = \Theta_p - \eta\nabla_{\Theta_p}\left(\mathcal{L}_i^l + \mathcal{L}_j^l\right)$
Step 3: Training the student weights Θ_s
3.1 Initializing $\Theta_t = \Theta_p$, $\Theta_s = \Theta_p$
3.2 Training and selecting the best Θ_s
a. Generating synthetic images $\mathbf{X}^{in}$ and $\mathbf{X}^{out}$ by cropping and pasting $\mathbf{X}^l$ and $\mathbf{X}^u$
b. Generating pseudo GT $\tilde{\mathbf{Y}}^u$ by $f(\mathbf{X}^u;\Theta_t)$
c. Generating synthetic GT $\mathbf{Y}^{in}$ and $\mathbf{Y}^{out}$ by cropping and pasting $\mathbf{Y}^l$ and $\tilde{\mathbf{Y}}^u$
* 2d. Computing synthetic losses with a mask $\mathbf{M}$
* 3$\mathcal{L}^{in} = \mathcal{L}_{\text{seg}}(f(\mathbf{X}^{in};\Theta_s),\mathbf{Y}^{in}) \odot \mathbf{M} + \alpha \times \mathcal{L}_{\text{seg}}(f(\mathbf{X}^{in};\Theta_s),\mathbf{Y}^{in}) \odot (\mathbf{1}-\mathbf{M})$
* 3$\mathcal{L}^{out} = \mathcal{L}_{\text{seg}}(f(\mathbf{X}^{out};\Theta_s),\mathbf{Y}^{out}) \odot (\mathbf{1}-\mathbf{M}) + \alpha \times \mathcal{L}_{\text{seg}}(f(\mathbf{X}^{out};\Theta_s),\mathbf{Y}^{out}) \odot \mathbf{M}$
* 2e. Computing labeled losses
* 3$\mathcal{L}_i^l = \mathcal{L}_{\text{seg}}(f(\mathbf{X}_i^l;\Theta_p),\mathbf{Y}_i^j), \quad \mathcal{L}_j^l = \mathcal{L}_{\text{seg}}(f(\mathbf{X}_j^l;\Theta_p),\mathbf{Y}_j^l)$
* 2f. Updating Θ_s
* 3$\Theta_s = \Theta_s - \eta\left(\nabla_{\Theta_s}\left(\mathcal{L}^{in} + \mathcal{L}^{out} + \gamma\left(\mathcal{L}_i^l + \mathcal{L}_j^l\right)\right)\right)$
* 2g. EMA Updating Θ_t
* 3$\Theta_t = \lambda \times \Theta_t + (1-\lambda)\Theta_s$

3.2. Pre-Trained Weight

In Step 3 of Algorithm 1, the weights Θ_t of the teacher network T need to be initialized with the pre-trained weights Θ_p. The pre-trained weights are trained solely on labeled images in Step 2, which is the same as in typical supervised learning. By setting Θ_p as the initial values for Θ_t and Θ_s, semi-supervised learning based on BCP becomes possible.

3.3. Bidirectional Copy–Paste for Synthetic Images

The method of generating synthetic images using BCP is as follows. First, a mask $\mathbf{M}$ of the same size as the images is created. $\mathbf{M}$ consists of zeros and ones, where the area of ones becomes the region to be copied. Then, $\mathbf{M}$ is applied to Equations (1) and (2) to generate $\mathbf{X}^{in}$ and $\mathbf{X}^{out}$.

$$\mathbf{X}^{in} = \mathbf{X}_j^l \odot \mathbf{M} + \mathbf{X}_p^u \odot (\mathbf{1}-\mathbf{M}), \tag{1}$$

$$\mathbf{X}^{out} = \mathbf{X}_q^u \odot \mathbf{M} + \mathbf{X}_i^l \odot (\mathbf{1}-\mathbf{M}), \tag{2}$$

where $\mathbf{X}_i^l$ and $\mathbf{X}_j^l$ are the i-th and j-th labeled images, respectively ($i \neq j$), and $\odot$ represents element-wise multiplication. $\mathbf{X}_p^u$ and $\mathbf{X}_q^u$ are the p-th and q-th unlabeled images, respectively ($p \neq q$). $\mathbf{1}$ represents a matrix of the same size as $\mathbf{M}$, with all elements being 1. Figure 3 provides a detailed example of a sample image generated through BCP.

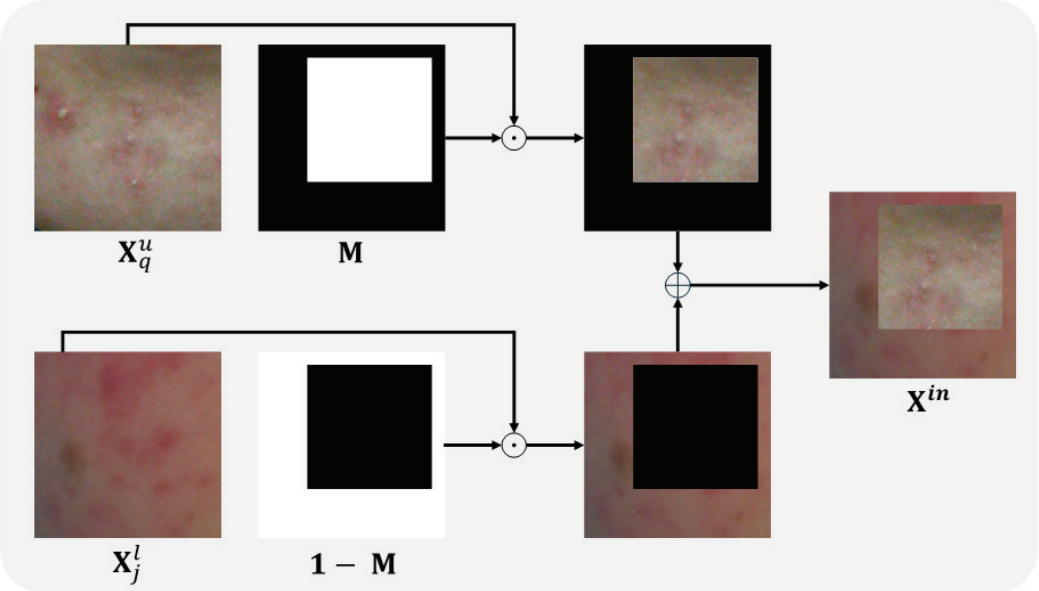


Figure 3. An example of creating $\mathbf{X}^{in}$ by applying BCP to an unlabeled image $\mathbf{X}_q^u$ and a labeled image $\mathbf{X}_j^l$. In mask $\mathbf{M}$, 1 represents the white area, and 0 represents the black area. $\mathbf{1}$ represents a matrix of the same size as $\mathbf{M}$, with all elements being 1. $\odot$ is element-wise multiplication, and $\oplus$ is element-wise addition.

3.4. Pseudo Synthetic Ground Truth for Supervisory Signals

The pseudo GT for unlabeled images is generated through the teacher network $\boldsymbol{T}$. First, the prediction values for the unlabeled images $\mathbf{X}_p^u$ and $\mathbf{X}_q^u$ are extracted using the equation below.

$$\mathbf{P}_p^u = f(\mathbf{X}_p^u; \Theta_t), \mathbf{P}_q^u = f(\mathbf{X}_q^u; \Theta_t), \tag{3}$$

where f is a network model.

Then, as in Equation (3), a binarized pseudo GT is generated using the equation below.

$$\tilde{\mathbf{Y}}_p^u(i,j) = \begin{cases} 1 & \text{if } \mathbf{P}_p^u(i,j) > 0.5 \\ 0 & \text{otherwise} \end{cases}, \quad \tilde{\mathbf{Y}}_q^u(i,j) = \begin{cases} 1 & \text{if } \mathbf{P}_q^u(i,j) > 0.5 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where i and j are coordinates.

Next, the formula below is applied to the ground truths $\mathbf{Y}_i^l$ and $\mathbf{Y}_j^l$ of $\mathbf{X}_i^l$ and $\mathbf{X}_j^l$, respectively, to generate $\mathbf{Y}^{in}$ and $\mathbf{Y}^{out}$, which are synthetic ground truths.

$$\mathbf{Y}^{in} = \mathbf{Y}_j^l \odot \mathbf{M} + \tilde{\mathbf{Y}}_p^u \odot (\mathbf{1} - \mathbf{M}), \tag{5}$$

$$\mathbf{Y}^{out} = \tilde{\mathbf{Y}}_q^u \odot \mathbf{M} + \mathbf{Y}_i^l \odot (\mathbf{1} - \mathbf{M}). \tag{6}$$

3.5. Semi-Supervised Loss Computation

To calculate the training loss for the student network $\boldsymbol{S}$, the labeled images ($\mathbf{X}_i^l$, $\mathbf{X}_j^l$) and synthetic images ($\mathbf{X}^{in}$, $\mathbf{X}^{out}$) are each inferred through the student network $\boldsymbol{S}$ as $\mathbf{Q}_i^l$, $\mathbf{Q}_j^l$, $\mathbf{Q}^{in}$, and $\mathbf{Q}^{out}$, respectively.

$$\mathbf{Q}^{in} = f(\mathbf{X}^{in}; \Theta_s), \quad \mathbf{Q}^{out} = f(\mathbf{X}^{out}; \Theta_s), \quad \mathbf{Q}_i^l = f(\mathbf{X}_i^l; \Theta_s), \quad \mathbf{Q}_j^l = f(\mathbf{X}_j^l; \Theta_s). \tag{7}$$

Then, the training loss for each corresponding GT is calculated using Equations (8) through (10). $\mathcal{L}_{\text{seg}}$ is the linear combination of Dice loss and cross-entropy.

$$\mathcal{L}^{in} = \mathcal{L}_{\text{seg}}(\mathbf{Q}^{in}, \mathbf{Y}^{in}) \odot \mathbf{M} + \alpha \times \mathcal{L}_{\text{seg}}(\mathbf{Q}^{in}, \mathbf{Y}^{in}) \odot (\mathbf{1} - \mathbf{M}), \tag{8}$$

$$\mathcal{L}^{out} = \mathcal{L}_{\text{seg}}(\mathbf{Q}^{out}, \mathbf{Y}^{out}) \odot (\mathbf{1} - \mathbf{M}) + \alpha \times \mathcal{L}_{\text{seg}}(\mathbf{Q}^{out}, \mathbf{Y}^{out}) \odot \mathbf{M}, \tag{9}$$

$$\mathcal{L}_i^l = \mathcal{L}_{\text{seg}}(\mathbf{Q}_i^l, \mathbf{Y}_i^j), \quad \mathcal{L}_j^l = \mathcal{L}_{\text{seg}}(\mathbf{Q}_j^l, \mathbf{Y}_j^l), \tag{10}$$

where α represents the weight of the unlabeled images. The final training loss is calculated using Equation (11).

$$\mathcal{L} = \mathcal{L}^{in} + \mathcal{L}^{out} + \gamma(\mathcal{L}_i^l + \mathcal{L}_j^l), \tag{11}$$

where γ is a parameter that adjusts the weight of the purely supervised loss. Using the above loss, the student network $\boldsymbol{S}$ is ultimately updated. Subsequently, an Exponential Moving Average (EMA) update is performed on the teacher network $\boldsymbol{T}$ using Equation (12).

$$\Theta_t = \lambda \times \Theta_t + (1 - \lambda)\Theta_s, \tag{12}$$

where Θ_t and Θ_s represent the parameters of the teacher network $\boldsymbol{T}$ and the student network $\boldsymbol{S}$, respectively.

4. Experimental Results

In this section, we analyze the performance of the proposed method for acne segmentation and compare it with previous semi-supervised learning methods. First, we describe the experimental setup and then proceed to compare the acne segmentation performance with previous semi-supervised methods.

4.1. Experimental Setup

To validate the performance of our proposed acne segmentation method, we acquired images of acne from facial images taken with skin diagnostic equipment [30]. Each acne image is cropped around the acne, as shown in Figure 1, and scaled to 256 × 256. We collected a total of 2000 acne images, of which 1600 were designated as the training set and the remaining 400 as the evaluation set. The optimizer used for network training was a stochastic gradient descent (SGD) with a learning rate of 0.01, momentum of 0.9, and weight decay set to 0.0001. The batch size was set to 24, comprising 12 labeled and 12 unlabeled data. Based on BCP [17], alpha was set to 0.5, and lambda was set to 0.99. The size of the area for mask 1 was set to 2/3 of the input image. Gamma was set to 0.5 according to our ablation study. Pre-training iterations were set to 10k, and semi-supervised learning iterations were set to 30k. The training evaluation was compared using Dice score and Jaccard index. All experiments were conducted on an RTX 4090, Ubuntu 20.04, Pytorch 2.1.1.

4.2. Comparison of Results

4.2.1. Comparison between Synthetic Images and Labeled Images for Pre-Trained Weight

Pre-trained weights are trained through supervised learning. In the original approach, BCP was applied to labeled images to create synthetic images for training pre-trained weights. Our method, however, utilizes labeled images directly without applying BCP for training pre-trained weights. Thus, a comparison between these two approaches was initially conducted.

Table 1 presents the results of training with different proportions of labeled data at 3% and 7%. As shown in Table 1, generating pre-trained weights with labeled images demonstrated superiority in three metrics over using synthetic images. Therefore, we opted to train with labeled images, which, overall, provided better performance for generating pre-trained weights Θ_p.

Table 1. Comparison between synthetic images and labeled images for generating pre-trained weights Θ_p.

Loss Type	Ratio		Metrics	
	Labeled	Unlabeled	Dice Score	Jaccard Index
Synthetic loss	3%	0%	0.4423	0.3108
Labeled loss	3%	0%	**0.4570**	**0.3203**
Synthetic loss	7%	0%	0.4784	0.3425
Labeled loss	7%	0%	**0.4951**	**0.3517**

4.2.2. Semi-Supervised Learning Comparison

We compared the semi-supervised learning performance of our proposed method with previous methods. Table 2 lists the performance of our proposed method against comparison methods across various metrics. Our method was trained in a semi-supervised manner, as proposed in Section 3, based on the pre-trained weights learned in Section 4.2.1. As shown in Table 2, our proposed method exhibited the highest performance. This suggests that training with both BCP-based synthetic images and labeled images simultaneously provided mutual benefits, leading to the superior performance of our proposed method. Figure 4 presents examples of results from each method when using 7% labeled images.

Table 2. Comparison of acne segmentation performance of the proposed method and previous semi-supervised learning methods.

Method	Ratio		Metrics	
	Labeled	Unlabeled	Dice Score	Jaccard Index
SS-Net [28]	3%	97%	0.4732	0.3333
BCP [17]	3%	97%	0.5054	0.3617
Ours	3%	97%	**0.5251**	**0.3777**
SS-Net [28]	7%	93%	0.5162	0.3750
BCP [17]	7%	93%	0.5357	0.3912
Ours	7%	93%	**0.5603**	**0.4117**

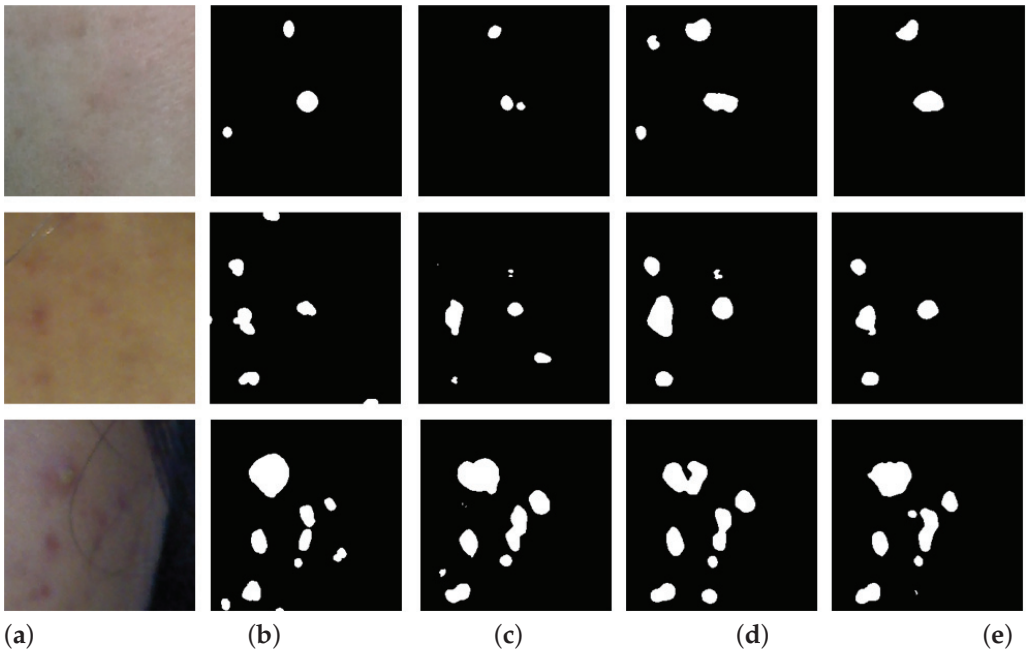
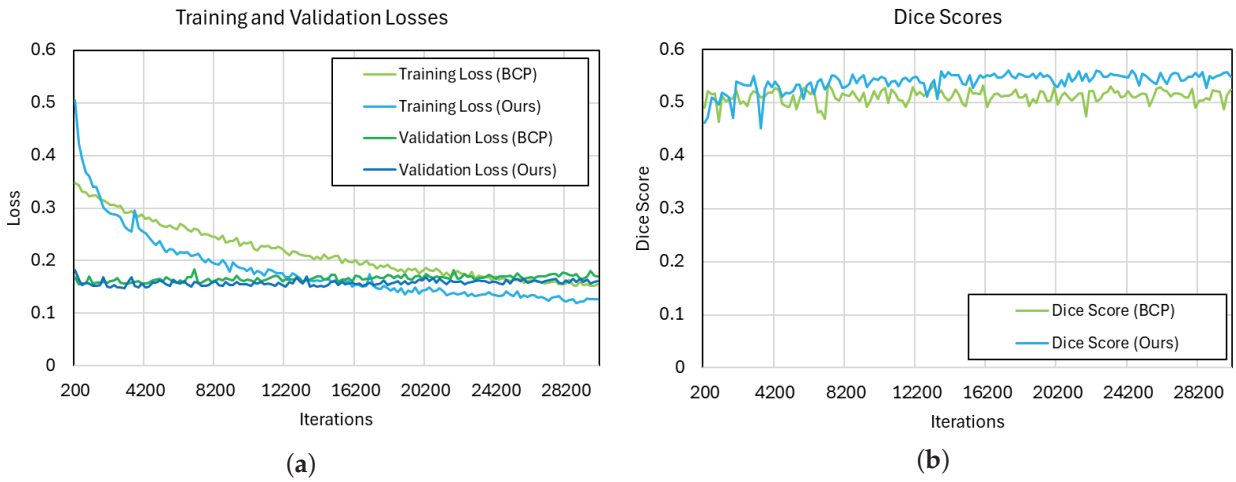


Figure 4. Examples of acne segmentation results from the compared semi-supervised methods. (**a**) represents the input images, (**b**) is the ground truth. (**c**) is the result of SS-Net, (**d**) is the result of BCP, and (**e**) is the result of the proposed method.

Additionally, to further compare our method with BCP, we examined the training and validation loss and the Dice score. We set the proportion of labeled images at 7% and calculated the loss using each method for comparison, as shown in Figure 5. In our method, because labeled loss is added, the initial loss is higher than BCP. However, as training progresses, it becomes lower than BCP. Especially during training, the validation loss is mostly lower than BCP. Therefore, by comparing the Dice score, we can confirm that the proposed method's acne segmentation performance is clearly higher than that of BCP.



(**a**) (**b**)

Figure 5. Comparison of training and validation loss and Dice score for our method and the BCP method. (**a**) shows the training and validation losses for each method, and (**b**) shows the Dice scores.

5. Ablation Study

In this section, we analyze the acne segmentation performance based on the gamma value used to fuse synthetic loss and labeled loss, and the number of parameters in U-Net, within the proposed method. Additionally, we compared the acne segmentation performance of the proposed method with BCP using the public database ACNE04.

5.1. Performance Variation according to γ and the Number of Channels in U-Net

In this section, we present the results of acne segmentation according to different γ values used in Equation (11) for combining synthetic image loss ($\mathcal{L}^{in} + \mathcal{L}^{out}$) and labeled image loss ($\mathcal{L}_i^l + \mathcal{L}_j^l$). We tested three scenarios with γ values of 0.1, 0.5, and 1.0. Generally, using $\gamma = 0.5$ resulted in superior overall performance, shown in Table 3. While there were differences in some scores, high performance was demonstrated in the Dice score and Jaccard indexes, which consider overall performance. Therefore, in our experiments, we used $\gamma = 0.5$.

Table 3. Comparison of acne segmentation performance based on γ.

γ	Ratio		Metrics	
	Labeled	**Unlabeled**	**Dice Score**	**Jaccard Index**
0.1	3%	97%	0.5177	0.3693
0.5	3%	97%	**0.5251**	**0.3777**
1.0	3%	97%	0.5205	0.3753
0.1	7%	93%	0.5522	0.4060
0.5	7%	93%	**0.5603**	**0.4122**
1.0	7%	93%	0.5588	0.4117

Next, we experimented with increasing the number of channels in the first encoder of the UNet-BN used in our experiments. While BCP set the number of channels at 16, we conducted comparative experiments with increased numbers at 32 and 64. Table 4 shows the acne segmentation performance when the number of channels was increased. As the number of channels increased, performance metrics also improved. Therefore, based on using 7% labeled images for acne segmentation, compared to using BCP, the Dice score increased by 0.0424 and the Jaccard index by 0.0359.

Table 4. Comparison of acne segmentation performance based on changes in the number of channels in the first encoder.

# Channels	Ratio		Metrics	
	Labeled	Unlabeled	Dice Score	Jaccard Index
16 (BCP [17])	3%	97%	0.5054	0.3617
16 (ours)	3%	97%	0.5251	0.3777
32 (ours)	3%	97%	0.5394	0.3912
64 (ours)	3%	97%	**0.5458**	**0.3965**
16 (BCP [17])	7%	93%	0.5357	0.3912
16 (ours)	7%	93%	0.5603	0.4117
32 (ours)	7%	93%	0.5709	0.4233
64 (ours)	7%	93%	**0.5781**	**0.4271**

5.2. Acne Segmentation Performance on ACNE04

In this subsection, we aim to compare acne segmentation performance using the public acne database ACNE04 [22]. However, ACNE04 does not have annotations for semantic segmentation. Instead, it provides bounding boxes for object detection. Using this bounding box information, we generated pseudo ground truth for semantic segmentation, as shown in Figure 6, by drawing circles passing through the center of each side of the bounding boxes. Because of the diversity of acne shapes, they do not accurately reflect the actual boundaries of acne lesions. As a result, this experiment focused on understanding the trend in performance differences with BCP rather than precise accuracy.

Originally, the ACNE04 database contained a total of 1457 images. However, some images have poor quality or contain watermarks. We excluded these and selected 1108 images that had good quality and similar shooting conditions. We selected 222 images for validation and the remaining ones for training. We cropped the selected images as described in Section 4, yielding 1600 training patches and 400 validation patches. Table 5 displays the results of semi-supervised learning using the proposed method and BCP. While the proposed method outperforms BCP, it yields a somewhat reduced improvement margin when compared to our acne database. This limitation is attributed to the crude pseudo ground truth and the characteristics of the images in ACNE04, which will be further analyzed in the Discussion section. Nevertheless, using both synthetic loss and labeled loss simultaneously helped to improve acne segmentation performance.

Table 5. Comparison of acne segmentation results using semi-supervised learning with ACNE04's pseudo ground truth.

Methods	Ratio		Metrics	
	Labeled	Unlabeled	Dice Score	Jaccard Index
BCP [17]	3%	97%	0.5103	0.3642
Ours	3%	97%	**0.5159**	**0.3702**
BCP [17]	7%	93%	0.5664	0.4141
Ours	7%	93%	**0.5749**	**0.4212**

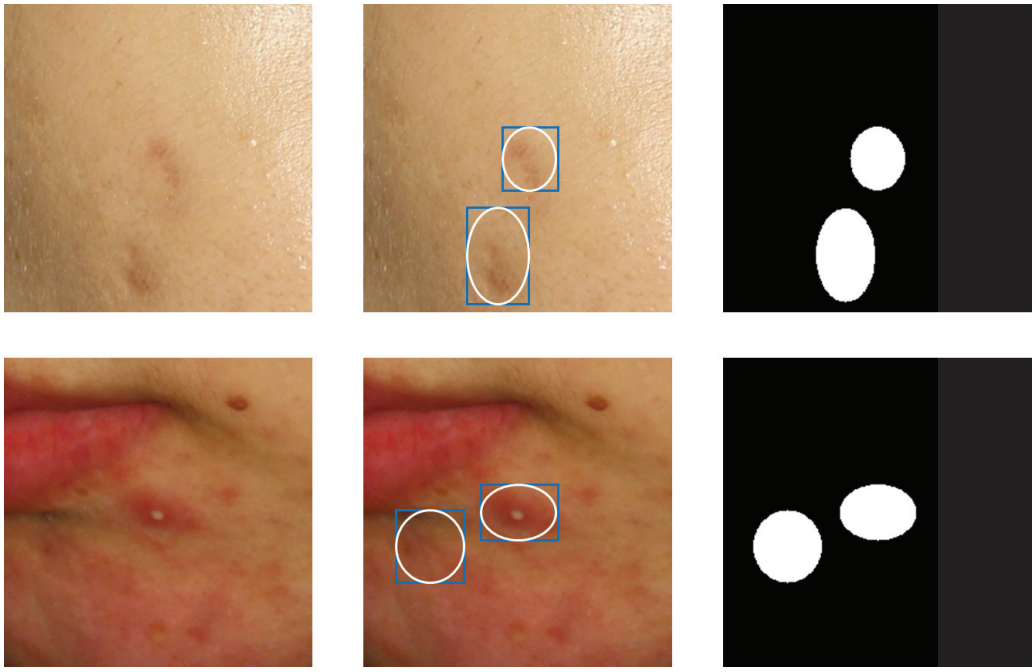


Figure 6. Example of creating semantic segmentation ground truth using ACNE04's bounding boxes. (**a**) is the original image, and (**b**) shows the blue boxes indicating acne with bounding boxes. By drawing ellipses inside the bounding boxes, pseudo ground truth for acne is generated as shown in (**c**).

6. Discussion

The proposed method showed improved performance for our data compared to the original BCP method. However, as observed in the ablation study in Section 5, there was a decrease in acne segmentation performance improvement in ACNE04. To analyze this, we compared synthetic images composed from each dataset, as depicted in Figure 7. Our data include a variety of skin colors and diverse lighting conditions, while the ACNE04 dataset used in the experiments is predominantly composed of East Asian skin tones. Therefore, even when creating synthetic images using ACNE04, as shown in Figure 7, the images composited in the foreground exhibit less disparity and lower color gradation compared to those composed with our data. Nevertheless, actual human skin tones vary widely in color depending on race and environmental conditions. Therefore, it is expected that our method will demonstrate superior performance in actual acne segmentation compared to the original BCP approach.

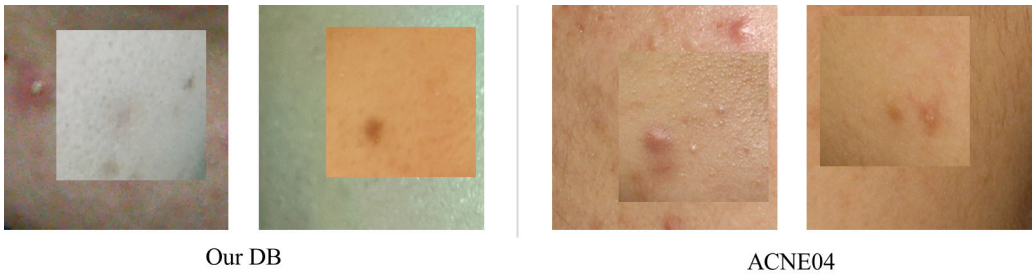


Figure 7. Comparison of synthetic images. The two images on the left are synthetic images generated from our database, and those on the right are from ACNE04. Our database reflects a variety of skin tones and lighting, resulting in a significant color difference between the duplicated inner images and the background images. In contrast, the ACNE04 images are synthesized in relatively similar colors.

7. Conclusions

In this paper, a semi-supervised learning method for training acne segmentation models is proposed. The original BCP method, which calculates the loss solely on synthetic images, was insufficient for detecting acne across diverse skin tones. To address this, the training process was enhanced by including original images to improve overall acne segmentation performance. The proposed method was compared with previous semi-supervised learning methods on acne patch images and demonstrated superior performance based on evaluation metrics such as the Dice score and Jaccard index. However, the performance improvement was less significant for ACNE04, where the skin color and lighting are similar. Since actual human skin color and lighting vary, performance improvements like those in the main results of this paper are expected in real applications. Future research aims to apply the proposed method across the medical imaging field to enhance performance in various areas.

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

Funding: This research was financially supported by Deeptech Incubator Projects Startups 1000+ (No. 20228951) funded by Korea Institute of Startup & Entrepreneurship Development (KISED).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Severance Hospital (2022-31-1043).

Informed Consent Statement: Facial images were collected with informed consent from all subjects, in accordance with all relevant national regulations and institutional policies, and approved by the authors' institutional committees.

Data Availability Statement: Data are contained within the article.

Acknowledgments: This research was conducted with the support of the Custom Growth Ladder (Scale-Up)—Multichannel Device Project, 2024 High-Growth Software Club Business, DIPS 1000+ (Deeptech Incubator Projects Startups 1000+) and the "HPC Support" Project, supported by the Ministry of Science and ICT and NIPA.

Conflicts of Interest: Authors Semin Kim , Huisu Yoon and Jongha Lee were employed by the company lululab.

References

1. Mekonnen, B.; Hsieh, T.; Tsai, D.; Liaw, S.; Yang, F.; Huang, S. Generation of Augmented Capillary Network Optical Coherence Tomography Image Data of Human Skin for Deep Learning and Capillary Segmentation. *Diagnostics* **2021**, *11*, 685. [CrossRef] [PubMed]
2. Bekmirzaev, S.; Oh, S.; Yo, S. RethNet: Object-by-object learning for detecting facial skin problems. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
3. Yoon, H.; Kim, S.; Lee, J.; Yoo, S. Deep-Learning-Based Morphological Feature Segmentation for Facial Skin Image Analysis. *Diagnostics* **2023**, *13*, 1894. [CrossRef] [PubMed]
4. Lee, J.; Yoon, H.; Kim, S.; Lee, C.; Lee, J.; Yoo, S. Deep learning-based skin care product recommendation: A focus on cosmetic ingredient analysis and facial skin conditions. *J. Cosmet. Dermatol.* **2024**. [CrossRef] [PubMed]
5. Budhi, G.; Adipranata, R.; Gunawan, A. Acne segmentation and classification using region growing and self-organizing map. In Proceedings of the 2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT), Denpasar, Indonesia, 26–29 September 2017. [CrossRef]
6. Alamdari, N.; Tavakolian, K.; Alhashim, M.; Fazel-Rezai, R. Detection and classification of acne lesions in acne patients: A mobile application. In Proceedings of the 2016 IEEE International Conference on Electro Information Technology (EIT), Grand Forks, ND, USA, 19–21 May 2016. [CrossRef]
7. Yadav, N.; Alfayeed, S.; Khamparia, A.; Pandey, B.; Thanh, D.; Pande, S. HSV model-based segmentation driven facial acne detection using deep learning. *Expert Syst.* **2022**, *39*, e12760. [CrossRef]

8. Rashataprucksa, K.; Chuangchaichatchavarn, C.; Triukose, S.; Nitinawarat, S.; Pongprutthipan, M.; Piromsopa, K. Acne detection with deep neural networks. In Proceedings of the 2020 2nd International Conference on Image Processing and Machine Vision, Bangkok, Thailand, 5–7 August 2020; pp. 53–56. [CrossRef]
9. Huynh, Q.; Nguyen, P.; Le, H.; Ngo, L.; Trinh, N.; Tran, M.; Nguyen, H.; Vu, N.; Nguyen, A.; Suda, K.; Tsuji, K.; et al. Automatic Acne Object Detection and Acne Severity Grading Using Smartphone Images and Artificial Intelligence. *Diagnostics* **2022**, *12*, 1879. [CrossRef] [PubMed]
10. 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.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
11. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
12. Min, K.; Lee, G.; Lee, S. ACNet: Mask-aware attention with dynamic context enhancement for robust acne detection. In Proceedings of the 2021 IEEE International Conference On Systems, Man, And Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 2724–2729.
13. Junayed, M.; Islam, M.; Anjum, N. A Transformer-Based Versatile Network for Acne Vulgaris Segmentation. In Proceedings of the 2022 Innovations In Intelligent Systems And Applications Conference (ASYU), Antalya, Turkey, 7–9 September 2022; pp. 1–6.
14. Kim, S.; Lee, C.; Jung, G.; Yoon, H.; Lee, J.; Yoo, S. Facial Acne Segmentation based on Deep Learning with Center Point Loss. In Proceedings of the 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, 22–24 June 2023; pp. 678–683.
15. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
16. Yun, S.; Han, D.; Oh, S.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
17. Bai, Y.; Chen, D.; Li, Q.; Shen, W.; Wang, Y. Bidirectional copy-paste for semi-supervised medical image segmentation. In Proceedings of the IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 11514–11524.
18. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. *arXiv* **2019**, arXiv:1912.04958.
19. Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; Aila, T. Training Generative Adversarial Networks with Limited Data. *arXiv* **2020**, arXiv:2006.06676.
20. Sankar, A.; Chaturvedi, K.; Nayan, A.; Hesamian, M.; Braytee, A.; Prasad, M. Utilizing Generative Adversarial Networks for Acne Dataset Generation in Dermatology. *BioMedInformatics* **2024**, *4*, 1059–1070. [CrossRef]
21. Kim, S.; Lee, J.; Lee, C.; Lee, J. Improving Facial Acne Segmentation through Semi-Supervised Learning with Synthetic Images. *J. Korea Multimed. Soc.***2024**, *27*, 241–249. [CrossRef]
22. Wu, X.; Wen, N.; Liang, J.; Lai, Y.K.; She, D.; Cheng, M.M.; Yang, J. Joint acne image grading and counting via label distribution learning. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
23. Kim, S.; Yoon, H.; Lee, J.; Yoo, S. Semi-automatic labeling and training strategy for deep learning-based facial wrinkle detection. In Proceedings of the 2022 IEEE 35th International Symposium On Computer-Based Medical Systems (CBMS), Shenzen, China, 21–23 July 2022; pp. 383–388.
24. Kim, S.; Yoon, H.; Lee, J.; Yoo, S. Facial wrinkle segmentation using weighted deep supervision and semi-automatic labeling. *Artif. Intell. Med.* **2023**, *145*, 102679. [CrossRef] [PubMed]
25. Kang, S.; Lozada, V.; Bettoli, V.; Tan, J.; Rueda, M.; Layton, A.; Petit, L.; Dréno, B. New Atrophic Acne Scar Classification: Reliability of Assessments Based on Size, Shape, and Number. *J. Drugs Dermatol. JDD* **2016**, *15*, 693–702. [PubMed]
26. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E.; Kurakin, A.; Li, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
27. Chen, Y.; Tan, X.; Zhao, B.; Chen, Z.; Song, R.; Liang, J.; Lu, X. Boosting semi-supervised learning by exploiting all unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7548–7557.
28. Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; Cai, J. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2022; pp. 34–43.
29. Yang, L.; Qi, L.; Feng, L.; Zhang, W.; Shi, Y. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7236–7246.
30. Lumini. Lumini KIOSK V2 Home Page. 2020. Available online: https://www.lulu-lab.com/bbs/write.php?bo_table=product_en&sca=LUMINI+KIOSK (accessed on 9 July 2020).

diagnostics

MDPI

Article

IMPA-Net: Interpretable Multi-Part Attention Network for Trustworthy Brain Tumor Classification from MRI

Yuting Xie [1,2], Fulvio Zaccagna [3,4], Leonardo Rundo [5], Claudia Testa [6,7], Ruifeng Zhu [8], Caterina Tonon [1,2], Raffaele Lodi [1,2] and David Neil Manners [2,9,*]

1 Department of Biomedical and Neuromotor Sciences, University of Bologna, 40126 Bologna, Italy; yuting.xie2@unibo.it (Y.X.); caterina.tonon@unibo.it (C.T.); raffaele.lodi@unibo.it (R.L.)
2 Functional and Molecular Neuroimaging Unit, IRCCS Istituto delle Scienze Neurologiche di Bologna, Bellaria Hospital, 40139 Bologna, Italy
3 Department of Imaging, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0SL, UK; fz247@cam.ac.uk
4 Department of Radiology, University of Cambridge, Cambridge CB2 0QQ, UK
5 Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy; lrundo@unisa.it
6 INFN Bologna Division, Viale C. Berti Pichat, 6/2, 40127 Bologna, Italy
7 Department of Physics and Astronomy, University of Bologna, 40127 Bologna, Italy
8 Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, 41125 Modena, Italy; reefing.z@gmail.com
9 Department for Life Quality Studies, University of Bologna, 40126 Bologna, Italy
* Correspondence: davidneil.manners@unibo.it

Abstract: Deep learning (DL) networks have shown attractive performance in medical image processing tasks such as brain tumor classification. However, they are often criticized as mysterious "black boxes". The opaqueness of the model and the reasoning process make it difficult for health workers to decide whether to trust the prediction outcomes. In this study, we develop an interpretable multi-part attention network (IMPA-Net) for brain tumor classification to enhance the interpretability and trustworthiness of classification outcomes. The proposed model not only predicts the tumor grade but also provides a global explanation for the model interpretability and a local explanation as justification for the proffered prediction. Global explanation is represented as a group of feature patterns that the model learns to distinguish high-grade glioma (HGG) and low-grade glioma (LGG) classes. Local explanation interprets the reasoning process of an individual prediction by calculating the similarity between the prototypical parts of the image and a group of pre-learned task-related features. Experiments conducted on the BraTS2017 dataset demonstrate that IMPA-Net is a verifiable model for the classification task. A percentage of 86% of feature patterns were assessed by two radiologists to be valid for representing task-relevant medical features. The model shows a classification accuracy of 92.12%, of which 81.17% were evaluated as trustworthy based on local explanations. Our interpretable model is a trustworthy model that can be used for decision aids for glioma classification. Compared with black-box CNNs, it allows health workers and patients to understand the reasoning process and trust the prediction outcomes.

Keywords: decision support; interpretability; trustworthiness; deep neural networks; brain tumor classification; multi-part attention

Citation: Xie, Y.; Zaccagna, F.; Rundo, L.; Testa, C.; Zhu, R.; Tonon, C.; Lodi, R.; Manners, D.N. IMPA-Net: Interpretable Multi-Part Attention Network for Trustworthy Brain Tumor Classification from MRI. *Diagnostics* **2024**, *14*, 997. https://doi.org/10.3390/diagnostics14100997

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 18 April 2024
Revised: 8 May 2024
Accepted: 9 May 2024
Published: 11 May 2024

1. Introduction

Brain cancer is one of the ten leading causes of death globally among men and women [1,2]. The World Health Organization estimates the 5-year survival rate is only 21% for people aged 40 and over [2]. In most clinical scenarios, LGGs are well-differentiated, slow-growing lesions, while HGGs are usually aggressive with dismal prognosis [3,4]. Survival rates differ markedly for different tumor grades. Identifying tumor grade at an

early stage is a major unmet need; it contributes to formulating better treatment strategies and enhances the overall quality of life of patients.

Magnetic resonance (MR) imaging is a non-invasive technique that remains the standard of care for brain tumor diagnosis and treatment planning in clinical practice [5,6]. It provides a reasonably good delineation of the gliomas and conveys biological information on the tumor location, size, necrosis, edema tissue, the mass effect, and breakdown of the blood–brain barrier (which results in contrast enhancement in post-contrast-enhanced T_1-weighted (ceT_1w) MR images) [6]. In general, LGGs are less invasive. They usually have well-defined boundaries and homogeneous tumor cores without prominent mitosis, necrosis, and microvascular proliferation [6–9]. HGGs always show more mass effect. They usually show microscopic peritumoral white matter tract invasion. The demonstration of this diffuse infiltration is an important discriminating feature for the accurate glioma diagnosis [6].

Diagnosis of brain tumors from MR images is a time-consuming and challenging task that requires professional knowledge and careful observation. As alternatives, various automated diagnosis approaches have been developed to assist radiologists in the interpretation of the brain MR images and reduce the likelihood of misdiagnosis. Convolutional neural networks (CNNs) provide a powerful technology for medical data analysis [10]. CNN-based deep learning architectures can extract important low-level and high-level features automatically from the given training dataset of sufficient variety and quality [11]; they embed the phase of feature extraction and classification into a self-learning procedure, allowing fully automatic classification without human interaction, which can be applied to the problem of tumor diagnosis.

Over the last decade, methods using CNNs have been extensively investigated for brain tumor classification due to their outstanding performance with very high accuracy in a research context [12,13]. The differential classification of HGG and LGG is a comparatively simple task that has been tackled in numerous different ways using different CNN methods, and the best-performing models have demonstrated close to 100% performance [10]. For example, Khazaee et al. [14] used a pre-trained EfficientNetB0 for HGG and LGG classification. The model achieved a mean classification accuracy of 98.87%. Chikhalikar et al. [15] proposed a custom CNN model to classify the type of tumor present in MRI images, achieving an accuracy of 99.46%. The authors in [16] used transfer learning with stacking InceptionResNetV2, DenseNet121, MobileNet, Incep-tionV3, Xception, VGG16, and VGG19 for the same classification task. The average classification accuracy for the test dataset reached 98.06%. Zhuge et al. [17] utilized a pre-trained ResNet50. The classification accuracy of the proposed model reached 96.3%.

The above CNN-based methods all achieved remarkable performance on automated HGG and LGG classification. However, MR images are unlikely to be artifact-free [18], and the lesion signal measured by MRI is typically mixed with nuisance sources. The above-mentioned black-box CNNs may learn confounding sources from MR images for decision making, and the health outcomes cannot easily gain the trust of physicians or patients because the evidence is unknown [6,19].

The lack of transparency and interpretability concerning the decision-making process still limits their development into clinical practice [12,19,20]. Visualizing the features that are faithful to the underlying lesion is crucial to ensuring the interpretability and trustworthiness of classification outcomes. Interpretability is the ability to provide explanations in terms understandable to a human [21], based on their domain knowledge related to the task, or common knowledge, according to the task characteristics. The need for interpretability has already been stressed by many papers [21–23], emphasizing cases where lack of interpretability may be harmful. Can we explain why algorithms go wrong? When things go well, do we know why and how to exploit them further?

In order to deploy a system in practice, it is necessary to present classification results in such a way that they are acceptable to end users. This is only possible if users trust the decision-making process, which, as a consequence, must be transparent and interpretable.

To date, a limited number of saliency-based interpretable methods have suggested different frameworks to improve the interpretability and trustworthiness of CNNs for brain tumor classifications [24–27]. We divide the previous interpretable approaches into two categories: object-level methods and pixel-level/part-level methods.

At the coarsest level, there are models that have been proposed to offer object-level explanations for brain tumor classification tasks, such as a class activation mapping method GradCAM [24,25] that highlights that entire object as the explanation behind the tumor predictions. The authors in [25] proposed a pre-trained ResNet-50 CNN architecture to classify three posterior fossa tumors and explained the classification decision by using GradCAM. The heatmap generated by the GradCAM technique can identify the area of emphasis and help visualize where the classification model looks for individual predictions.

At a finer level, there are a few interpretable techniques that have been applied to explain the brain tumor classification results with pixel-level/part-level explanations, such as pixel-level interpretable algorithms SHAP, Guided Backpropagation (GBP) [24], and a part-level interpretable model called LIME. Authors in [27] explained the tumor predictions made by the CNN model with SHAP and LIME methods. The SHAP algorithm explains the individual prediction by computing the contribution of each pixel on a predicted image to the prediction using Shapley values to understand what are the main pixels that affect the output of the model [28]. The LIME algorithm is a counterfactual explanation method that approximates the classification behavior of a complex neural network using a simpler, more understandable model without exploring the model itself [29]. In the study, the authors segmented the input image into superpixels and made small disturbances around each superpixel to figure out the contribution/importance of each superpixel to the prediction result. Another study conducted by Pereira et al. [24] utilized GradCAM and GBP maps to provide insights into the regions that support the prediction to perform quality assessment of tumor grade prediction between HGG and LGG. The GBP is a gradient-based visualization method that can visualize which pixels in the input image are more informative for the correct classification.

The above methods identify the most important pixels or objects of an image as the explanation for the prediction outcomes. To some extent, they verify the validity of the classification models. Nevertheless, it is worth stressing that knowing the most important pixels or objects of an image that determined a specific prediction does not always amount to a good-quality explanation.

Ideally, networks should be able to explain the reasoning process behind each individual decision, and this process, ideally, would be similar to that used by a radiologist, who looks at specific features of the MR image relevant to the task. For example, if a doctor classifies a tumor as HGG, this decision always relies mainly on the high-level class-representative features or properties, like the tumor's irregularity, the necrotic area, or the enhancing ring [30].

The objectives of this study were to build an interpretable multi-part attention [31] network (IMPA-Net) for brain tumor classification to unbox the model and the reasoning process of individual predictions with understandable MR imaging features. The proposed IMPA-Net, motivated by [32], provides both global and local explanations for brain tumor classification on MRI images. Figure 1 gives a more detailed illustration of the connections and distinctions between the two explanations. The global explanation is represented by a group of feature patterns that the model learns and uses for the classification. The quality of the feature patterns can be used to evaluate the ability and reliability of the model on the classification task. The local explanation interprets the reasoning process of an individual prediction by comparing the prototypical parts of the image with feature patterns. It can be used to evaluate the trustworthiness of individual predictions.

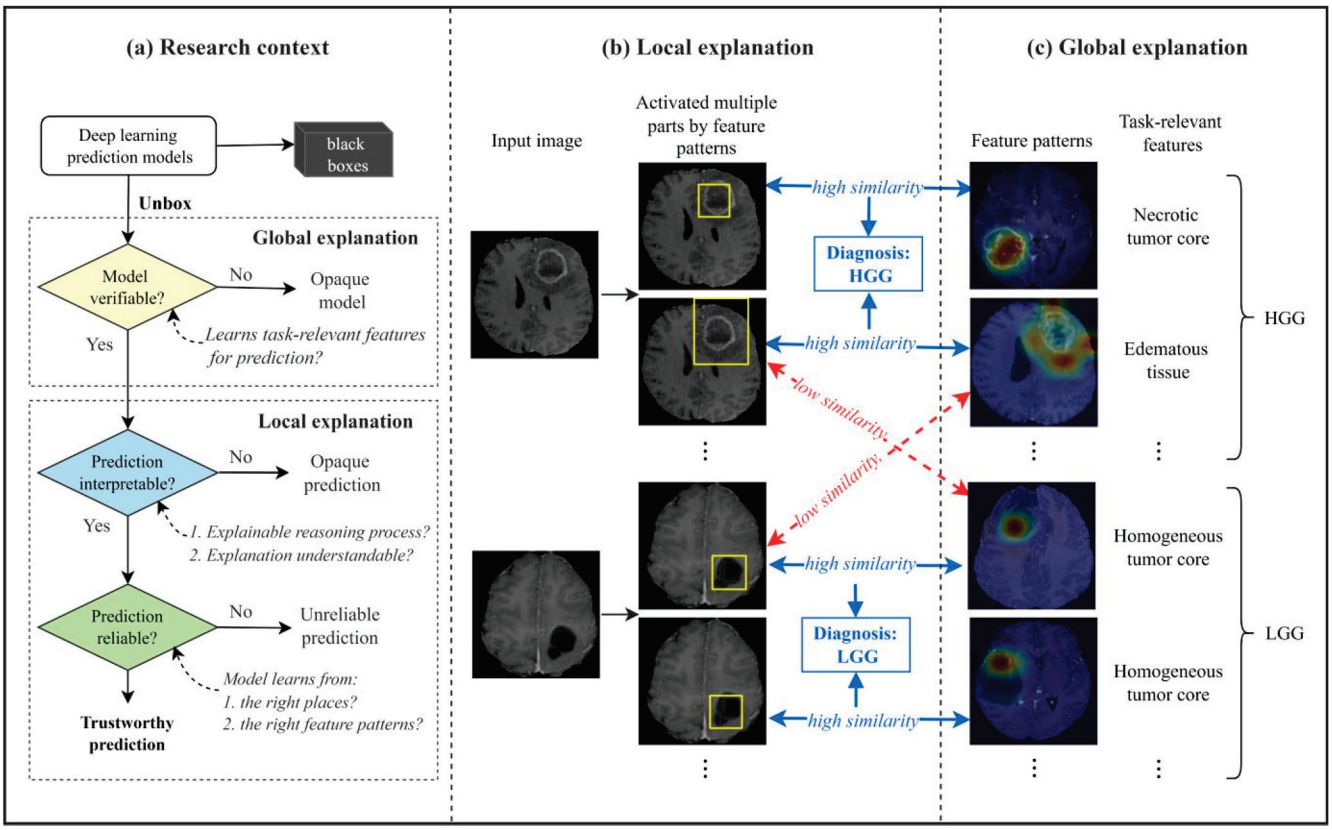


Figure 1. Global and local explanations provided by the proposed IMPA-Net. (**a**) Research context illustrates the importance and basic ideas of global and local explanations for deep learning-based brain tumor classification. It outlines the problems in this research field that the proposed IMPA-Net attempts to address; (**b**) local explanation: given an input image, IMPA-Net compares the activated parts of the input image with the feature patterns and thereby predicts the tumor grade; (**c**) global explanation can be interpreted as the class-representative features the entire model learns to distinguish two classes.

The main contribution of this paper is that it addresses the black-box problems of CNN classification models for glioma diagnosis by developing a model with the following characteristics:

(i) The first multi-part interpretable model that can provide both global and local explanations for brain tumor classification, enabling better human–machine collaboration for decision aid.
(ii) It presents the reasoning process of individual predictions to show how the model arrives at the decision making in this context, allowing health workers to evaluate the reliability of the prediction outcomes.
(iii) It allows the prediction results to be interpreted in a clinical context.
(iv) It highlights the most relevant information for predictions based on medical disease-related features that can be understood and interpreted by clinicians and patients.

The remainder of the paper is structured as follows. Section 3 gives a detailed introduction to the dataset, the proposed interpretable multi-part attention network, and the experimental setup. Results are given in Section 3. Section 4 evaluates the performance of the proposed method on both aspects of its classification and explanation. Section 5 concludes the key findings of this study. Section 6 concludes the proposed work and discusses the future research directions.

2. Materials and Methods

The overall workflow of the development and evaluation of the proposed methodology is shown in Figure 2. Input brain MRI images are firstly pre-processed by resizing, normalization, and cropping, and then three augmentation methods, including rotation, shearing, and skewing are performed to produce the training dataset. The proposed methodology classifies the input image by comparing its prototypical patches with pre-learned feature patterns of classes HGG and LGG. In this stage, feature patterns of both classes are optimized and produced. The quality of the feature patterns is evaluated in the next step on aspects of their interpretability, class representability, and correctness, and then poor-quality feature patterns are excluded in the local explanation process. In the next stage, local explanations of individual predictions are given to illustrate how the model arrives at the final decisions, and each case will be evaluated based on whether it satisfies two basic conditions identified for reliability assessment. Finally, the proposed model is evaluated on both aspects of its performance (classification and explanation), including classifier performance, global explanation evaluation, local explanation evaluation (correctness and confidence), and user evaluation.

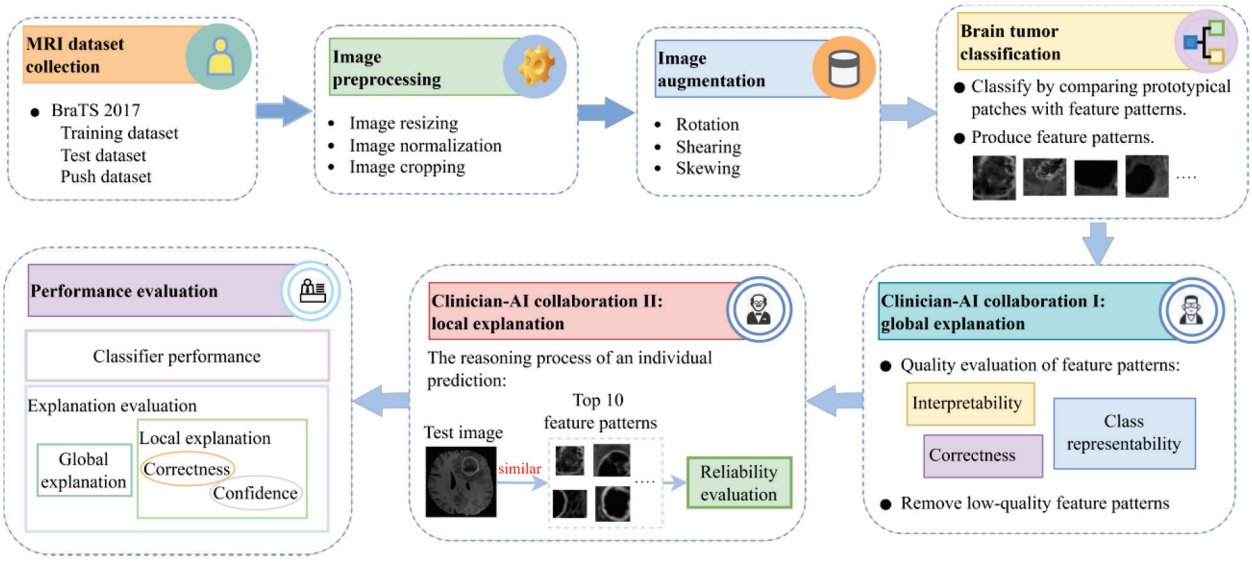


Figure 2. The overall workflow of the development and evaluation of the proposed methodology.

2.1. Data and Image Processing

We trained and evaluated our network on data from the BraTS 2017 database [33–35]. The dataset contains 285 routine-acquired 3T multimodal clinical MRI scans from multiple institutions, comprising 210 patients with pathologically confirmed HGG and 75 patients with LGG. All images from the dataset were pre-processed by co-registration to the same anatomical template, interpolation to the same resolution (1 mm^3), and skull stripping [33].

Slices that contain gliomas were extracted from each patient's MRI scan. Considering the enhancing ring in post-contrast-enhanced T_1-weighted (ceT_1w) MR is an important discriminating feature for accurate tumor diagnosis between HGG and LGG [6], in our experiments, only ceT_1w MR images were considered. The dataset was then partitioned into a training dataset (70%) and a testing dataset (30%). A push dataset of 60 images was randomly selected from the training dataset (30 images for each class).

All images were normalized by Z-score normalization and converted to PNG format, and then the background pixels were cropped to focus feature learning on the brain areas instead of the whole image. Moreover, the images were resized to 224 × 224 to fit the model's training configurations.

2.2. Data Augmentation

To increase the size and variability of the training dataset, data augmentation methods were performed, including twice rotating in the axial imaging plane by a random amount between 20° left and 20° right, shearing by a random amount between 10° left and right twice in the transverse direction, and skewing by tilting the images left/right by a random amount (magnitude = 0.2) twice. In this way, the training dataset is augmented six-fold, resulting in 6228 images (3546 HGG, 2682 LGG).

2.3. Interpretable Convolutional Neural Network

Figure 3 gives an overview of the proposed IMPA-Net, which consists of a feature extractor, multi-part attention (MPA), and similarity-based classifier. Images are first propagated into convolutional layers for feature extraction, with a structure selected from VGG16. In the proposed classification model, we chose VGG16 as the feature extractor as it combines simplicity, ease of implementation, and fine-tuning capability with adequate feature extraction effectiveness and generalization ability. The pre-trained VGG16 model is suitable for transfer learning or fine-tuning as a feature extractor for brain tumor classification tasks [12]. A non-linear activation function ReLU is used for all convolutional layers. Then, these convolutional layers are followed by a multi-part attention module for similarity calculation between CNN outputs and the feature patterns pre-learned by the model. In particular, our network tries to find evidence for an image (such as the pre-processed HGG image in Figure 3) to be of class HGG by comparing its prototypical patches with learned feature patterns of class HGG and LGG, as illustrated in the similarity correlation units. This comparison produces a map of similarity scores of each feature pattern, which is upsampled and superimposed on the input image to see which part of the input image is activated by each feature pattern. The activation maps are then propagated into a max-pooling layer, producing a single similarity score for each comparison. Finally, the model classifies the input image based on the top 10 similarity scores. The output $S_{c_{HGG}}$ denotes the weighted sum of top-10 similarity scores generated by the multi-part attention module.

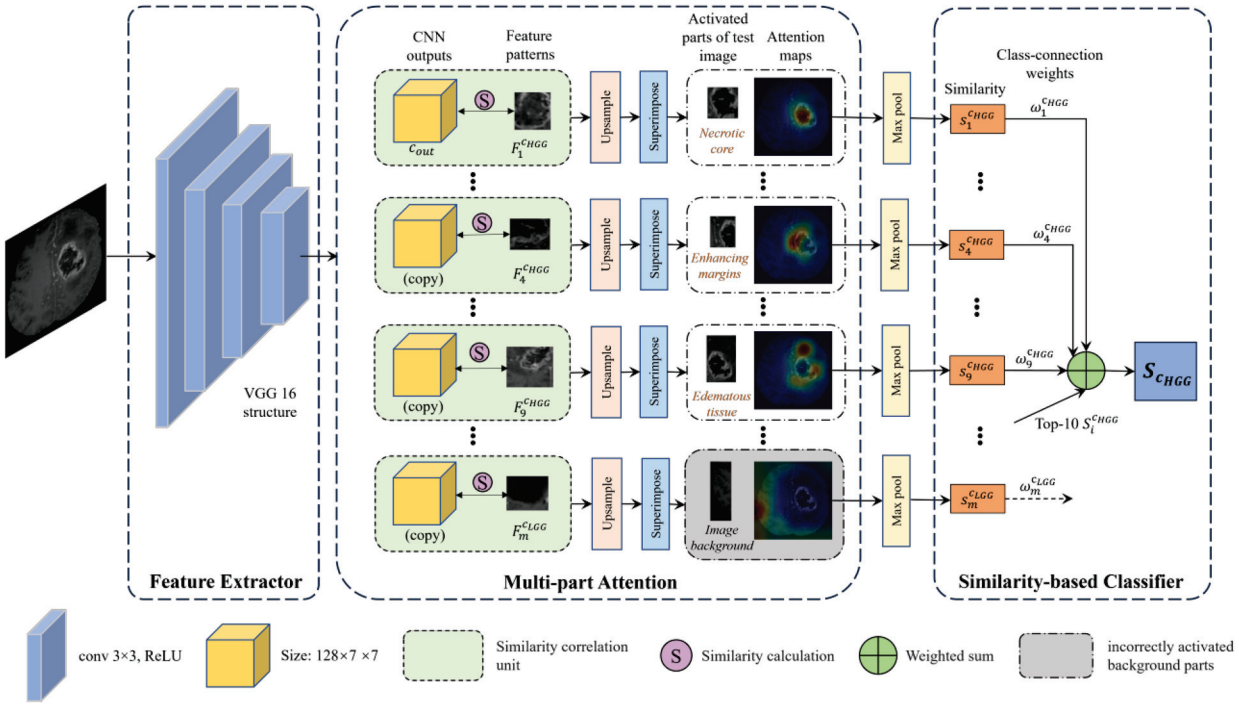


Figure 3. Schematic diagram of the proposed IMPA-Net. It consists of three modules: a feature extractor, a multi-part attention block, and a similarity-based classifier. The feature patterns within the multi-part attention block are learned from the push dataset during the training phase.

2.3.1. Feature Extractor

The architecture consists of a regular convolutional neural network for feature extraction with a structure selected from VGG16 (kernel size 3×3), followed by two additional 1×1 convolutional layers. All these convolutional layers (f) use a ReLU with a non-linear activation function.

For a given pre-processed input image x (such as the HGG sample image in Figure 3), the convolutional layers f extract useful features from x to use for prediction, whose output $c_{out} = f(x)$ have spatial dimension $D \times 7 \times 7$, where D is the number of the output channels of the last convolutional layer.

2.3.2. Multi-Part Attention

In our experiments, we allocated a pre-determined number of feature patterns $FP = \left\{ fp_i^{c_j} \right\}_{i=1}^{m}$, $m = 50$ for each class, where c_j ($j \in \{\text{HGG}, \text{LGG}\}$) represents the class identity of the feature pattern and i is the index of that feature pattern among all feature patterns of class c_j. So that for each class, 50 feature patterns are learned and produced by the model from a push dataset. This dataset consists of a pre-determined number of MRI images that are randomly selected from the training dataset. The shape of each pattern is $D \times h \times w$, where $h \times w < 7 \times 7$. In our experiments, h and w are set to 1. The depth of each feature pattern is the same as that of c_{out} but the height and width are smaller than those of the c_{out}, each feature pattern will be supposed to represent some representative activation pattern in a patch of the convolutional output c_{out}, which in turn will correspond to some prototypical image patch in the original training image.

In our network, every feature patch can be considered as a representative pattern of one image from the push dataset, and these feature patterns are supposed to direct attention to enough medical semantic content for recognizing a class [36]. As a schematic illustration of the multi-part attention for the HGG sample image in Figure 3, the first feature pattern $fp_1^{c_{HGG}}$ corresponds to the necrotic tumor core of an HGG training image, and the fourth feature pattern $fp_4^{c_{HGG}}$ enhancing tumor margin of an HGG training image, and the ninth feature pattern $fp_9^{c_{HGG}}$ the edematous area of an HGG image.

The similarity correlation units SCU in a multi-part attention module computes the L2 distance between the CNN outputs and the feature patterns, as shown in Equation (1). The i_{th} similarity correlation unit $SCU_i^{c_j}$ of class c_j calculates the squared Euclidean distances between feature patterns $fp_i^{c_j}$ and each patch $\tilde{c_{out}}$ generated from the convolutional outputs $\boldsymbol{c_{out}}$ and then inverts the distances to similarity scores. Mathematically, the similarity correlation unit $SCU_i^{c_j}$ calculates the following:

$$dist\left(\tilde{c_{out}},\ fp_i^{c_j}\right) = \left\| \tilde{c_{out}}, fp_i^{c_j} \right\|_2,\ \ \tilde{c_{out}} \in patches(\boldsymbol{c_{out}}), \tag{1}$$

$$sim\left(\tilde{c_{out}},\ fp_i^{c_j}\right) = \log\left(\frac{dist\left(\tilde{c_{out}},\ fp_i^{c_j}\right)^2 + 1}{dist\left(\tilde{c_{out}},\ fp_i^{c_j}\right)^2 + \varepsilon}\right), \tag{2}$$

$$SCU_i^{c_j}(c_{out}) = \max_{\tilde{c_{out}} \in \text{ patches } (c_{out})} sim\left(\tilde{c_{out}},\ fp_i^{c_j}\right), \tag{3}$$

These similarity scores calculated by Equation (2) define an activation map, which retains the spatial relation of the convolutional output c_{out}. The activation map can be unsampled to the size of the input image to visualize the part of the input image that looks most similar to the feature pattern [36]. In Figure 3, the similarity score between the first feature patterns $fp_1^{c_{HGG}}$, a an HGG necrotic tumor core, and the most activated patch of the input image of a an HGG is $s_1^{c_{HGG}}$. The similarity score between the fourth feature pattern $fp_4^{c_{HGG}}$, an HGG enhancing tumor margin, and the most activated patch of the input image is $s_4^{c_{HGG}}$. The third feature pattern $fp_9^{c_{HGG}}$, an HGG edematous area, activated mostly on the

edematous tissue of the HGG sample image, with a similarity score of $s_9^{c_{HGG}}$. This shows that our model finds that the necrotic tumor core of the HGG sample image has a stronger presence than that of enhancing tumor margin in the input image.

Equation (2) indicates that the similarity is monotonically decreasing with respect to the squared Euclidean distance, that is, the highest similarity score of the similarity correlation unit $SCU_i^{c_j}$ comes when $\tilde{c_{out}}$ is the closest patch to $fp_i^{c_j}$. In activation maps, warmer values indicate higher similarity between the learned feature patterns and the parts of the input image activated by the feature pattern, which is enclosed in the yellow rectangles on the superimposed source images. Then, the activation maps produced by similarity scores are max pooled to reduce to a single similarity score $s_i^{c_j}$ for each feature pattern $fp_i^{c_j}$. Hence, if the similarity score of the i_{th} similarity correlation unit $SCU_i^{c_j}$ is high, it indicates that there is a patch in the input image that is very similar to the i_{th} feature pattern of class c_j in the latent space, and that the activated patch contains a similar pattern to that represented in the i_{th} feature pattern.

2.3.3. Similarity-Based Classifier

Finally, in the classifier block, the top 10 ranking similarity scores are multiplied by the class-connection weight matrix $\omega_i^{c_j}$ to produce the output logit to class c_j. The matrix $\omega_i^{c_j}$ represents the relationship between feature patterns and the logit of the class. Higher class-connection values refer to higher representability of the feature pattern to its class.

$$S_{c_j} = \sum_{i=1}^{10} \omega_i^{c_j} \cdot s_i^{c_j}, j \in \{\text{HGG}, \text{LGG}\} \tag{4}$$

2.4. Model Training

The training of the proposed model is divided into three stages: stochastic gradient descent (SGD) of layers before the classifier layer, projection and optimization of feature patterns, and optimization of class-connection weights.

2.4.1. Stochastic Gradient Descent (SGD) of Layers before the Classifier Layer

The architecture aims to learn meaningful and teak-relevant features that can be used to distinguish between HGG and LGG, where the most important patches for the classification task are clustered (in Euclidean distance) around similar feature patterns of the 'correct' class and separated from feature patterns from a different class [36]. To learn these features, an iterative algorithm SGD is used to simultaneously optimize the parameters of the convolutional layers f (f_{conv}) in the feature extractor and the feature pattern $FP = \left\{ fp_i^{c_j} \right\}_{i=1}^{m}$ in the multi-part attention module via back propagation. In this step, the weight matrix (class connection values) $\omega_i^{c_j}$ of the last layer in the classifier block is frozen.

Formally, let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of training images, $Y = \{y_1, y_2, \ldots, y_n\}$ be the set of the corresponding labels. The optimization problem to be solved here is to minimize the defined loss function that incorporates the cross-entropy loss (CELoss), cluster loss (ClstLoss), and separation loss (SepLoss):

$$Loss = \frac{1}{n} \sum_{k=1}^{n} \text{CELoss}(f \circ SCU \circ f(x_k), y_k) + r_1 \text{ClstLoss} + r_2 \text{SepLoss} \tag{5}$$

where ClstLoss and SepLoss are

$$\text{ClstLoss} = \frac{1}{n} \sum_{k=1}^{n} \underset{i : fp_i^{c_j} \in FP^{y_k}}{\operatorname{argmin}} \underset{\tilde{f(x_k)} \in \text{patches}(f(x_k))}{\operatorname{argmin}} \left\| \tilde{f(x_k)} - fp_i^{c_j} \right\|_2^2 \tag{6}$$

$$\text{SepLoss} = -\frac{1}{n} \sum_{k=1}^{n} \underset{i : fp_i^{c_j} \notin FP^{y_k}}{\operatorname{argmin}} \underset{\tilde{f(x_k)} \in \text{patches}(f(x_k))}{\operatorname{argmin}} \left\| \tilde{f(x_k)} - fp_i^{c_j} \right\|_2^2, \tag{7}$$

The CELoss penalizes misclassification during the training process, and the aim is to minimize CELoss to give better classifications. The ClstLoss is minimized to encourage the prototypical parts to cluster around the correct class, see Equation (6), whereas the SepLoss is minimized to separate the prototypical parts from the incorrect class; see Equation (7).

2.4.2. Projection of Feature Patterns

To visualize which parts of the training images from the push dataset are used as feature patterns, the network projects every feature pattern $fp_i^{c_j}$ onto the closest patch of the output $f({x_k}^{c_j})$ that has the smallest distance from $fp_i^{c_j}$, and the closest patch has the same class c_j as that of $fp_i^{c_j}$ [32]. The reason is that the patch of training image ${x_k}^{c_j}$ that corresponds to $fp_i^{c_j}$ should be the one that $fp_i^{c_j}$ activates most strongly on. We can visualize the part of ${x_k}^{c_j}$ on which $fp_i^{c_j}$ has the strongest activation by forwarding ${x_k}^{c_j}$ through a trained network. Mathematically, for feature pattern $fp_i^{c_j}$ of class c_j ($j \in \{\text{HGG}, \text{LGG}\}$), the network performs the following update:

$$fp_i^{c_j} = \underset{patch, patch \in patches(f(x^{c_j}))}{\operatorname{argmin}} \| patch - fp_i^{c_j} \|_2, y_k = c_j \tag{8}$$

2.4.3. Optimization of Class-Connection Weights

In this stage, all the parameters from the convolutional layers and multi-part attention blocks are frozen, and a convex optimization on the class-connection weight matrix $\omega_i^{c_j}$ of the last layer is performed. To rely only on positive connections between feature patterns and logits, the negative connection $\omega_i^{c_j}$ is set to 0 for all to reduce the reliance of the model on a negative reasoning process of the form "this image is of class HGG because it is not of class LGG.". Mathematically, we perform this step to optimize

$$\min_{\omega_i^{c_j}} \frac{1}{n} \sum_{k=1}^{n} CELoss(f \circ SCU \circ f(x_k), y_k) + \lambda \sum_{c_j : fp_i^{c_j} \notin FP^{y_k}} \left| \omega_i^{(k,c_j)} \right|, \tag{9}$$

2.5. Experimental Setup

All the experiments were conducted on a PC with an Intel Core i7-6700K 4.00 GHz processor running Ubuntu 18.04.6 with one NVIDIA GeForce RTX 2060, using Python 3.9.7 and PyTorch 1.10.1.

The parameters of the convolutional layers from the VGG16 model were pre-trained on ImageNet [37], and the parameters of the additional convolutional layers were initialized with Kaiming uniform methods [38]. The parameters of the two additional convolutional layers are trained and optimized with the learning rate 3×10^{-3} for 5 epochs, while the pre-trained parameters and biases are fixed. In the following joint training stage, the parameters of all convolutional layers are optimized from epoch 6, and the model performs feature pattern projection every 20 epochs, that is, epochs 20, 40, 60, 80, and 100, and the convex optimization of the last layer is performed after each feature pattern projection process for 20 iterations with learning rate 10^{-4}.

The other hyperparameters are learning rate for layers pre-trained on ImageNet: 10^{-4} and learning rate for feature pattern optimization: 3×10^{-3}. For VGG16, we set $D = 128$ as the number of channels in a similarity correlation unit.

3. Results

3.1. Global Explanation

Global explanation can be interpreted as the class-representative features the entire model uses to distinguish two classes. Figure 4 shows six learned feature patterns and their activation maps for each class. It can be seen that all feature patterns localize important distinguishing features of both classes. The feature patterns of HGG that have higher

responses in contrast-enhancing tumors as a classification feature agrees with the actual imaging characteristics of HGG [6]; the feature patterns that focus on the necrotic tumor core that present heterogeneous high signal and the edematous areas are also important disease-representative features of HGG [6]; the feature patterns of LGG present higher responses on the homogeneous tumor cores and the non-enhancing tumor margins [7–9]. It is worth mentioning that those localized medical features can be understood and interpreted by the users, and thus, our framework can help provide global explanations in a human-understandable manner.

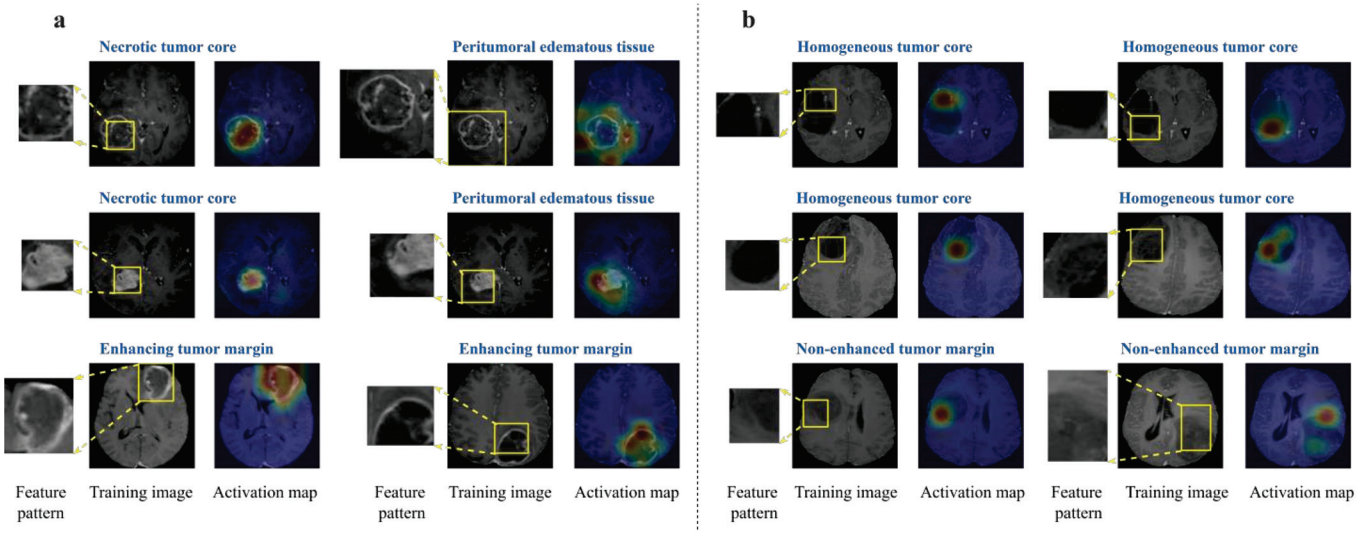


Figure 4. Six learned feature patterns and activation maps of HGG (**a**) and LGG (**b**) selected to represent different clinically relevant discriminative features of each class learned by the model. Training image where feature pattern comes from (feature pattern in box); Activation map (warmer colors indicate higher activation).

3.2. Local Explanation: Individual Predictions

The local explanation of individual predictions has to satisfy two conditions in order for its prediction explanations to be considered trustworthy and reliable; that is, all feature patterns that present the 10 highest similarity scores are from the class of the test image, and the concept of each top-10 feature pattern is consistent with that of the activated prototypical patch.

Figure 5 shows the reasoning process of our interpretable model in reaching a prediction on a test image of an HGG. As shown in the activation maps, the highest responses were found on the tumor core activated by the top and 2nd ranked feature patterns of class HGG (with similarity scores 8.143 and 8.105, respectively), the 3rd ranked feature pattern on the tumor enhancing margins, the 6th, 8th, and 9th ranked feature patterns on the edematous tissues.

The network correctly classifies the tumor as an HGG according to the ground truth. Furthermore, it provides the evidence of this prediction outcome with multi-part attention between patches of the test image and feature patterns as the tumor is classified as an HGG because prototypical patches of the test image, including its necrotic tumor core, enhancing margins, and edematous tissue was found to have higher similarity (top 10) with feature patterns from HGG class. The evidence is evaluated to be trustworthy according to the two reliability criteria, that is, all top-10 feature patterns are from the HGG class, and the concept of each top-10 feature pattern is consistent with that of the localized prototypical patch.

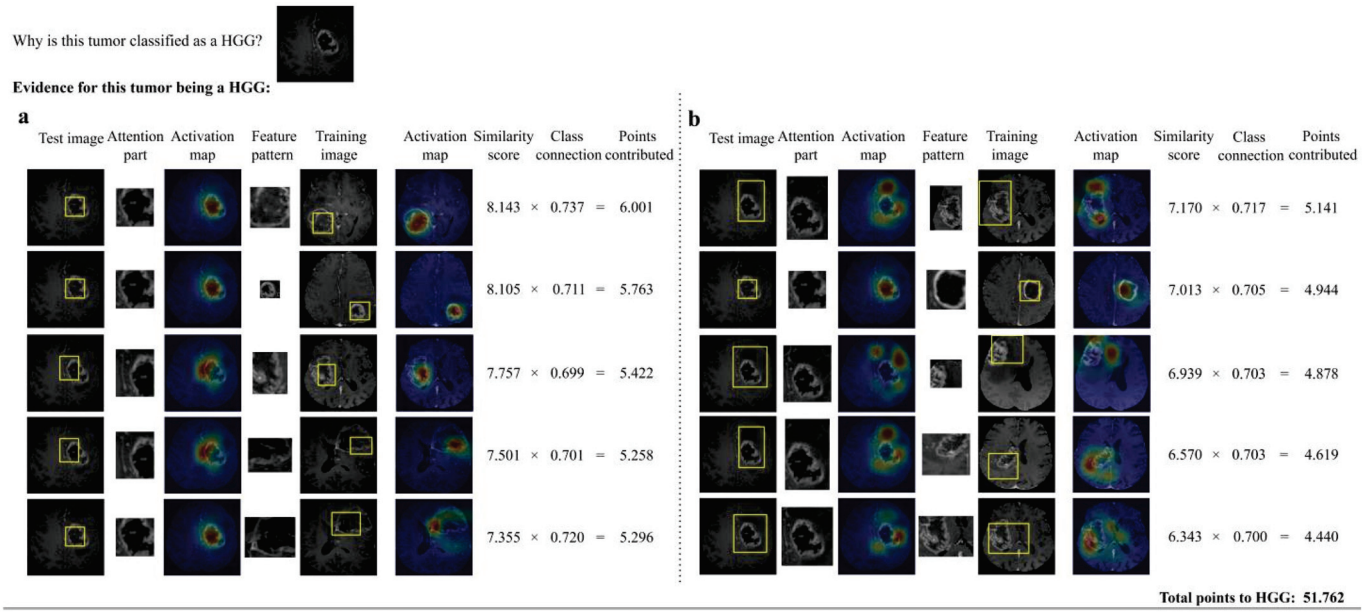


Figure 5. The reasoning process of our network for deciding the grade of a tumor. There are ten rows, split into two groups for ease of presentation: (**a**) top 1~#5th part attention between a patch of the test image and feature pattern, (**b**) #6th~#10th part attention between a patch of the test image and feature pattern. Each row is organized as follows: in the leftmost column a yellow rectangle generated by the proposed model is superimposed on the test image, showing a part that looks like a feature pattern; second column, an enlargement of the part of the test image considered by the model to look similar to the feature pattern (shown in col. 4); third column: activation maps indicating how similar each featured pattern resembles part of the test image, in which warmer color indicates higher responses; fifth column: training images where feature pattern comes from; sixth column: corresponding activation maps. The final columns quantify the result of the comparison. Column 7: similarity score between the localized prototypical part of the test image (col. 2) and the feature pattern (col. 4). Column 8: class connection values generated by the proposed model correspond to the class-connection weight connection between the feature patterns and the logit of class. Column 9: weighted similarity scores between the localized prototypical patches of the test image with top-10 feature patterns.

Figure 6 shows the reasoning process for reaching a classification decision on a test image of an LGG. As shown in the third column, the highest responses were found on the tumor core of the LGG image activated by two 'tumor core' feature patterns (similarity score of 7.420 and 7.332, respectively), the 3rd and 4th ranked feature patterns on the tumor margins. The network correctly classifies the tumor as an LGG. The explanation is the network classifies the tumor as an LGG because prototypical patches of the test image, including its homogeneous tumor core and non-enhancing tumor margins, were found to have higher similarity (top 10) with feature patterns from the LGG class. Those medical feature patterns can be understood and interpreted by the users, and thus, our framework can help provide global explanations in a human-understandable manner. The evidence for the prediction is evaluated to be trustworthy according to the two reliability criteria.

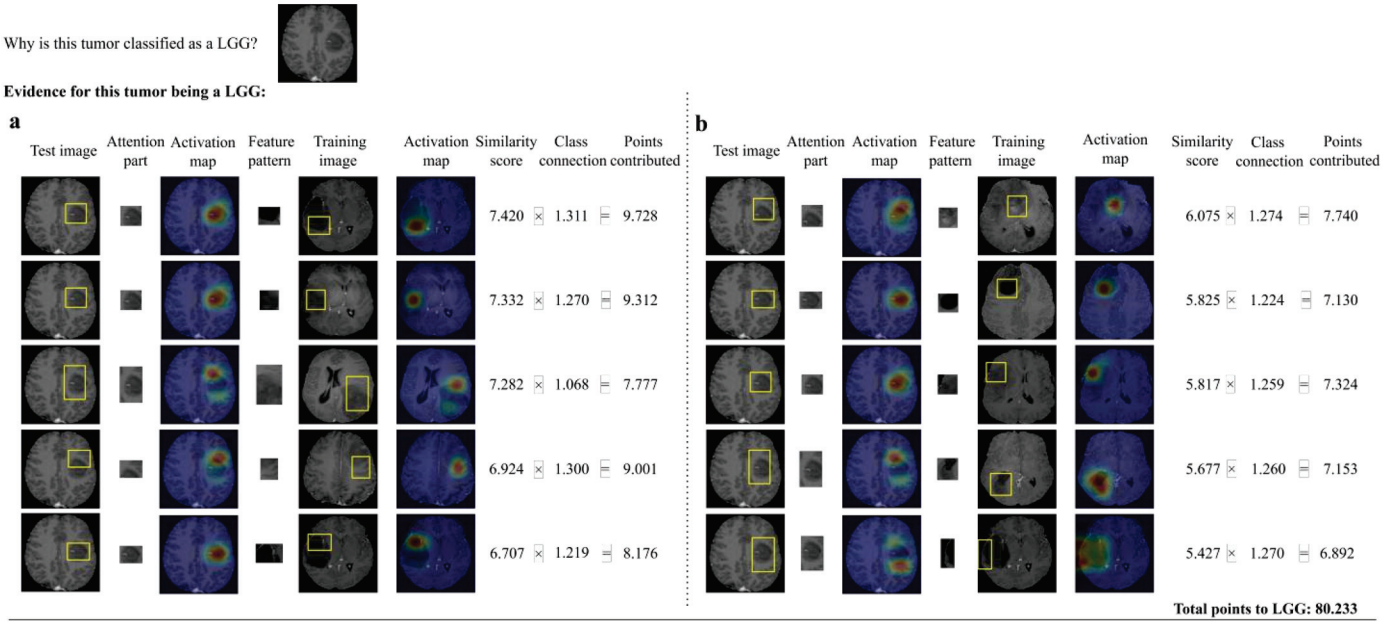


Figure 6. Example output showing the reasoning process of our network in deciding the grade of an LGG tumor, (**a**) top 1~#5th part attention between a patch of the test image and feature pattern, (**b**) #6th~#10th part attention between a patch of the test image and feature pattern.

4. Performance Evaluation

4.1. Classification Performance

Statistic metrics for classification performance, including accuracy (ACC), precision (PRE), specificity (SPE), sensitivity (SEN), and F_1-score, were calculated both for the interpretable decision-aid system described in this work and the baseline model, whose architecture consisted of the same convolutional layers without the intermediate multi-part attention module and similarity-based classifier. Correct predictions were further evaluated on their reliability based on local explanations to obtain reliable prediction accuracy to assess the trustworthiness of the model.

Table 1 presents the comparison of the classification performance of our interpretable model (before and after the exclusion of 'background' feature patterns) with the baseline model trained on the same dataset. Results show that the interpretable model is slightly less accurate than the baseline model and that the exclusion of the 'background' feature patterns improved the classification accuracy by 6.53%.

Table 1. Comparison of the classification performance of our interpretable model with the baseline model.

Model	Performance Metrics				
	ACC	**PRE**	**SPE**	**SEN**	**F_1 Score**
Baseline model	97.30%	99.18%	98.96%	96.03%	0.9758
Our model before exclusion	85.59%	89.17%	86.46%	84.92%	0.8699
Our model after exclusion	92.12%	94.65%	93.23%	91.27%	0.9293

4.2. Explanation Performance

4.2.1. Global Explanation Evaluation

Once trained, the system provides global explanations in the form of a set of feature patterns that identify image features characteristic of the classes to be predicted. Each of the feature patterns learned by the system was evaluated on whether it corresponds to a feature of the class (HGG or LGG) that it is supposed to represent and whether the area with the highest response (red) is located within the tumor or tissue altered by the

presence of the tumor. A feature pattern is considered invalid if its most activated area is situated in the background regions, namely healthy tissue, ventricles, non-brain tissue, or image background.

Within all feature patterns, two apparent duplicates were found of the LGG class. Thirteen invalid 'background' feature patterns (6 HGG and 7 LGG) were found to have higher responses in regions irrelevant to the classification task (e.g., low-signal ventricles and high-intensity background areas). The accuracy of global explanation, defined as the fraction of learned feature patterns that focus on task-relevant regions, was 86%. The initial assessment process was conducted by one author (Y.T.X). In cases of ambiguity, feature patterns were reviewed by other authors (F.Z, L.R), and the final evaluation was arrived at by consensus. Considering the impact of invalid feature patterns on local explanation, those 'background' feature patterns were excluded in the further local analysis process.

Figure 7 evaluates the representability of two feature patterns that have the largest class connection weight of each class. The similarity score between the feature pattern (class connection of 0.737 to HGG) and the prototypical patch from the tumor core of the first HGG sample image ranks #2 with a similarity of 8.020 (max. 8.782) and #4 with a similarity of 7.653 (max. 8.379) with the prototypical patch from the tumor core of the second sample image, showing its high representativity of class HGG. The feature pattern of LGG with the highest class-connection value (1.311) also shows high representativity of class LGG; the similarity scores with the localized patches rank first among 10 feature patterns for two LGG sample images.

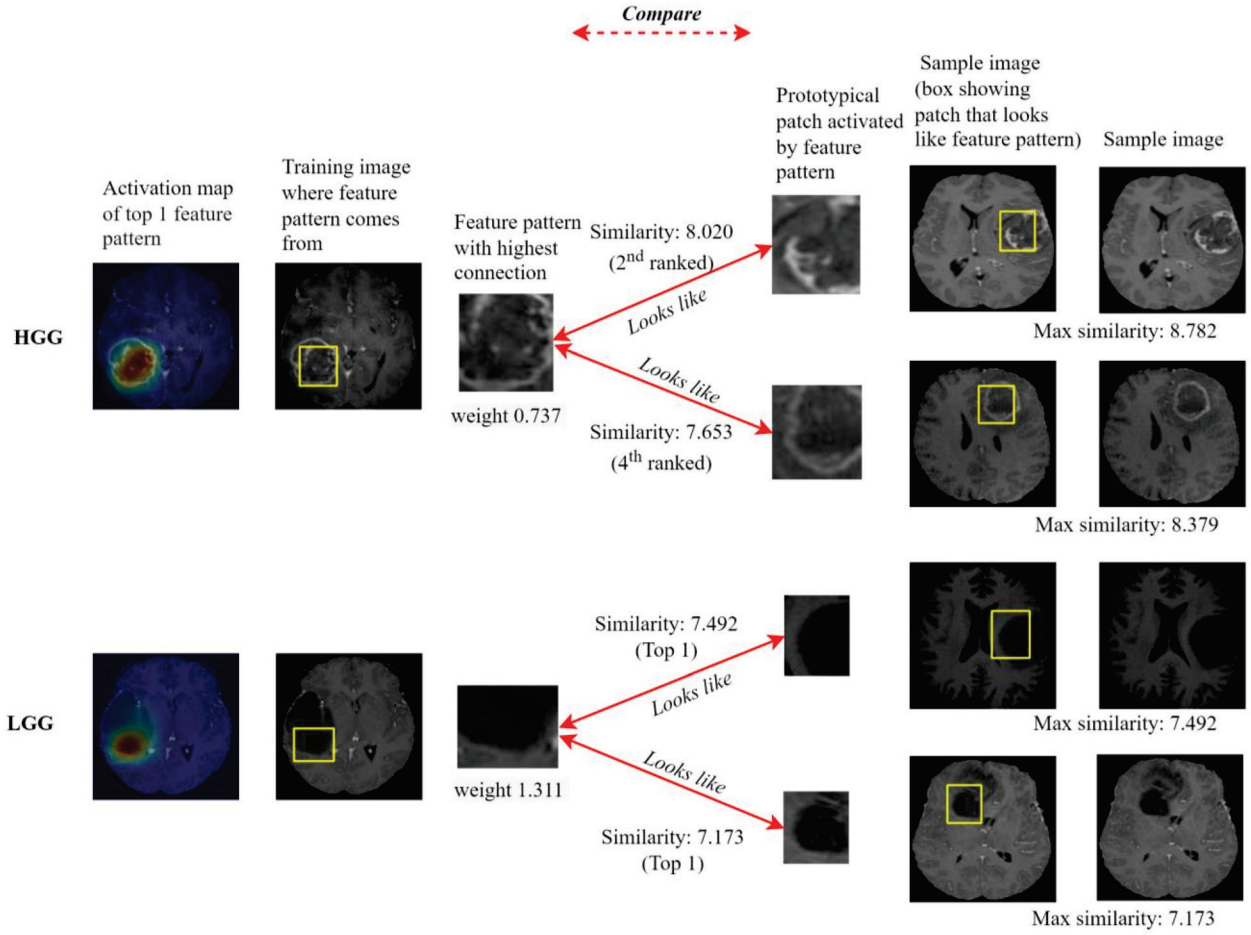


Figure 7. Representability of the feature patterns. The explanations on two input MR images are shown for the feature pattern that has the largest class-connection weight on each class.

4.2.2. Local Explanation Evaluation

The reliability of an individual prediction was evaluated based on whether its local explanation satisfies two basic reliability conditions, namely that the reasoning process should be both confident and correct.

Confidence in the reasoning process. Confidence in the reasoning process can be evaluated by examining the output of the local explanation. For each case in the test set, the number of feature patterns corresponding to a correctly or incorrectly identified tumor type was counted among those with the 10 highest similarity scores. The results were averaged and summarized in Table 2. Results demonstrate that the inconsistency of the feature patterns with the class of test images has a high impact on the classification performance of the model (comparison between wrong predictions and correct predictions, mean 8.62, 0.34, respectively) and a small impact on the reliability of the predictions (comparison between correct predictions and unreliable predictions).

Table 2. Summary of the number of feature patterns (top 10) consistent or inconsistent with the actual class of test images among all test cases, unreliable predictions, wrong predictions, and correct predictions (TP and TN predictions), summarized as mean (standard deviation) of the fraction of feature patterns among the top 10 that match or mismatch to the actual class of test images.

Class of Feature Pattern	Class of Test Image							
	All Test Cases		Unreliable Predictions [1]		Wrong Predictions [2]		Correct Predictions [3]	
	HGG	LGG	HGG	LGG	HGG	LGG	HGG	LGG
HGG	9.29 (2.27)	0.90 (2.26)	9.87 (0.41)	1.56 (1.17)	2.09 (1.27)	8.62 (1.26)	9.98 (0.17)	0.34 (0.84)
LGG	0.71 (2.27)	9.10 (2.26)	0.13 (0.41)	8.44 (1.17)	7.91 (1.27)	1.38 (1.26)	0.02 (0.17)	9.66 (0.84)

Note: [1] Unreliable predictions are cases among {TP, TN} predictions that are evaluated to be unreliable according to the two identified reliability criteria. [2] Wrong predictions are {FP, FN}. [3] Correct predictions are {unreliable predictions, reliable predictions}.

Correctness of the reasoning process. A correct reasoning process is defined as one in which the concept of the activated prototypical patch is consistent with that of the feature pattern. Table 3 summarizes the number of incorrectly activated background patches by top 10 feature patterns among all test images, unreliable predictions, wrong predictions, and correct predictions.

Table 3. The numbers of incorrectly activated background patches by the top 10 feature patterns were given as mean (standard deviation) of the fraction of feature patterns among the top 10 that mismatched the actual class of test images.

Concept of Activated Patch	Class of Test Image							
	All Test Images		Unreliable Predictions		Wrong Predictions		Correct Predictions	
	HGG	LGG	HGG	LGG	HGG	LGG	HGG	LGG
Image background area	0.33 (0.89)	0.46 (1.40)	1.37 (1.34)	1.85 (2.41)	1.46 (1.44)	1.23 (1.83)	0.23 (0.74)	0.40 (1.35)
	0.39 (1.14)		1.61 (1.96)		1.37 (1.57)		0.30 (1.06)	
Brain background area	0.65 (1.55)	0.97 (3.19)	2.61 (2.07)	4.46 (3.67)	2.32 (2.30)	1.00 (0.71)	0.43 (1.28)	0.97 (2.51)
	0.75 (11.96)		3.55 (3.11)		1.83 (1.96)		0.67 (1.93)	

Wilcoxon Signed-Ranks tests were used to assess the effect of incorrectly activated background patches on the number of mismatched feature patterns among the top 10 ranked, an indicator of prediction reliability, comparing image background areas and brain background (i.e., healthy tissue or CSF), for each classification class both separately and jointly.

Considering all test images, image background showed a significantly lower influence (p-value < 0.05) on reliability compared to brain background (W = 1244.5, p-value < 0.001),

and the same pattern was repeated considering only the HGG test images (W = 411.0, p-value = 0.002) or the LGG test images (W = 214.0, p-value = 0.005). Dividing the test images based on correct predictions (reliable predictions and unreliable predictions, HGG: W = 171.5, p-value = 0.011; LGG: W = 114.5, p-value = 0.003), wrong predictions (HGG: W = 53.5, p-value = 0.095 ($p > 0.05$), LGG: W = 17.5, p-value = 0.943 ($p > 0.05$)) and unreliable predictions (same values as correct predictions), only in the second group did these two sources of error show no difference in their effect.

Tables 2 and 3 indicate the necessity and importance of unboxing the inference process of CNN models for brain tumor classification. This allows health workers to screen out unreliable 'correct' predictions that might have been learned from irrelevant regions for decision making.

5. Discussion

This work proposed an interpretable multi-part attention network for brain tumor classification. In detail, the widely used VGG16 was built with a specific interpretable architecture to ensure good enough classification performance for the BRATS 2017 dataset. The model was evaluated in terms of both classification and explainability perspectives. Results demonstrated the model produced accurate tumor classification, and the classification accuracy is on par with some of the best-performing CNN models. Furthermore, the proposed framework is able to provide higher quality explanations for HGG and LGG classification, including global explanation and local explanation.

In detail, global explanation is interpreted as a set of feature patterns the model learns from to classify HGG and LGG. The quality of the feature patterns in terms of their validity and representativity was evaluated by radiologists to see if they were valid evidence for decision aids. Results demonstrated the model learns from the class-representative features of both classes for the classification task, and the HGG feature patterns have higher responses in the contrast-enhancing tumor, necrotic tumor core, and the edematous areas as classification evidence; this agrees with the actual imaging characteristics of HGG. The LGG feature patterns present higher responses on the homogeneous tumor cores and the non-enhancing tumor margins.

Another important advantage of the proposed model is the local explanation it presents for individual predictions. Background areas, such as the ventricles, were found to be activated by the 'tumor core' feature patterns of the LGG class. These background patches are not faithful features to the underlying lesion. Therefore, unboxing the reasoning process is necessary; it allows the clinicians and patients to screen out 'unreliable' correct predictions.

The local explanation of individual explanations was also evaluated by radiologists to see if it is reliable and acceptable for decision-making support. This form of reliability evaluation and model tuning is not available in the development of "black box" networks or the interpretable models mentioned above. According to the findings, the developed solution provided positive outcomes regarding the brain tumor classification and explanation targeted in this study.

Considering the limitations of the present study, these can be divided into methodological limitations in the construction of the network and limitations in the contextualization of the results.

It is reasonable to suppose that network construction limitations contribute to the lower classification accuracy of the proposed interpretable model compared with the baseline model. This discrepancy could be attributed to the model's classification inference process, which is greatly influenced by the feature patterns obtained from the randomly generated push dataset. In future work, optimizing the selection of the push dataset may help to improve the classification accuracy of the model. It is also possible that the training data augmentation process could be optimized, as some recent evidence suggests that, even though we used very widely used augmentation methods, the inclusion of image orientations not found in the testing set does not improve the generalizing ability of the model [39].

Regarding interpretation of the results, we did not find other interpretable deep learning methods applied to brain tumor classification based on the same dataset, and we cannot confirm the degree to which the 86% reliability obtained by the model would be considered acceptable by the health workers. Further collaboration with medical practitioners is important for the practical assessment of our model. Considering possible future developments or our work, several possible extensions are clear. The data modalities could be extended to incorporate a greater variety of structural images, such as T1w, T2w, and FLAIR, as well as more targeted sequences, including amide proton transfer [40] and MR spectroscopy [41]. It is also important to consider whether findings in the BraTS2017 dataset carry over into other datasets. For example, many clinical scanners continue to use lower field strengths. Publicly available data sets such as MNIBITE [42] and the recent ReMIND [43] could be leveraged to test IMPA-Net with 1.5-T data.

6. Conclusions

An interpretable classification model based on CNN was developed for brain tumor classification to enhance the interpretability and trustworthiness of the model and the health outcomes. The proposed model visualizes the features the model learns and uses for the classification task. It unboxes the reasoning process of individual predictions and explains the outcomes in a human-understandable manner, allowing clinicians and patients to understand and evaluate the reliability of predictions.

In future investigations, alternative datasets encompassing a greater variety of sequences and settings, will be included to improve the classification performance and the generality of the work. Further discussions on the quality of decision aids are also necessary to determine whether they improved decision making and outcomes for patients facing treatment or screening decisions and to explore the applicability of IMPA-Net in other medical imaging tasks.

Author Contributions: Conceptualization, C.T. (Claudia Testa), D.N.M., R.Z., F.Z., L.R. and Y.X.; methodology, R.Z. and Y.X.; hardware and software, R.Z. and Y.X.; formal analysis, C.T. (Caterina Tonon), C.T. (Claudia Testa), D.N.M., F.Z. and L.R.; investigation, C.T. (Claudia Testa), D.N.M., F.Z. and L.R.; resources, C.T. (Caterina Tonon), R.Z. and R.L.; data curation, D.N.M. and Y.X.; writing—original draft preparation, Y.X.; writing—review and editing, C.T. (Caterina Tonon), C.T. (Claudia Testa), D.N.M., F.Z., L.R. and R.Z.; supervision, C.T. (Caterina Tonon), C.T. (Claudia Testa) and D.N.M.; funding acquisition, C.T. (Caterina Tonon), R.L. and D.N.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the China Scholarship Council (grant number: 202008320283). The publication of this article was supported by the "Ricerca Corrente" funding from the Italian Ministry of Health.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The first author takes full responsibility for the analyses, interpretation, and conduct of the research. The underlying codes are available from the first author upon reasonable request. The data are publicly available at https://www.med.upenn.edu/sbia/brats2017/data.html (accessed on 10 May 2024).

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

References

1. Cancer Research UK. Brain, Other CNS and Intracranial Tumours Statistics. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brain-other-cns-and-intracranial-tumours/incidence#collapseTen#heading-One (accessed on 6 December 2023).
2. Louis, D.N.; Perry, A.; Wesseling, P.; Brat, D.J.; Cree, I.A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.K.; Pfister, S.M.; Reifenberger, G.; et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary. *Neuro. Oncol.* **2021**, *23*, 1231–1251. [CrossRef] [PubMed]
3. Norden, A.D.; Drappatz, J.; Wen, P.Y. Malignant Gliomas in Adults. *Blue Books Neurol.* **2010**, *36*, 99–120. [CrossRef]

4. Wirsching, H.G.; Weller, M. Glioblastoma. In *Malignant Brain Tumors: State-of-the-Art Treatment*; Springer International Publishing: Cham, Switzerland, 2016; pp. 265–288; ISBN 3319498649.
5. Fink, J.R.; Muzi, M.; Peck, M.; Krohn, K.A. Multimodality Brain Tumor Imaging: MR Imaging, PET, and PET/MR Imaging. *J. Nucl. Med.* **2015**, *56*, 1554–1561. [CrossRef] [PubMed]
6. Villanueva-Meyer, J.E.; Mabray, M.C.; Cha, S. Current Clinical Brain Tumor Imaging. *Clin. Neurosurg.* **2017**, *81*, 397–415. [CrossRef] [PubMed]
7. Grier, J.T.; Batchelor, T. Low-Grade Gliomas in Adults. *Oncologist* **2006**, *6*, 681–693. [CrossRef] [PubMed]
8. Ganz, J.C. Low Grade Gliomas. *Prog. Brain Res.* **2022**, *268*, 271–277. [CrossRef] [PubMed]
9. Forst, D.A.; Nahed, B.V.; Loeffler, J.S.; Batchelor, T.T. Low-Grade Gliomas. *Oncologist* **2014**, *19*, 403–413. [CrossRef]
10. Shen, D.; Wu, G.; Suk, H.-I. Deep Learning in Medical Image Analysis. *J. Imaging* **2021**, *7*, 74. [CrossRef] [PubMed]
11. Yasaka, K.; Akai, H.; Kunimatsu, A.; Kiryu, S.; Abe, O. Deep Learning with Convolutional Neural Network in Radiology. *Jpn. J. Radiol.* **2018**, *36*, 257–272. [CrossRef]
12. Xie, Y.; Zaccagna, F.; Rundo, L.; Testa, C.; Agati, R.; Lodi, R.; Manners, D.N.; Tonon, C. Convolutional Neural Network Techniques for Brain Tumor Classification (from 2015 to 2022): Review, Challenges, and Future Perspectives. *Diagnostics* **2022**, *12*, 1850. [CrossRef]
13. Nazir, M.; Shakil, S.; Khurshid, K. Role of Deep Learning in Brain Tumor Detection and Classification (2015 to 2020): A Review. *Comput. Med. Imaging Graph.* **2021**, *91*, 101940. [CrossRef] [PubMed]
14. Khazaee, Z.; Langarizadeh, M.; Ahmadabadi, M.E.S. Developing an Artificial Intelligence Model for Tumor Grading and Classification, Based on MRI Sequences of Human Brain Gliomas. *Int. J. Cancer Manag.* **2022**, *15*, e120638. [CrossRef]
15. Chikhalikar, A.M.; Dharwadkar, N.V. Model for Enhancement and Segmentation of Magnetic Resonance Images for Brain Tumor Classification. *Pattern Recognit. Image Anal.* **2021**, *31*, 49–59. [CrossRef]
16. El Hamdaoui, H.; Benfares, A.; Boujraf, S.; El Houda Chaoui, N.; Alami, B.; Maaroufi, M.; Qjidaa, H. High Precision Brain Tumor Classification Model Based on Deep Transfer Learning and Stacking Concepts. *Indones. J. Electr. Eng. Comput. Sci.* **2021**, *24*, 167–177. [CrossRef]
17. Zhuge, Y.; Ning, H.; Mathen, P.; Cheng, J.Y.; Krauze, A.V.; Camphausen, K.; Miller, R.W. Automated Glioma Grading on Conventional MRI Images Using Deep Convolutional Neural Networks. *Med. Phys.* **2020**, *47*, 3044–3053. [CrossRef]
18. Kaufman, L.; Kramer, D.M.; Crooks, L.E.; Ortendahl, D.A. Measuring Signal-to-Noise Ratios in MR Imaging. *Radiology* **1989**, *173*, 265–267. [CrossRef] [PubMed]
19. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for Xai in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
20. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [CrossRef]
21. Zhang, Y.; Tino, P.; Leonardis, A.; Tang, K. A Survey on Neural Network Interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *5*, 726–742. [CrossRef]
22. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, *32*, 4793–4813. [CrossRef]
23. Vellido, A. The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care. *Neural Comput. Appl.* **2020**, *32*, 18069–18083. [CrossRef]
24. Pereira, S.; Meier, R.; Alves, V.; Reyes, M.; Silva, C.A. Automatic Brain Tumor Grading from MRI Data Using Convolutional Neural Networks and Quality Assessment. *Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **2018**, *11038*, 106–114. [CrossRef] [PubMed]
25. Artzi, M.; Redmard, E.; Tzemach, O.; Zeltser, J.; Gropper, O.; Roth, J.; Shofty, B.; Kozyrev, D.A.; Constantini, S.; Ben-Sira, L. Classification of Pediatric Posterior Fossa Tumors Using Convolutional Neural Network and Tabular Data. *IEEE Access* **2021**, *9*, 91966–91973. [CrossRef]
26. Marmolejo-Saucedo, J.A.; Kose, U. Numerical Grad-Cam Based Explainable Convolutional Neural Network for Brain Tumor Diagnosis. *Mob. Networks Appl.* **2022**. [CrossRef]
27. Gaur, L.; Bhandari, M.; Razdan, T.; Mallik, S.; Zhao, Z. Explanation-Driven Deep Learning Model for Prediction of Brain Tumour Status Using MRI Image Data. *Front. Genet.* **2022**, *13*, 822666. [CrossRef]
28. Thomson, W.; Roth, A.E. The Shapley Value: Essays in Honor of Lloyd S. Shapley. *Economica* **1991**, *58*, 123–124. [CrossRef]
29. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should i Trust You?" Explaining the Predictions of Any Classifier. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2016**, *13*, 1135–1144. [CrossRef]
30. Pintelas, E.; Liaskos, M.; Livieris, I.E.; Kotsiantis, S.; Pintelas, P. Explainable Machine Learning Framework for Image Classification Problems: Case Study on Glioma Cancer Prediction. *J. Imaging* **2020**, *6*, 37. [CrossRef] [PubMed]
31. Niu, Z.; Zhong, G.; Yu, H. A Review on the Attention Mechanism of Deep Learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
32. Chen, C.; Li, O.; Tao, C.; Barnett, A.J.; Su, J.; Rudin, C. This Looks like That: Deep Learning for Interpretable Image Recognition. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
33. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef] [PubMed]

34. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing The Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features. *Sci. Data* **2017**, *4*, 170117. [CrossRef] [PubMed]
35. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv* **2018**, arXiv:1811.02629.
36. Singh, G.; Yow, K.C. These Do Not Look like Those: An Interpretable Deep Learning Model for Image Recognition. *IEEE Access* **2021**, *9*, 41482–41493. [CrossRef]
37. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034. [CrossRef]
39. Reyes, D.; Sánchez, J. Performance of Convolutional Neural Networks for the Classification of Brain Tumors Using Magnetic Resonance Imaging. *Heliyon* **2024**, *10*, e25468. [CrossRef] [PubMed]
40. Guo, P.; Unberath, M.; Heo, H.Y.; Eberhart, C.G.; Lim, M.; Blakeley, J.O.; Jiang, S. Learning-Based Analysis of Amide Proton Transfer-Weighted MRI to Identify True Progression in Glioma Patients. *NeuroImage Clin.* **2022**, *35*, 103121. [CrossRef] [PubMed]
41. Ranjith, G.; Parvathy, R.; Vikas, V.; Chandrasekharan, K.; Nair, S. Machine Learning Methods for the Classification of Gliomas: Initial Results Using Features Extracted from MR Spectroscopy. *Neuroradiol. J.* **2015**, *28*, 106–111. [CrossRef] [PubMed]
42. Laurence, M.; Rolando, F.D.M.; Kevin, P.; David, A.; Claire, H.; D Louis, C. Online Database of Clinical MR and Ultrasound Images of Brain Tumors. *Med. Phys.* **2012**, *39*, 3253–3261. [CrossRef]
43. Juvekar, P.; Dorent, R.; Ogl, F.K.; Torio, E.; Barr, C.; Rigolo, L.; Galvin, C.; Jowkar, N.; Kazi, A.; Haouchine, N.; et al. ReMIND: The Brain Resection Multimodal Imaging Database. *medRxiv* **2023**. [CrossRef]

diagnostics

MDPI

Article

Three-Dimensional Measurement of the Uterus on Magnetic Resonance Images: Development and Performance Analysis of an Automated Deep-Learning Tool

Daphné Mulliez [1,*], Edouard Poncelet [1], Laurie Ferret [2], Christine Hoeffel [3], Blandine Hamet [1], Lan Anh Dang [1], Nicolas Laurent [1] and Guillaume Ramette [1]

[1] Service d'Imagerie de la Femme, Centre Hospitalier de Valenciennes, 59300 Valenciennes, France; poncelet.edouard@gmail.com (E.P.); blandine.hamet@gmail.com (B.H.); dng.lananh@gmail.com (L.A.D.); laurent-n@ch-valenciennes.fr (N.L.); ramette.g@gmail.com (G.R.)
[2] Unité de Recherche Clinique, Centre Hospitalier de Valenciennes, 59300 Valenciennes, France; ferret-l@ch-valenciennes.fr
[3] Service de Radiologie, Hôpital Maison Blanche, Avenue du Général Koenig, 51092 Reims, France; choeffel-fornes@chu-reims.fr
* Correspondence: daphnemllz@gmail.com

Abstract: Uterus measurements are useful for assessing both the treatment and follow-ups of gynaecological patients. The aim of our study was to develop a deep learning (DL) tool for fully automated measurement of the three-dimensional size of the uterus on magnetic resonance imaging (MRI). In this single-centre retrospective study, 900 cases were included to train, validate, and test a VGG-16/VGG-11 convolutional neural network (CNN). The ground truth was manual measurement. The performance of the model was evaluated using the objective key point similarity (OKS), the mean difference in millimetres, and coefficient of determination R^2. The OKS of our model was 0.92 (validation) and 0.96 (test). The average deviation and R^2 coefficient between the AI measurements and the manual ones were, respectively, 3.9 mm and 0.93 for two-point length, 3.7 mm and 0.94 for three-point length, 2.6 mm and 0.93 for width, 4.2 mm and 0.75 for thickness. The inter-radiologist variability was 1.4 mm. A three-dimensional automated measurement was obtained in 1.6 s. In conclusion, our model was able to locate the uterus on MRIs and place measurement points on it to obtain its three-dimensional measurement with a very good correlation compared to manual measurements.

Keywords: deep learning; convolutional neural network; artificial intelligence; uterus; measurement; MRI

Citation: Mulliez, D.; Poncelet, E.; Ferret, L.; Hoeffel, C.; Hamet, B.; Dang, L.A.; Laurent, N.; Ramette, G. Three-Dimensional Measurement of the Uterus on Magnetic Resonance Images: Development and Performance Analysis of an Automated Deep-Learning Tool. *Diagnostics* **2023**, *13*, 2662. https://doi.org/10.3390/diagnostics13162662

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 3 July 2023
Revised: 8 August 2023
Accepted: 10 August 2023
Published: 12 August 2023

1. Introduction

Several variations can be observed in the size of the female genitalia, especially the uterus [1]. First, there are individual physiological factors such as age (pre-pubertal phase, time of genital activity, menopause) and such as the natural changes in the aspect of the uterus occurring during gestation. A uterine deformation could also be caused by personal endometrial or myometrial pathologies. Ultimately, there is a significant inter-individual variability depending on everyone regarding the size of the uterus [2].

Uterus measurements are undoubtedly useful for assessing both the treatment and follow-ups of gynaecological patients. Evaluation of the size of the uterus helps describe the development and senescence of the organs, choose the best procedures, and assist surgical procedures such as laparoscopy or laparotomy.

From a clinical perspective, knowing the size of the uterus is useful to diagnose delayed or precocious puberty [3]. Data on the sagittal uterine length and body-to-cervix ratio can be used to conclude whether the internal genitalia are of pubertal or non-pubertal morphology [4]. Also, describing the size of the uterus in the case of polymyomatous pathology is an indicator of the position of the uterus in the patient's abdomen and enables clinicians to appreciate the potential urinary and digestive repercussions [5]. Furthermore,

measuring total uterine length is essential before inserting an Intra Uterine Device (IUD). It helps to determine where to stop IUD insertion and avoid insertion problems such as perforation [6]. The width of the uterine cavity is measured to ensure that T-shaped devices fit properly to the cavity, thus avoiding dimensional disproportions that could lead to problems of effectiveness or material displacement [7].

From a surgical perspective, knowing the size of the uterus helps to estimate the best approach for removing the organ when a hysterectomy may be indicated (laparoscopy, laparotomy, vaginal approach...) [8]. Also, it is necessary to measure the transverse diameter of the uterus before endometrial ablation by thermal balloon, cryotherapy, radiofrequency or microwave energy to avoid damaging the endocervical canal [9]. For the NovaSure electrosurgical technique in particular, the radiofrequency controller needs to know the values for the intercornal width of the uterine fundus and the length of the uterine cavity, which must be between 40 and 65 mm [1]. Techniques for treating genital prolapse by promontofixation will also depend on the size of the uterus [10].

In the obstetrical field, measuring uterine length provides important information to monitor the progress of the pregnancy and to know whether fetal growth is progressing correctly [11]. If the size of the uterus is too small or too large, this may lead to an ultrasound check on fetal growth, the amount of amniotic fluid, and the appearance of the placenta. Measurements of the uterus can help decide whether curettage is necessary in the event of a failed pregnancy (spontaneous miscarriage or voluntary termination of pregnancy). As the width of the endometrial cavity is correlated with gestational age, measuring it at the time of abortion could help avoid many surgical procedures [12]. Researchers have postulated that by using the mean of cavity width and area plus two standard deviations as the upper limit of cavity size (width 50 mm, area 60 cm^2), then only 44% of patients in their study would have required curettage [1]. Measuring uterine length is also particularly useful for locating the most suitable intrauterine site for the development of fertilized eggs for in vitro fertilisation. A study by Egbase et al. showed that implantation and pregnancy rates were higher when the total length of the uterus was between 70 and 90 mm [13]. They placed the embryos 5 mm from the uterine fundus.

For all these reasons, an accurate description of uterine measurements must be made in the magnetic resonance (MR) examination of the female pelvis [14].

Usually, the measurement is performed manually by radiologists, wasting medical time and leading to inter-operator variability [15].

Artificial intelligence (AI) is a scientific discipline based on the premise that the faculties of human intelligence can be simulated by a machine. Machine learning consists in creating algorithms capable of learning how to model functions and make predictions using a process of generalisation of the data induced. Deep learning is a sub-type of machine learning, using neural networks that are particularly effective in image processing, as they can learn spatial determinants.

Artificial intelligence has become a potential solution to assist segmentation [16,17]. Thus, AI can develop an automatic segmentation tool for given structures and enable significant improvements in radiological workflows [18].

Deep learning models have been successful in providing automatic segmentation for different organs such as the prostate, kidneys, and heart [19–22].

Measuring the uterus is a more challenging task due to anatomical variability and complex contrasts with surrounding tissues, but also in relation to pathologies such as endometriosis or myomas that distort the contours of the uterus [11].

Along with ultrasound, magnetic resonance imaging (MRI) is the best imaging modality for the diagnosis of pelvic pathologies in women. Furthermore, it provides better reproducibility because it is not observer dependent.

A previous study presents a method for automatic uterus segmentation in MRI for patients with uterine fibroids undergoing ultrasound-guided HIFU therapy. They used a 3D nnU-Net model in order to automate uterine volumetry for tracking changes after therapy [23].

Another recent study showed that using a combination of deep learning reconstruction and a sharpening filter markedly increases the image quality of SSFSE of the uterus to the level of the PROPELLER sequence [24].

The aim of our work was to develop, validate, and test an automated deep learning tool for MR images, to implement an automatic measurement of the three-dimensional size of the uterus, and to evaluate its performance compared with the manual measurements of radiologists.

2. Material and Methods

This retrospective study of model creation was approved by the Independent Ethics Committee under the reference CHV-2022-006. All patients were informed of the use of their medical data according to the legal framework imposed by CNIL MR-004. In addition, all data were pseudonymized beforehand.

2.1. Data Acquisition

All women over 18 years of age who underwent pelvic MRI, including sagittal and axial T2-weighted images, in the women's imaging department of Valenciennes Hospital (France) between September 2021 and March 2022 were retrospectively collected from the Institutional Picture Archiving and Communication System (Electronic Medical Record Entreprise, VEPRO AG, Version 8.2, Pfungstadt, Germany).

Pregnant women were excluded, as were MR images with severe motion artefacts, a highly deviated uterus, and subserous myomatous pathology (FIGO VI and VII, due to important deformation of the uterus).

These examinations were performed using two MR units, either at 1.5 T (SIGNA artist) or 3 T (SIGNA premier) (General Electric Healthcare, Cleveland, OH, USA). The acquisition parameters are listed in Table 1.

Table 1. Magnetic resonance imaging (MRI) acquisition parameters.

Parameters \ **MRI**	**General Electric 1.5 T, SIGNA Artist, 2021**		**General Electric 1.5 T, SIGNA Artist, 2020**		**General Electric 3 T, SIGNA Premier, 2019**	
Plane	Sagittal	Axial	Sagittal	Axial	Sagittal	Axial
TE (ms)	100–120		100–120		115–120	
TR (ms)	5000–1100		4000–10,000		4000–13,000	
Number of excitations (Nex)	2		2		1.5	2
Field of view (mm)	(393 × 260)–(408 × 220)	363 × 240	393 × 260	(360–240)–(410 × 270)	332 × 220	
Frequency (Hz)	41.67		41.67		50	
Slice thickness (mm)	3.5–4.0		3.5–4.0		3.0–3.5	
Interslice gap (mm)	3.5		3.5		0.5–3.0	

Patients were randomly assigned to training (80%) and validation sets (20%) without any overlap.

An additional set of MR images acquired between July and August 2021 was used for external validation using the same inclusion and exclusion criteria (test set).

2.2. Data Labelling

One radiologist (DM) used artificial intelligence software specifically designed for medical imaging (Cleverdoc V1.9.0 platform, Cleverdoc Entreprise, Lille, France) for labelling.

When the uterus was not seen, a label of "not observable" was attributed to the examen. When visible, it was targeted by a square area (a label box was drawn on the uterus). Three consecutive cuts were annotated with points corresponding to the measurements of the uterus (Figure 1). Thickness and length were noted on sagittal T2-weighted images. The length was measured using two different methods: one major axis defined by two points,

and the other defined by three points passing through the fundus, the cervix-isthmus junction, and the exocervix (Figure 2). Radiologists often prefer the three-point measurement when the uterus is significantly flexed. The width was labelled on the axial T2-weighted images. These manual measurements were used to define the ground truth.

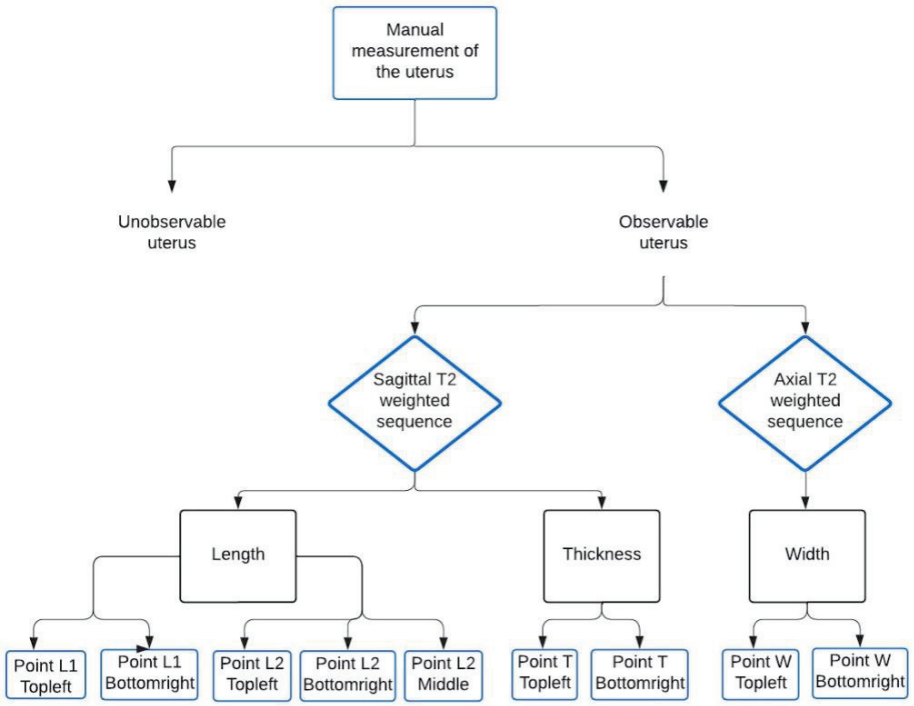


Figure 1. Flowchart of the data labelling.

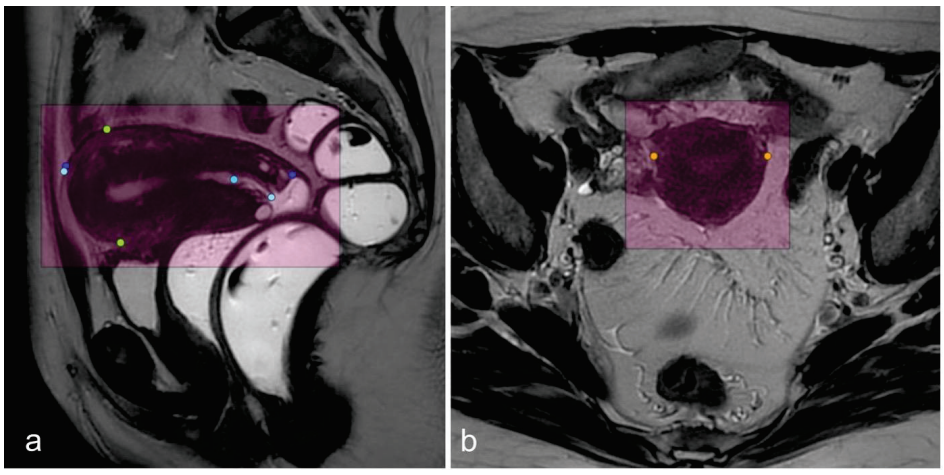


Figure 2. (**a**): Sagittal T2-weighted image with the label "observable" (purple), length label with two points (dark blue), length label with three points (light blue), thickness label (green). (**b**): axial T2-weighted image with label class "observable" (purple) and width label (orange). The vagina and the rectum were filled with ultrasound gel in order to respond to the clinical indication of this case.

2.3. Model

The overall pipeline works due to two separate models working successively.

First, a box model finds the uteri. Then, a key point model is used to place the measurement points on the cropped version of the image by placing two points for each class. This is achieved by splitting each class into two subclasses. The point for each subclass is determined by its position (top-left or bottom-right). For a length composed of three points, the midpoint is given in a separate preprocessing step.

We used convolutional neural network (CNN) architectures with an encoder or decoder pattern.

Encoder

The encoder network was composed of VGG-16 (box model) and VGG-11 (key point model). It receives an image with a size of 224 × 224 pixels as input. Subsequently, it passes through five blocks which are separated by a Max Pooling layer (Kernel = 2.2; Stride size = 2; No padding) that halves the height and width of the features (Figure 3).

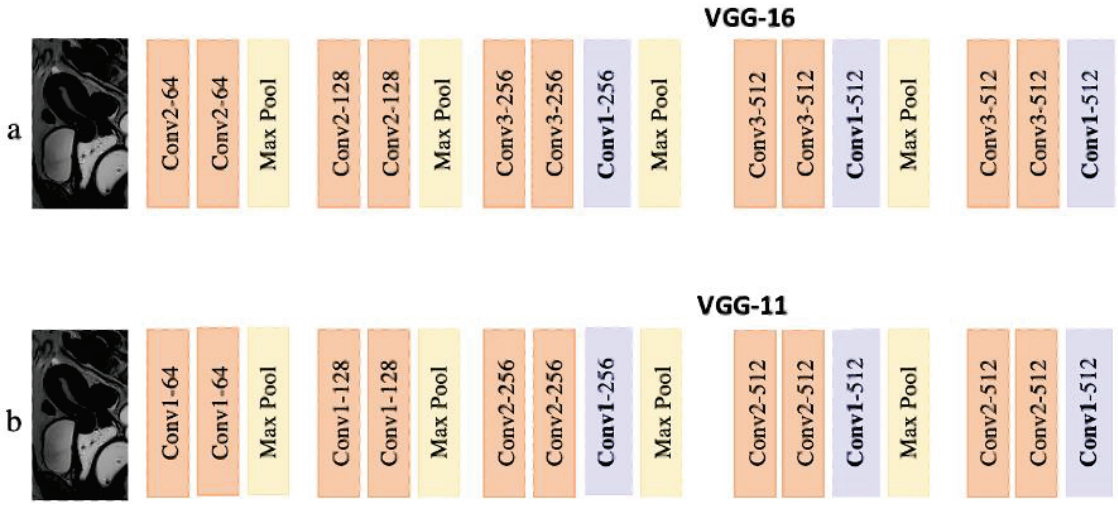


Figure 3. (**a**). VGG-11- modified architecture of the encoder for the box model. (**b**). VGG-16 modified architecture of the encoder for the keypoint model.

Decoder

Our model's decoders are single-instance boxes and key point detectors, which produce one box and one key point instance for each output class, respectively.

The input of the head is a four-dimensional vector of shape. The outputs are four values per class (x, y, w, h) for the position of the box and two values per class (x, y) for key point positions.

2.4. Training

We applied data augmentation techniques such as vertical and horizontal flipping, random rotation of a multiple of 90°, and translation of up to 0.1. Furthermore, for the key point model, we applied more techniques, such as changing the brightness and contrast, blurring the image, applying Gaussian noise, or reversing the image's colours.

We sorted our data based on the classes we wanted to train in ("uterus" classes for the box model and "keypoint" classes for the key point model). We applied a balancer that always kept a batch size of 10 with an equivalent number of items for each class in order to ensure the proper distribution of losses and metrics.

The box model was run for 100 epochs (220,783 iterations), and the key point model was run for 250 epochs (40,320 iterations).

The Adam optimiser was used with a learning rate of 0.0001 to optimise the weight of the model.

2.5. Statistics

2.5.1. Model Training

We kept track of the model's losses and calculated the metrics: The intersection over union (IoU) to evaluate the accuracy of box positions and objective key point similarity (OKS) for each key point class. This metric quantifies the closeness of the predicted key point location by using the ground-truth key point. The closer the predicted key point is to

the ground truth, the closer the OKS approach is to 1. Above 0.80, the model is considered very good. This metric is calculated as follows:

$$\mathrm{OKS} = \exp\left(-\frac{d^2}{2s^2k^2}\right) \quad (1)$$

where d is the distance between the ground truth key point and predicted key point, s is the area of the bounding box divided by the total image area, and k is the per-key point constant that controls the fall-off.

2.5.2. Model Testing

Four experts with experience in genital imaging (EP, GR, BH, and LD with 12, 2, and 1 years of experience, respectively) manually measured the size of the uterus in three dimensions on every MR image of the test set. The radiologists were blinded to the measurements made by others and to the machine. The same viewer (ViewerCleverdoc1.9.0) was used.

To evaluate the performance of our deep learning tool, we calculated the absolute average difference (in millimetres) between the measurements of one dataset and those of another. The coefficient of determination R^2 was used to reflect how well the AI measurements matched those of the radiologists using a linear regression model. If R^2 is equal to 1, the algorithm obtains strictly identical measurements to those of the radiologists.

The Cleverdoc V1.9.0 platform (Lille, France) was used for the statistical analyses.

3. Results

A total of 845 MRI scans were collected. Moreover, 45 patients were excluded: 37 because of myomas, 6 because of pregnancies, 3 because of a highly deviated uterus, and 2 because of poor image quality. The characteristics of the patients are shown in Table 2.

Table 2. Characteristics of the patients whose data were included for training, validation, and testing of the model.

	Training and Validation Set (*n* = 800)	**Test Set (*n* = 100)**
Age (mediane (interquartiles))	45 (33–58)	47 (34–56)
Gel vaginal markup		
No	436 (65%)	60 (60%)
Yes	364 (45%)	40 (40%)
Uterus position		
Anteflexed	704 (88%)	93 (93%)
Retroflexed	96 (12%)	7 (7%)
MRI without pelvic pathology	177 (22%)	26 (26%)
Subperitoneal endometriosis	123 (15%)	13 (13%)
Adenomyosis	116 (14%)	12 (12%)
Myomas (FIGO 0—V)	124 (15%)	19 (19%)
Cervical cancer	23 (3%)	2 (2%)
Endometrial pathology	75 (9%)	10 (10%)
Ovarian pathology	165 (21%)	16 (16%)
Hysterectomy	50 (6%)	-
Uterine malformation	7 (0.9%)	1 (1%)
Other (static disorder, no-gynaecological pathology…)	82 (10%)	13 (13%)

From the 800 included patients for training and validation, 4800 sets were obtained (three consecutive slices centred on the uterus for each sagittal and axial sequence).

An additional external cohort of 100 MR images was used for the model testing (Figure 4).

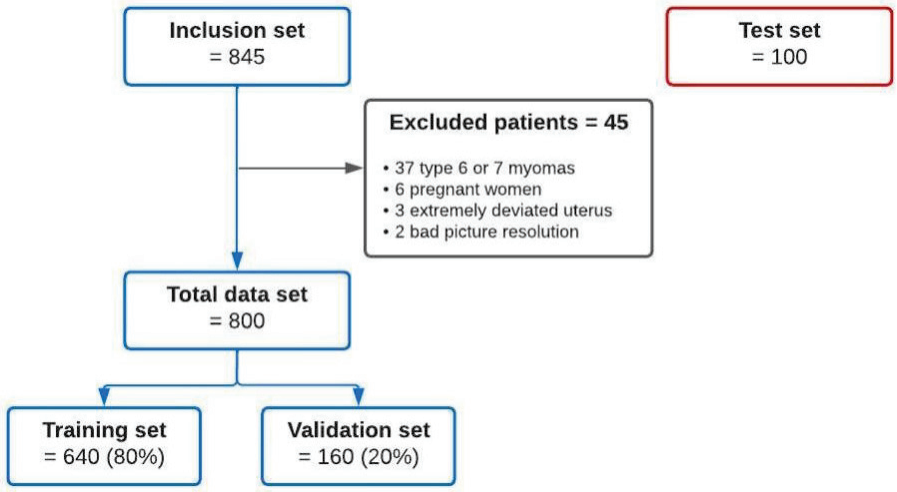


Figure 4. Overview of the workflow for training and testing the automated convolutional neural network (CNN) tool for measurement of the uterus.

3.1. Validation Performance (Initial Dataset)

During the validation phase, the algorithm was able to locate the uterus and the measurement key points with excellent accuracy.

With the measurement by DM as ground truth, the mean OKS was 0.92, ranging from 0.90 and 0.94 (Table 3). The OKS was calculated using the cropped images obtained using the box model.

Table 3. Objective key point similarity (OKS) values of the algorithm for each measurement key point, with the measurement by DM as ground truth.

Key Point	Length2 Top left (L1)	Length2 Bottom right (L2)	Length2 Middle (L3)	Length1 Top left (L4)	Length2 Bottom right (L5)	Width Top left (W1)	Width Bottom right (W2)	Thickness Top left (T1)	Thickness Bottom right (T2)	Average (av)
OKS	0.92	0.90	0.94	0.90	0.90	0.94	0.93	0.92	0.93	0.92

3.2. Test Performance (External Dataset)

We observed an improvement in the accuracy of our model when we switched to a new unknown cohort for testing. With the average of radiologists' measurements as ground truth, the mean OKS of our DL tool was 0.96, ranging from 0.95 and 0.98, as reported in Table 4. The OKS was calculated using full-size images.

Table 4. Objective key point similarity (OKS) values of the algorithm and of each radiologist, for each measurement key point, with the average of measurements by radiologists as ground truth.

Key point	Length2 Top left (L1)	Length2 Bottom right (L2)	Length2 Middle (L3)	Length1 Top left (L4)	Length1 Bottom right (L5)
GR	0.96	0.96	0.98	0.95	0.95
ED	0.97	0.97	0.98	0.96	0.95
LD	0.96	0.96	0.97	0.97	0.96
BH	096	0.96	0.97	0.96	0.95
AI	0.95	0.95	0.97	0.96	0.95

Table 4. *Cont.*

Key point	Width Top left (W1)	Width Bottom right (W2)	Thickness Top left (T1)	Thickness Bottom right (T2)	Average (av)
GR	0.99	0.98	0.94	0.94	0.96
ED	0.99	0.98	0.95	0.95	0.97
LD	0.99	0.98	0.95	0.95	0.97
BH	0.98	0.97	0.95	0.95	0.96
AI	0.97	0.98	0.95	0.94	0.96

Regarding the execution speed, it took less than 5 min for the model to extract all measurements, that is, one three-dimensional measurement in approximately 1.6 s. In comparison, the average time for a three-dimensional measurement by a radiologist was clocked at 37.89 s.

3.2.1. Correlation between Manual and Automated Measurements

Out of the 100 MR images of the test set, the average deviation between AI measurements and radiologists' measurements was 3.6 mm (±6.6 standard deviation (SD)). The distribution of the gaps was as follows: 3.9 mm for two-point length, 3.7 mm for three-point length, 2.6 mm for width, and 4.2 mm for thickness (Figure 5). The details of the absolute measurements are listed in Table 5.

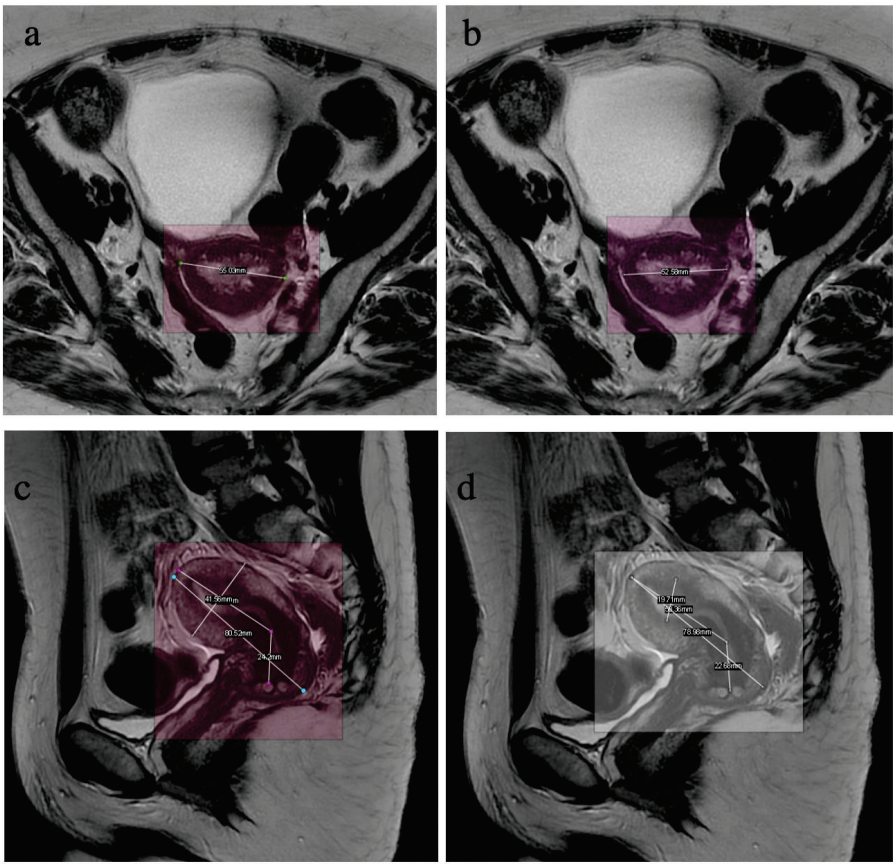


Figure 5. Sample results of the test phase. Width measurements on a T2-weighted axial sequence by radiologist (GR) (**a**) and by the algorithm (**b**). Lengths and thickness measurements on a T2-weighted sagittal sequence by a radiologist (DM) (**c**) and by the algorithm (**d**). We can notice the algorithm's error in detection of the contours for the thickness measurement.

Table 5. Statistics of uterine dimension measurements by the radiologists (ground truth) and by the algorithm (AI).

	Minimum (mm)		Maximum (mm)		Median (mm)		Average (mm)		Standard Deviation (SD)	
	Ground Truth	AI	Ground Truth	AI	Ground Truth	AI	Ground Truth	AI	Ground Truth	AI
Length 1 (2 points)	44.73	1.59	123.19	127.94	76.46	76.93	79.95	78.09	17.13	19.28
Length 2 (3 points)	42.14	46.04	123.96	119.75	77.23	74.69	79.43	76.19	16.43	15.46
Thickness	19.78	6.3	73.56	74.08	39.23	36.90	39.89	36.23	11.13	11.95
Width	32.16	35.95	93.65	90.98	52.72	52.58	54.42	54.60	12.35	11.06

The R^2 coefficients of determination between the algorithm's measurements and the average of the radiologists' measurements were 0.93 for two-point length, 0.94 for three-point length, 0.93 for width, and 0.75 for thickness, as shown in Figure 6.

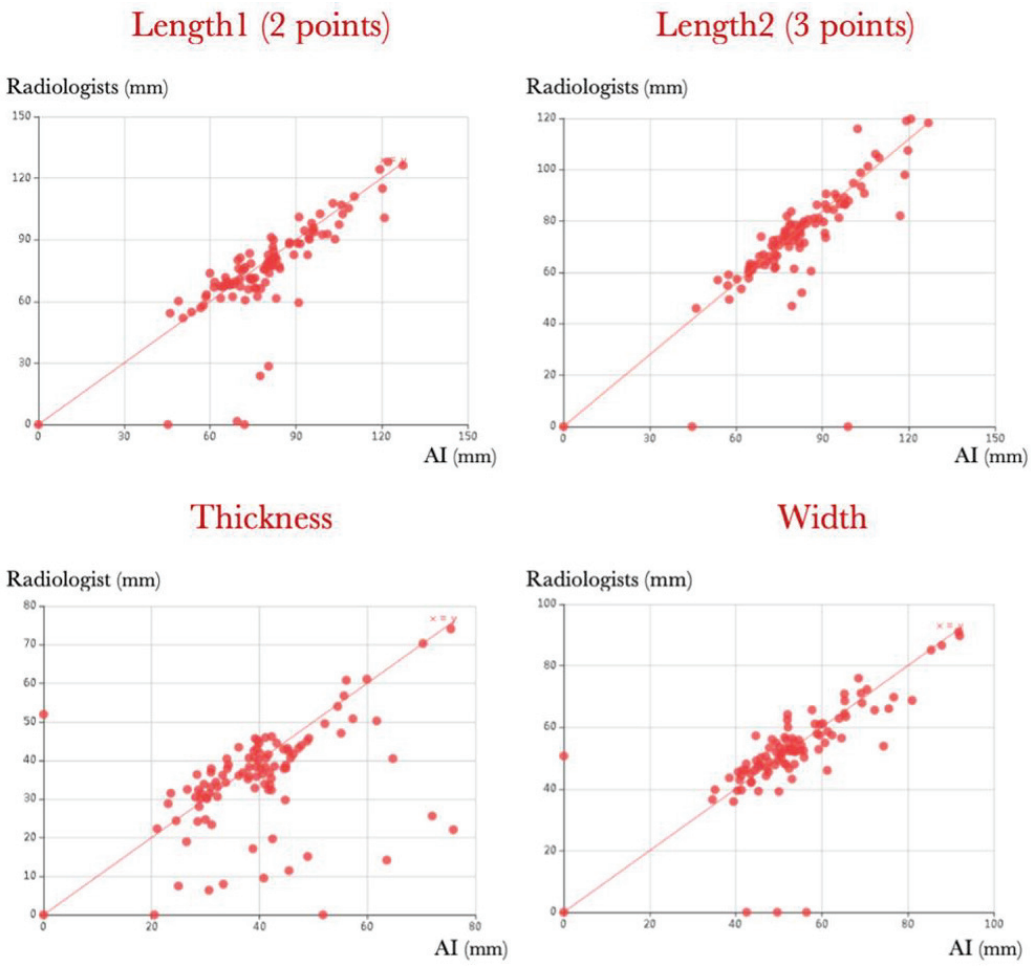


Figure 6. Scatter plots showing the correspondence between artificial intelligence (AI) measurements (in abscissa) and the average of the radiologist's measurements (in ordinate) in millimetres.

3.2.2. Variability between Measurements by Radiologists

The mean difference in measurements between all radiologists was 1.4 mm, detailed as follows: 1.27 mm for two-point length, 2.2 mm for three-point length, 1.14 mm for width, and 0.93 mm for thickness. The differences between each measure by one radiologist and the average of the others are presented in Table 6.

Table 6. Average deviation (AD) in millimeters between all radiologists.

Measures	Length2	Length2	Width	Thickness
EP	2.11	1.37	0.97	0.93
LD	2.12	1.2	1.12	0.92
BH	2.23	1.42	1.31	1.08
GR	2.36	1.11	1.15	0.8
Average	2.2	1.27	1.14	0.93

We performed a secondary analysis to highlight the distribution of the AI errors. Nineteen out of 200 images (100 axial + 100 sagittal) had an absolute deviation (averaged over all image measurements) discreetly greater than 8 mm. This represents less than 10% of the total number of examinations.

4. Discussion

We successfully achieved our goal of developing an artificial intelligence algorithm that is able to locate the uterus in pelvic MR examinations, place measurement key points on it, and provide its three-dimensional measurement with satisfactory accuracy.

The OKS was close to 1, improving from 0.92 (validation) to 0.96 (test). These results can be explained by the fact that the OKS of the validation phase were calculated based on the cropped images, whereas those of the test phase were calculated from the full-size images. The larger the image, the smaller the positioning error.

However, the algorithm remains highly performant when encountering new brand images. One of the strengths of our study is that our network was tested in an external cohort which did not have a selection bias applied, except for subserous myomas. This performance favours the generalisation of this model.

To the best of our knowledge, only one study of segmentation of images of the uterus has been conducted to date. Kurata et al. evaluated a U-net architecture to contour the uterus on MR images in the sagittal plane [25]. They reached an average DSC score (dice similarity coefficient, which can be compared to OKS) of 0.82. This study included 122 patients with or without uterine disorders. Our model was optimised by using a substantially larger training database of 800 patients.

In parallel, for men, a wide range of studies have been carried out on the automatic segmentation of images of the prostate, with similar results [26,27]. For example, Alexander Ushinsky et al. trained a customized hybrid 3D-2D U-Net CNN architecture on manually segmented MR images and had a DSC score of 0.898 [28].

However, it is more complex for an AI tool to locate and segment images of the uterus than of the prostate because the uterus can have different positions, bends, or shapes. Moreover, the uterus is surrounded by many elements (colon, bladder, and ovaries).

Another highlight of our study is that our training dataset was strengthened by the clinical heterogeneity of its cases, both in terms of pathological conditions and patient preparation. It included examinations for cervical cancer, endometriosis, and an ultrasound gel. This suggests that the performance of our CNN would be robust in prospective clinical settings.

Most studies on automated segmentation have used volumetric models or U-Net architectures [29,30]. In contrast, our network's performance was achieved with the VGG-11/16 architectures. This is a major aspect that led to the success of our algorithm. This model is more suitable for distance measurements because it is specifically designed to locate an organ and place measurement points on it. To do so, our pipeline operates using two different models (box and key point models).

The average deviation between the AI measurements and those of the radiologists was 3.6 mm ($\pm$6.6 SD), while the inter-radiologist variability was 1.4 mm. However, the R^2 coefficient was approaching 0.94 for lengths and width, meaning the coherence remained

extremely strong between the radiologists and AI. For thickness, however, the R^2 coefficient was 0.75, owing to the algorithm being challenged by the junctional zone in rare cases.

The speed of our system is a major advantage over the time required for manual segmentation. In our experience, it takes a radiologist 37.89 s to measure a uterus in three dimensions, set against 1.6 s for the algorithm. Our VGGnet may increase the throughput.

Our algorithm has the ability to overthrow a basic task, thus saving radiologists time for significant intellectual tasks.

Our study had a few limitations that should be acknowledged. First, this was a retrospective, monocentric study. The database was created using three MRI scanners (General Electric Healthcare, Valenciennes Hospital, France). The generalisability to other centres or MRI equipment has not yet been established. We subsequently included images obtained using the same T2-weighted acquisition protocols. We can imagine a comparative study of the performance of the algorithm between different MRI parameters or protocols.

In this first approach to developing a three-dimensional measurement software, we preferred to exclude pregnant women to make easier the algorithm training. In a further study, we could try to include examinations with pregnant uteruses to make our algorithm more inclusive. This could provide useful information for obstetricians.

We can easily imagine a clear application of our AI tool in daily practice. The measurements of the algorithm can be displayed on the image server or automatically added to reports. Valuable information on uterine size could help gynaecologists and surgeons in their daily practice. A patient's state of genital activity could be precisely described, as well as certain pathological conditions such as myomatous disease. Gynaecological procedures such as IUD insertion, hysterectomy, endometrial resection, or promontoxiation could be facilitated. These uterine measurements could also certainly be used in the obstetrical follow-up of patients.

Subsequent studies are required to prospectively validate our network in a clinical setting. We could consider further studies using the same pipeline to measure endometrial thickness or ovarian dimensions.

5. Conclusions

To conclude, we validated our approach of fully automated measurements of the uterus in MRIs.

Our VGG-16/11-based convolutional neural network is able to precisely locate the uterus and place measurement key points on it with excellent accuracy. From these points, the three-dimensional measurement of the uterus is obtained. The average difference in measurement between IA and radiologists remains inconsequential, even though it slightly exceeds the inter-radiologist variability.

This provides a useful and performant tool that can easily be applied in clinical practice as an alternative to time-consuming manual tracing.

Author Contributions: Conceptualization, D.M., E.P. and G.R.; methodology, D.M., E.P. and G.R.; validation, E.P., G.R., B.H., L.A.D.; formal analysis, L.F.; resources, D.M., G.R.; writing—original draft preparation, D.M., E.P., G.R.; writing—review and editing, D.M., E.P., G.R., N.L., C.H.; supervision, E.P., C.H. and N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Commission Ethique de la Recherche Clinique (CERCl) of Valenciennes Hospital, France (reference CHV-2022-006, date of approval 24 March 2022).

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

Data Availability Statement: All data was collected from the Institutional Picture Archiving and Communication System (EMR Manager, VEPRO AG, Version 8.2, Germany) of Valenciennes Hospital, France.

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

Abbreviations

AD	Average Deviation
AI	Artificial Intelligence
CNN	Convolutional Neural Network
DL	Deep Learning
DSC	Dice Similarity Coefficient
IoU	Intersection over Union
ML	Machine Learning
MR	Magnetic Resonance
OKS	Objective Key point Similarity
SD	Standard Deviation

References

1. Goldstuck, N. Assessment of uterine cavity size and shape: A systematic review addressing relevance to intrauterine procedures and events. *Afr. J. Reprod. Health* **2012**, *16*, 130–139.
2. Bridges, N.A.; Cooke, A.; Healy, M.J.; Hindmarsh, P.C.; Brook, C.G. Growth of the uterus. *Arch. Dis. Child.* **1996**, *75*, 330–331. [CrossRef]
3. Talarico, V.; Rodio, M.B.; Viscomi, A.; Galea, E.; Galati, M.C.; Raiola, G. The role of pelvic ultrasound for the diagnosis and management of central precocious puberty: An update. *Acta Biomed. Atenei Parm.* **2021**, *92*, e2021480.
4. Trotman, G.E. Delayed puberty in the female patient. *Curr. Opin. Obstet. Gynecol.* **2016**, *28*, 366. [CrossRef]
5. Adamou, H.; Amadou Magagi, I.; Oumarou Garba, S.; Habou, O. Acute intestinal obstruction due to extrinsic compression by previa myoma and ectopic pregnancy: A case report. *J. Med. Case Rep.* **2018**, *12*, 10. [CrossRef]
6. Weisberg, E. Insertion of Intrauterine Devices. *Med. J.* **1980**, *2*, 359–362. [CrossRef]
7. Chi, I.C.; Champion, C.B.; Wilkens, L.R. Cervical dilatation in interval insertion of an IUD Who requires it and does it lead to a high explulsion rate? *Contraception* **1987**, *36*, 403–415. [CrossRef] [PubMed]
8. Taylor, S.M.; Romero, A.A.; Kammerer-Doak, D.N.; Qualls, C.; Rogers, R.G. Abdominal hysterectomy for the enlarged myomatous uterus compared with vaginal hysterectomy with morcellation. *Am. J. Obstet. Gynecol.* **2003**, *189*, 1579–1582. [PubMed]
9. Iavazzo, C.; Salakos, N.; Bakalianou, K.; Vitoratos, N.; Vorgias, G.; Liapis, A. Thermal balloon endometrial ablation: A systematic review. *Arch. Gynecol. Obstet.* **2008**, *277*, 99–108. [CrossRef] [PubMed]
10. Acsinte, O.M.; Rabischong, B.; Bourdel, N.; Canis, M.; Botchorishvili, R. Laparoscopic Promontofixation in 10 Steps. *J. Minim. Invasive Gynecol.* **2018**, *25*, 767. [CrossRef] [PubMed]
11. Garfield, R.; Saade, G.; Buhimschi, C.; Buhimschi, I.; Shi, L.; Shi, S.; Chwalisz, K. Control and assessment of the uterus and cervix during pregnancy and labour. *Hum. Reprod. Update* **1998**, *4*, 673–695. [CrossRef] [PubMed]
12. Kojita, Y.; Matsuo, H.; Kanda, T.; Nishio, M.; Sofue, K.; Nogami, M.; Kono, A.K.; Hori, M.; Murakami, T. Deep learning model for predicting gestational age after the first trimester using fetal MRI. *Eur. Radiol.* **2021**, *31*, 3775–3782. [PubMed]
13. Egbase, P.E.; Al-Sharhan, M.; Grudzinskas, J.G. Influence of position and length of uterus on implantation and clinical pregnancy rates in IVF and embryo transfer treatment cycles. *Hum. Reprod.* **2000**, *15*, 1943–1946. [CrossRef]
14. Ludwin, A.; Martins, W.P. Correct measurement of uterine fundal internal indentation depth and angle: An important but overlooked issue for precise diagnosis of uterine anomalies. *Ultrasound Obstet. Gynecol.* **2021**, *58*, 497–499. [CrossRef] [PubMed]
15. Brouwer, C.L.; Steenbakkers, R.J.H.M.; van den Heuvel, E.; Duppen, J.C.; Navran, A.; Bijl, H.P.; Chouvalova, O.; Burlage, F.R.; Meertens, H.; Langendijk, J.A.; et al. 3D Variation in delineation of head and neck organs at risk. *Radiat. Oncol. Lond. Engl.* **2012**, *7*, 32. [CrossRef] [PubMed]
16. Cardenas, C.E.; Yang, J.; Anderson, B.M.; Court, L.E.; Brock, K.B. Advances in Auto-Segmentation. *Semin. Radiat. Oncol.* **2019**, *29*, 185–197. [CrossRef]
17. Kalantar, R.; Lin, G.; Winfield, J.M.; Messiou, C.; Lalondrelle, S.; Blackledge, M.D.; Koh, D.-M. Automatic Segmentation of Pelvic Cancers Using Deep Learning: State-of-the-Art Approaches and Challenges. *Diagnostics* **2021**, *11*, 1964. [CrossRef]
18. 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]
19. Han, S.; Hwang, S.I.; Lee, H.J. The Classification of Renal Cancer in 3-Phase CT Images Using a Deep Learning Method. *J. Digit. Imaging* **2019**, *32*, 638–643. [CrossRef]
20. 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. JASN* **2019**, *30*, 1514–1522. [CrossRef]
21. Sforazzini, F.; Salome, P.; Moustafa, M.; Zhou, C.; Schwager, C.; Rein, K.; Bougatf, N.; Kudak, A.; Woodruff, H.; Dubois, L.; et al. Deep Learning–based Automatic Lung Segmentation on Multiresolution CT Scans from Healthy and Fibrotic Lungs in Mice. *Radiol. Artif. Intell.* **2022**, *4*, e210095. [CrossRef] [PubMed]

22. Van Assen, M.; Muscogiuri, G.; Caruso, D.; Lee, S.J.; Laghi, A.; De Cecco, C.N. Artificial intelligence in cardiac radiology. *Radiol. Med. (Torino)* **2020**, *125*, 1186–1199. [CrossRef]
23. Theis, M.; Tonguc, T.; Savchenko, O.; Nowak, S.; Block, W.; Recker, F.; Essler, M.; Mustea, A.; Attenberger, U.; Marinova, M.; et al. Deep learning enables automated MRI-based estimation of uterine volume also in patients with uterine fibroids undergoing high-intensity focused ultrasound therapy. *Insights Imaging* **2023**, *14*, 1. [CrossRef]
24. Tsuboyama, T.; Onishi, H.; Nakamoto, A.; Ogawa, K.; Koyama, Y.; Tarewaki, H.; Tomiyama, N. Impact of Deep Learning Reconstruction Combined with a Sharpening Filter on Single-Shot Fast Spin-Echo T2-Weighted Magnetic Resonance Imaging of the Uterus. *Investig. Radiol.* **2022**, *57*, 379–386. [CrossRef] [PubMed]
25. Kurata, Y.; Nishio, M.; Kido, A.; Fujimoto, K.; Yakami, M.; Isoda, H.; Togashi, K. Automatic segmentation of the uterus on MRI using a convolutional neural network. *Comput. Biol. Med.* **2019**, *114*, 103438. [CrossRef]
26. Bhandary, S.; Kuhn, D.; Babaiee, Z.; Fechter, T.; Benndorf, M.; Zamboglou, C.; Grosu, R. Investigation and benchmarking of U-Nets on prostate segmentation tasks. *Comput. Med. Imaging Graph.* **2023**, *107*, 102241. [CrossRef] [PubMed]
27. Thimansson, E.; Bengtsson, J.; Baubeta, E.; Engman, J.; Flondell-Sité, D.; Bjartell, A.; Zackrisson, S. Deep learning algorithm performs similarly to radiologists in the assessment of prostate volume on MRI. *Eur. Radiol.* **2023**, *33*, 2519–2528. [CrossRef] [PubMed]
28. Ushinsky, A.; Bardis, M.; Glavis-Bloom, J.; Uchio, E.; Chantaduly, C.; Nguyentat, M.; Houshyar, R. A 3D-2D Hybrid U-Net Convolutional Neural Network Approach to Prostate Organ Segmentation of Multiparametric MRI. *Am. J. Roentgenol.* **2021**, *216*, 111–116. [CrossRef]
29. Krithika alias AnbuDevi, M.; Suganthi, K. Review of Semantic Segmentation of Medical Images Using Modified Architectures of UNET. *Diagnostics* **2022**, *12*, 3064. [CrossRef]
30. Gifford, R.; Jhawar, S.R.; Krening, S. Deep Learning Architecture to Improve Edge Accuracy of Auto-Contouring for Head and Neck Radiotherapy. *Diagnostics* **2023**, *13*, 2159. [CrossRef]

diagnostics

MDPI

Review

A Review of Recent Advances in Brain Tumor Diagnosis Based on AI-Based Classification

Reham Kaifi [1,2,3]

[1] Department of Radiological Sciences, College of Applied Medical Sciences, King Saud bin Abdulaziz University for Health Sciences, Jeddah City 22384, Saudi Arabia; kaifir@ksau-hs.edu.sa
[2] King Abdullah International Medical Research Center, Jeddah City 22384, Saudi Arabia
[3] Medical Imaging Department, Ministry of the National Guard—Health Affairs, Jeddah City 11426, Saudi Arabia

Abstract: Uncontrolled and fast cell proliferation is the cause of brain tumors. Early cancer detection is vitally important to save many lives. Brain tumors can be divided into several categories depending on the kind, place of origin, pace of development, and stage of progression; as a result, tumor classification is crucial for targeted therapy. Brain tumor segmentation aims to delineate accurately the areas of brain tumors. A specialist with a thorough understanding of brain illnesses is needed to manually identify the proper type of brain tumor. Additionally, processing many images takes time and is tiresome. Therefore, automatic segmentation and classification techniques are required to speed up and enhance the diagnosis of brain tumors. Tumors can be quickly and safely detected by brain scans using imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), and others. Machine learning (ML) and artificial intelligence (AI) have shown promise in developing algorithms that aid in automatic classification and segmentation utilizing various imaging modalities. The right segmentation method must be used to precisely classify patients with brain tumors to enhance diagnosis and treatment. This review describes multiple types of brain tumors, publicly accessible datasets, enhancement methods, segmentation, feature extraction, classification, machine learning techniques, deep learning, and learning through a transfer to study brain tumors. In this study, we attempted to synthesize brain cancer imaging modalities with automatically computer-assisted methodologies for brain cancer characterization in ML and DL frameworks. Finding the current problems with the engineering methodologies currently in use and predicting a future paradigm are other goals of this article.

Keywords: brain tumors; magnetic resonance imaging; computed tomography; computer-aided diagnostic and detection; deep learning; machine learning

Citation: Kaifi, R. A Review of Recent Advances in Brain Tumor Diagnosis Based on AI-Based Classification. *Diagnostics* **2023**, *13*, 3007. https://doi.org/10.3390/diagnostics13183007

Academic Editors: Dechang Chen, Wan Azani Mustafa and Hiam Alquran

Received: 23 June 2023
Revised: 14 September 2023
Accepted: 19 September 2023
Published: 20 September 2023

1. Introduction

The human brain, which serves as the control center for all the body's organs, is a highly developed organ that enables a person to adapt to and withstand various environmental situations [1]. The human brain allows people to express themselves in words, carry out activities, and express thoughts and feelings. Cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM) are the three major tissue components of the human brain. The gray matter regulates brain activity and comprises neurons and glial cells. The cerebral cortex is connected to other brain areas through white matter fibers comprising several myelinated axons. The corpus callosum, a substantial band of white matter fibers, connects the left and right hemispheres of the brain [2]. A brain tumor is a brain cell growth that is out of control and aberrant. Any unanticipated development may affect human functioning since the human skull is a rigid and volume-restricted structure, depending on the area of the brain involved. Additionally, it might spread to other organs, further jeopardizing human functions [3]. Early cancer detection makes the ability to plan effective treatment possible, which is crucial for the healthcare sector [4]. Cancer is difficult to cure,

and the odds of survival are significantly reduced if it spreads to nearby cells. Undoubtedly, many lives could be preserved if cancer was detected at its earliest stage using quick and affordable diagnostic methods. Both invasive and noninvasive approaches may be utilized to diagnose brain cancer. An incision is made during a biopsy to extract a lesion sample for analysis. It is regarded as the gold standard for the diagnosis of cancer, where pathologists examine several cell characteristics of the tumor specimen under a microscope to verify the malignancy.

Noninvasive techniques include physical inspections of the body and imaging modalities employed for imaging the brain [5]. In comparison to brain biopsy, other imaging modalities, such as CT scans and MRI images, are more rapid and secure. Radiologists use these imaging techniques to identify brain problems, evaluate the development of diseases, and plan surgeries [6]. However, brain scans or image interpretation to diagnose illnesses are prone to inter-reader variability and accuracy, which depends on the medical practitioner's competency [5]. It is crucial to accurately identify the type of brain disorder to reduce diagnostic errors. Utilizing computer-aided diagnostic (CAD) technologies can improve accuracy. The fundamental idea behind CAD is to offer a computer result as an additional guide to help radiologists interpret images and shorten the reading time for images. This enhances the accuracy and stability of radiological diagnosis [7]. Several CAT-based artificial intelligence techniques, such as machine learning (ML) and deep learning (DL), are described in this review for diagnosing tissues and segmenting tumors. The segmentation process is a crucial aspect of image processing. This approach includes a procedure for extracting the area that helps determine whether a region is infected. Using MRI images to segment brain tumors presents various challenges, including image noise, low contrast, loss borders, shifting intensities inside tissues, and tissue-type variation.

The most complex and crucial task in many medical image applications is detecting and segmenting brain tumors because it often requires much data and information. Tumors come in a variety of shapes and sizes. Automatic or semiautomatic detection/segmentation, helped by AI, is currently crucial in medical diagnostics. The medical professionals must authenticate the boundaries and areas of the brain cancer and ascertain where precisely it rests and the exact impacted locations before therapies such as chemotherapy, radiation, or brain surgery. This review examines the output from various algorithms that are used in segmenting and detecting brain tumors.

The review is structured as follows: Types of brain tumors are described in Section 2. The imaging modalities utilized in brain imaging are discussed in Section 3. The review algorithms used in the study are provided in Section 4. A review of the relevant state-of-the-art is provided in Section 5. The review is discussed in Section 6. The work's conclusion is presented in Section 7.

2. Types of Brain Tumors

The main three parts of the brain are the brain stem, cerebrum, and cerebellum [1]. The cerebellum is the second-largest component of the brain and manages bodily motor activities, including balance, posture, walking, and general coordination of movements. It is positioned behind the brain and connected to the brain stem. Internal white matter, tiny but deeply positioned volumes of gray matter, and a very thin gray matter outer cortex can all be found in the cerebellum and cerebrum. The brainstem links to the spinal cord. It is situated at the brain's base. Vital bodily processes, including motor, sensory, cardiac, repositories, and reflexes, are all under the control of the brainstem. Its three structural components are the medulla oblongata, pons, and midbrain [2]. A brain tumor is the medical term for an unexpected growth of brain cells [8]. According to the tumor's location, the kind of tissue involved, and whether they are malignant or benign, scientists have categorized several types of brain tumors based on the location of the origin (primary or secondary) and additional contributing elements [9]. The World Health Organization (WHO) categorized brain tumors into 120 kinds. This categorization is based on the cell's origin and behavior, ranging from less aggressive to greater aggressive. Even certain tumor

forms are rated, with grades I being the least malignant (e.g., meningiomas, pituitary tumors) and IV being the most malignant. Despite differences in grading systems that rely on the kind of tumor, this denotes the pace of growth [10]. The most frequent type of brain tumor in adults is glioma, which may be classified into HGG and LGG. The WHO further categorized LGG into I–II grade tumors and HGG into III–IV grade. To reduce diagnosing errors, accurate identification of the specific type of brain disorder is crucial for treatment planning. A summary of various types of brain tumors is provided in Table 1.

Table 1. Types of brain tumors.

Types of Tumors Based on	Type	Comment
Nature	Benign	Less aggressive and grows slowly
	Malignant	Life-threatening and rapidly expanding
Origin	Primary tumor	Originates in the brain directly
	Secondary tumor	This tumor develops in another area of the body like lung and breast before migrating to the brain
Grading	Grade I	Basically, regular in shape, and they develop slowly
	Grade II	Appear strange to the view and grow more slowly
	Grade III	These tumors grow more quickly than grade II cancers
	Grade IV	Reproduced with greater rate
Progression stage	Stage 0	Malignant but do not invade neighboring cells
	Stage 1	Malignant and quickly spreading
	Stage 2	
	Stage 3	
	Stage 4	The malignancy invades every part of the body

3. Imaging Modalities

For many years, the detection of brain abnormalities has involved the use of several medical imaging methods. The two brain imaging approaches are structural and functional scanning [11]. Different measurements relating to brain anatomy, tumor location, traumas, and other brain illnesses compose structural imaging [12]. The finer-scale metabolic alterations, lesions, and visualization of brain activity are all picked up by functional imaging methods. Techniques including CT, MRI, SPECT, positron emission tomography (PET), (FMRI), and ultrasound (US) are utilized to localize brain tumors for their size, location as well as shape, and other characteristics [13].

3.1. MRI

MRI is a noninvasive procedure that utilizes nonionizing, safe radiation [14] to display the 3D anatomical structure of any region of the body without the need for cutting the tissue. To acquire images, it employs RF pulses and an intense magnetic field [15].

The body is intended to be positioned within an intense magnetic field. The water molecules of the human body are initially in their equilibrium state when the magnets are off. The magnetic field is then activated by moving the magnets. The body's water molecules align with the magnetic field's direction under the effect of this powerful magnetic field [14]. Protons are stimulated to spin opposing the magnetic field and realign

by the application of a high RF energy pulse to the body in the magnetic field's direction. Protons are stimulated to spin opposing the magnetic field and realign by the application of a high RF energy pulse to the body in the magnetic field's direction. When the RF energy pulse is stopped, the water molecules return to their state of equilibrium and align with the magnetic field once more [14]. This causes the water molecules to produce RF energy, which the scanner detects and transforms into visual images [16]. The tissue structure determines the amount of RF energy the water molecules can use. As we can see in Figure 1, healthy brain has white matter (WM), gray matter (GM), and CSF, according to a structural MRI scan [17]. The primary difference between these tissues in a structural MRI scan is based on the amount of water they contain, with WM constituting 70% water and GM containing 80% water. The CSF fluid is almost entirely composed of water, as shown in Figure 1.

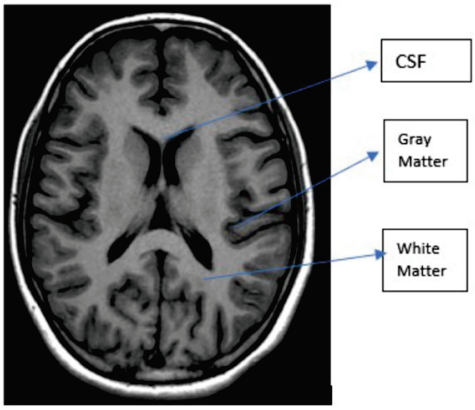


Figure 1. Healthy brain MRI image showing white matter (WM), gray matter (GM), and CSF [17].

Figure 2 illustrates the fundamental MRI planes used to visualize the anatomy of the brain: axial, coronal, and sagittal. Tl, T2, and FLAIR MRI sequences are most often employed for brain analysis [14]. A Tl-weighted scan can distinguish between gray and white matter. T2-weighted imaging is water-content sensitive and is therefore ideally suited to conditions where water accumulates within the tissues of the brain.

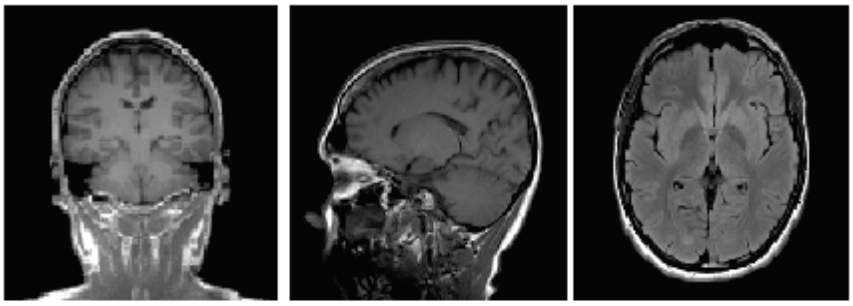


Figure 2. Fundamental MRI planes: (**a**) coronal, (**b**) sagittal, and (**c**) axial.

In pathology, FLAIR is utilized to differentiate between CSF and abnormalities in the brain. Gray-level intensity values in pixel spaces form an image during an MRI scan. The values of the gray-level intensity are dependent on the cell density. On T1 and T2 images of a tumor brain, the intensity level of the tumorous tissues differs [16]. The properties of various MRI sequences are shown in Table 2.

Table 2. Properties of various MRI sequences.

	T1	T2	Flair
White Matter	Bright	Dark	Dark
Gray Matter	Gray	Dark	Dark
CSF	Dark	Bright	Dark
Tumor	Dark	Bright	Bright

Most tumors show low or medium gray intensity on T1-w. On T2-w, most tumors exhibit bright intensity [17]. Examples of MRI tumor intensity level are shown in Figure 3.

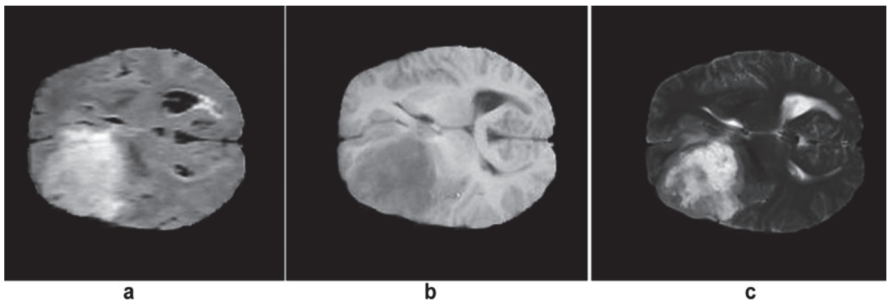


Figure 3. MRI brain tumor: (**a**) FLAIR image, (**b**) T1 image, and (**c**) T2 image [17].

Another type of MRI identified as functional magnetic resonance imaging (fMRI) [18] measures changes in blood oxygenation to interpret brain activity. An area of the brain that is more active begins to use more blood and oxygen. As a result, an fMRI correlates the location and mental process to map the continuing activity in the brain.

3.2. CT

CT scanners provide finely detailed images of the interior of the body using a revolving X-ray beam and a row of detectors. On a computer, specific algorithms are used to process the images captured from various angles to create cross-sectional images of the entire body [19]. However, a CT scan can offer more precise images of the skull, spine, and other bone structures close to a brain tumor, as shown in Figure 4. Patients typically receive contrast injections to highlight aberrant tissues. The patient may occasionally take dye to improve their image. When an MRI is unavailable, and the patient has an implantation like a pacemaker, a CT scan may be performed to diagnose a brain tumor. The benefits of using CT scanning are low cost, improved tissue classification detection, quick imaging, and more widespread availability. The radiation risk in a CT scan is 100 times greater than in a standard X-ray diagnosis [19].

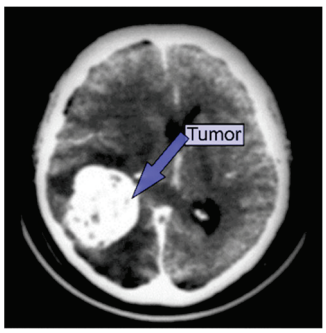


Figure 4. CT brain tumor.

3.3. PET

An example of a nuclear medicine technique that analyzes the metabolic activity of biological tissues is positron emission tomography (PET) [20]. Therefore, to help evaluate the tissue being studied, a small amount of a radioactive tracer is utilized throughout the procedure. Fluorodeoxyglucose (FDG) is a popular PET agent for imaging the brain. To provide more conclusive information on malignant (cancerous) tumors and other lesions, PET may also be utilized in conjunction with other diagnostic procedures like CT or MRI. PET scans an organ or tissue by utilizing a scanning device to find photons released by a radionuclide at that site [20]. The chemical compounds that are normally utilized by the specific organ or tissue throughout its metabolic process are combined with a radioactive atom to create the tracer used in PET scans, as shown in Figure 5.

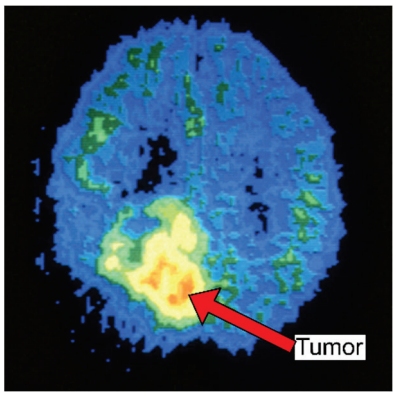


Figure 5. PET brain tumor.

3.4. SPECT

A nuclear imaging examination called a single-photon emission computed tomography (SPECT) combines CT with a radioactive tracer. The tracer is what enables medical professionals to observe the blood flow to tissues and organs [21]. A tracer is injected into the patient's bloodstream prior to the SPECT scan. The radiolabeled tracer generates gamma rays that the CT scanner can detect since it is radiolabeled. Gamma-ray information is gathered by the computer and shown on the CT cross-sections. A 3D representation of the brain can be created by adding these cross-sections back together [21].

3.5. Ultrasound

An ultrasound is a specialized imaging technique that provides details that can be useful in cancer diagnosis, especially for soft tissues. It is frequently employed as the initial step in the typical cancer diagnostic procedure [22]. One advantage of ultrasound is that a test can be completed swiftly and affordably without subjecting the patient to radiation. However, ultrasound cannot independently confirm a cancer diagnosis and is unable to generate images with the precise level of resolution or detail like a CT or MRI scan. A medical expert gently moves a transducer throughout the patient's skin across the region of the body being examined during a conventional ultrasound examination. A succession of high-frequency sounds is generated by the transducer, which "bounce off" the patient's interior organs. The ensuing echoes return to the ultrasound device, which then transforms the sound waves into a 2D image that may be observed in real-time on a monitor. According to [22], US probes have been applied in brain tumor resection. According to the degree of density inside the tissue being assessed, the shape and strength of ultrasonic echoes can change. An ultrasound can detect tumors that may be malignant because solid masses and fluid-filled cysts bounce sound waves differently.

4. Classification and Segmentation Method

As was stated in the introduction, brain tumors are a leading cause of death worldwide. Computer-aided detection and diagnosis refer to software that utilizes DL, ML, and computer vision for analyzing radiological and pathological images. It has been created to assist radiologists in diagnosing human disease in various body regions, including applications for brain tumors. This review explored different CAT-based artificial intelligence approaches, including ML and DL, for automatically classifying and segmenting tumors.

4.1. Classification Methods

A classification is an approach in which related datasets are grouped together according to common features. A classifier in classification is a model created for predicting the unique features of a class label. Predicting the desired class for each type of data is the fundamental goal of classification. Deep learning and machine learning techniques are used for the classification of medical images. The key distinction between the two types is the approach for obtaining the features used in the classification process.

4.1.1. Machine Learning

ML is a branch of AI that allows computers to learn without being explicitly programmed. Classifying medical images, including lesions, into various groups using input features has become one of the latest applications of ML. There are two types of ML algorithms: supervised learning and unsupervised learning [23]. ML algorithms learn from labeled data in supervised learning. Unsupervised learning is the process by which ML systems attempt to comprehend the interdata relationship using unlabeled data. ML has been employed to analyze brain cancers in the context of brain imaging [24]. The main stages of ML classification are image preprocessing, feature extraction, feature selection, and classification. Figure 6 illustrates the process architecture.

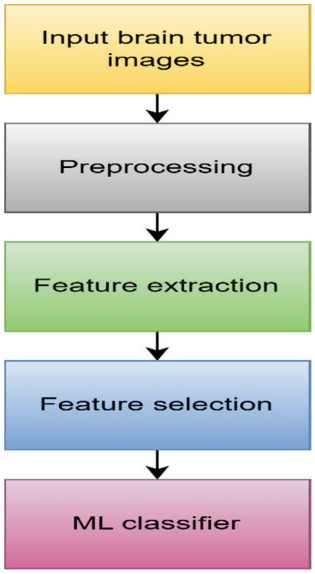


Figure 6. ML block diagram.

1. Data Acquisition

As previously noted, we can collect brain cancer images using several imaging modalities such as MRI, CT, and PET. This technique effectively visualizes aberrant brain tissues.

2. Preprocessing

Preprocessing is a very important stage in the medical field. Normally, noise enhancement or reduction in images occurs during preprocessing. Medical noise significantly reduces image quality, making them diagnostically inefficient. To properly classify medical

images, the preprocessing stage must be effective enough to eliminate as much noise as possible without affecting essential image components [25]. This procedure is carried out using a variety of approaches, including cropping, image scaling, histogram equalization, filtering using a median filter, and image adjusting [26].

3. Feature extraction

The process of converting images into features based on several image characteristics in the medical field is known as feature extraction. These features carry the same information as the original images but are entirely different. This technique has the advantages of enhancing classifier accuracy, decreasing overfitting risk, allowing users to analyze data, and speeding up training [27]. Texture, contrast, brightness, shape, gray level co-occurrence matrix (GLCM) [28], Gabor transforms [29], wavelet-based features [30], 3D Haralick features [31], and histogram of local binary patterns (LBP) [32] are some of the examples of the various types of features.

4. Feature selection

The technique attempts to arrange the features in ascending order of importance or relevance, with the top features being mostly employed in classification. As a result, multiple feature selection techniques are needed to reduce redundant information to discriminate between relevant and nonrelated features [33], such as PCA [34], genetic algorithm (GA) [35], and ICA [36].

5. ML algorithm

Machine learning aims to divide the input information into separate groups based on common features or patterns of behavior. KNN [35], ANN [37], RF [38], and SVM [39] are examples of supervised methods. These techniques include two stages: training and testing. During training, the data are manually labeled using human involvement. The model is first constructed in this step, after which it is utilized to determine the classes that are unlabeled in the testing stage. Application of the KNN algorithm works by finding the points that are closest to each other by computing the distance between them using one of several different approaches, including the Hamming, Manhatten, Euclidean, and Minkowski distances [35].

The support vector machine (SVM) technique is frequently employed for classification tasks. Every feature forming a data point in this approach, which represents a coordinate, is formed in a distinct n-space. As a result, the objective of the SVM method is to identify a boundary or line across a space with n dimensions, referred to as a hyperplane that separates classes [39]. There are numerous ways to create different hyperplanes, but the one with the maximum margin is the best. The maximum margin is the separation between the most extreme data points inside a class, often known as the support vectors.

4.1.2. Extreme Learning Machine (ELM)

Another new field that uses less computing than neural networks is evolutionary machine learning (EML). It is based on the real-time classification and regression technique known as the single-layer feed-forward neural network (SLFFNN). The input-to-hidden layer weights in the ELM are initialized randomly, whereas the hidden-to-output layer weights are trained to utilize the Moore–Penrose inverse method [40] to obtain a least-squares solution. As a result, classification accuracy is increased while net complexity, training time, and learning speed are all reduced.

Additionally, the hidden layer weights provide the network the capacity to multitask similar to other ML techniques such as KNN, SVM, and Bayesian networks [40]. As shown in Figure 7, the ELM network is composed of three levels, all of which are connected. Weights between the hidden and output layers can only vary, but the weights between the input and hidden layers are initially fixed at random and remain so during training.

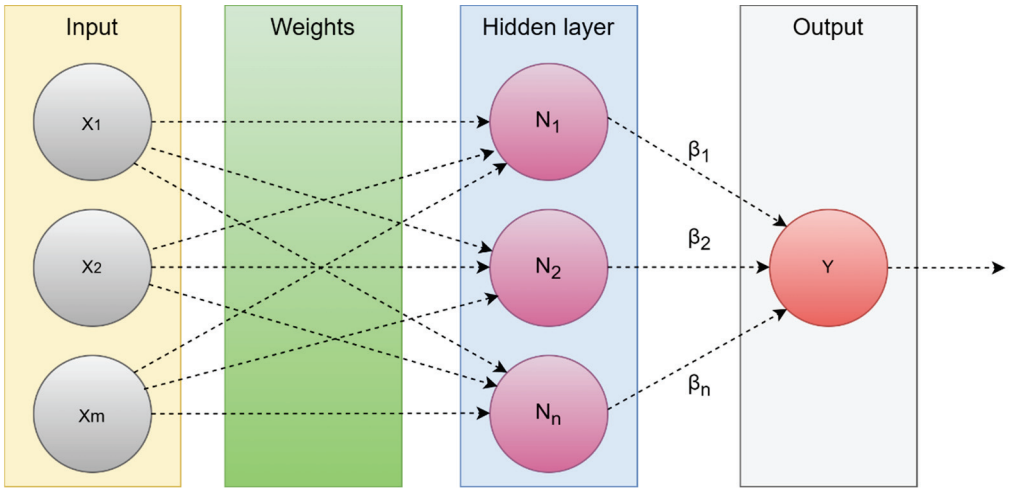


Figure 7. Extreme learning machine.

4.1.3. Deep Learning (DL)

Beginning a few years ago, deep learning, a branch of machine learning, has been utilized extensively to create automatic, semiautomatic, and hybrid models that can accurately detect and segment tumors in the shortest period possible [41]. DL can learn the features that are significant for a problem by utilizing a training corpus with sufficient diversity and quality. Deep learning [42] has achieved excellent success in tackling the issues of ML by combining the feature extraction and selection phases into the training process [43]. Deep learning is motivated by the comprehension of neural networks that exist within the human brain. DL models are often represented as a sequence of layers generated by a weighted sum of information from the previous layer. The data are represented by the first layer, while the output is represented by the last layer [44]. Deep learning models can tackle extremely difficult problems while often requiring less human interaction than conventional ML techniques because several layers make it possible to duplicate complex mapping functions.

The most common DL model used for the categorization and segmentation of images is a convolution neural network (CNN). In a hierarchical manner, CNN analyzes the spatial relationship of pixels. Convoluting the images with learned filters creates a hierarchy of feature maps, which is how this is accomplished. This convolution function is performed in several layers such that the features are translation- and distortion-invariant and hence accurate to a high degree [45]. Figure 8 illustrates the main process in DL.

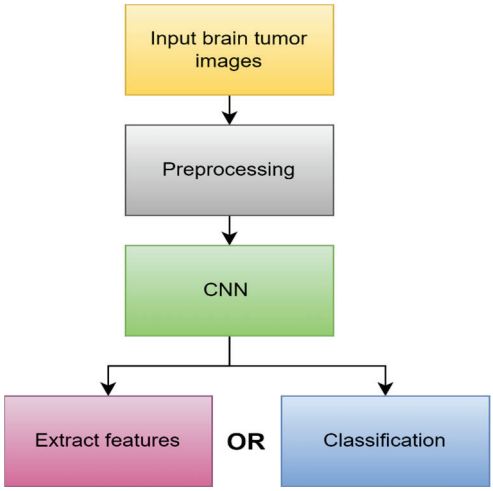


Figure 8. DL block diagram.

Preprocessing is primarily used to eliminate unnecessary variation from the input image and make training the model easier. More actions are required to extend beyond neural network models' limits, such as resizing normalization. All images must be resized before being entered into CNN classification models since DL requires inputs of a constant size [46]. Images that are greater than the desired size can be reduced by downscaling, interpolation, or cutting the background pixels [46].

Many images are required for CNN-based classification. Data augmentation is one of the most important data strategies for addressing issues with unequal distribution and data paucity [47].

CNN's architecture is composed of three primary layers: convolutional, pooling, and fully connected. The first layer is the main layer that is able to extract image features such as edges and boundaries. Based on the desired prediction results, this layer may automatically learn many filters in parallel for the training dataset. The first layer creates features, but the second layer oversees data reduction, which minimizes the size of those features and reduces the demand for computing resources. Every neuron in the final layer, which is a completely connected layer, is coupled to every neuron in the first layer. The layer serves as a classifier to classify the acquired feature vector of previous layers [48,49]. The approach that CNN uses is similar to how various neural networks work: it continually modifies its weights by taking an error from the output and inserting it as output to improve filters and weights. In addition, CNN standardizes the output utilizing a SoftMax function [50]. Many types of CNN architecture exist, including ResNet, AlexNet, and cascade-CNN, among others [51].

4.2. Segmentation Method

Brain tumor segmentation, which has been employed in some research, is an important step in improving disease diagnosis, evaluation, treatment plans, and clinical trials. The purpose of segmentation in tumor classification is to detect the tumor location from brain scans, improve representation, and allow quantitative evaluations of image structures during the feature extraction step [52]. Brain tumor segmentation can be accomplished in two ways: manually and completely automatically [53].

Manual tumor segmentation from brain scans is a difficult and time-consuming procedure. Furthermore, the artifacts created during the imaging procedure result in poor-quality images that are difficult to analyze. Additionally, due to uneven lesions, geographical flexibility, and unclear borders, manual detection of brain tumors is challenging. This section discusses several automated brain tumor segmentation strategies to help radiologists overcome these issues.

4.2.1. Region-Based Segmentation

A region in an image is a collection of related pixels that comply with specific homogeneity requirements, such as shape, texture, and pixel intensity values [54]. In a region-based segmentation, the image is divided into disparate areas to precisely identify the target region [55]. When grouping pixels together, the region-based segmentation takes into consideration the pixel values, such as gray-level variance and difference, as well as their spatial closeness, such as the Euclidean distance or region density. K-means clustering [56] and FCM [56] are the most techniques used in this method.

4.2.2. Thresholding Methods

The thresholding approach is a straightforward and effective way to separate the necessary region [57], but finding an optimum threshold in low-contrast images may be challenging.

Based on picture intensity, threshold values are chosen using histogram analysis [58]. There are two types of thresholding techniques: local and global. The global thresholding approach is the best choice for segmentation if the objects and the background have highly uniform brightness or intensity. The Gaussian distribution approach may be used to

obtain the ideal threshold value [59]. Otsu thresholding [38] is the popular method among these techniques.

4.2.3. Watershed Techniques

The intensities of the image are analyzed using watershed techniques [60]. Topological watershed [61], marker-based watershed [62], and image IFT watershed [63] are a few examples of watershed algorithms.

4.2.4. Morphological-Based Method

The morphology technique relies on the morphology of image features. It is mostly used for extracting details from images based on shape representation. Dougherty [64] defines dilation and erosion as two basic operations. Dilation is used to increase the size of an image. Erosion reduces the size of images.

4.2.5. Edge-Based Method

Edge detection is performed using variations in image intensity. Pixels at an edge are those where the image's function abruptly changes. Edge-based segmentation techniques include those by Sobel, Roberts, Prewitt, and Canny [65]. Reference [66] offers an enhanced edge detection approach for tumor segmentation. The development of an automated image-dependent thresholding is combined with the Sobel operator to identify the edges of the brain tumor.

4.2.6. Neural-Networks-Based Method

Neuronal network-based segmentation techniques employ computer models of artificial neural networks consisting of weighted connections between processing units (called neurons). At the connections, the weights act as multipliers. To acquire the coefficient values, training is necessary. The segmentation of medical images and other fields has made use of a variety of neural network designs. Some of the techniques utilized in the segmentation process include the multilayer perceptron (MLP), Hopfield neural networks (HNN) [67], back-propagation learning algorithm, SVM-based segmentation [68], and self-organizing maps (SOM) neural network [67].

4.2.7. DL-Based Segmentation

The primary strategy used in the DL-based segmentation of brain tumors technique is to pass an image through a series of deep learning structures before performing input image segmentation based on the deep features [69]. Many deep learning methods, such as deep CNNs, CNN, and others, have been suggested for segmenting brain tumors.

A deep learning system called semantic segmentation [70] arranges pixels in an image according to semantic categories. The objective is to create a dense pixel-by-pixel segmentation map of the image, and each pixel is given an assigned category or entity.

4.3. Performance Evaluation

An important component of every research work involves evaluating the classification and segmentation performance. The primary goal of this evaluation is to measure and analyze the model's capability for segmentation or diagnostic purposes. Segmentation is a crucial step in improving the diagnostic process, as we mentioned before, but for this to occur, the segmentation process must be as accurate as feasible. Additionally, to evaluate the diagnostic approach utilized while taking complexity and time into account [71].

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the main four elements in any analysis or to evaluate any segmentation or classification algorithm. A pixel that is accurately predicted to be assigned to the specified class in a segmentation method is represented by TP and TN based on the ground truth. Furthermore, FP is a result when the model predicts a pixel wrongly as not belonging to a specific class. A

false negative (FN) results when the model wrongly predicts a pixel belonging to a certain class [71].

TP in classification tasks refers to an image that is accurately categorized into a positive category based on the ground truth. Similar to this, the TN result occurs when the model properly classifies an image in the negative category. As opposed to that, FP results occur when the model wrongly assigns an image in the positive class while the actual datum is in the negative category. FN results occur when the model misclassifies an image while it belongs in the positive category. Through the four elements mentioned above, different performance measures enable us to expand the analysis.

Accuracy (ACC) measures a model's ability to correctly categorize all pixels/classes, whether they are positive or negative. Sensitivity (SEN) shows the percentage of accurately predicted positive images/pixels among all actual positive samples. It evaluates a model's ability to recognize relevant samples or pixels. The percentage of actual negatives that were predicted is known as specificity (SPE). It indicates a percentage of classes or pixels that could not be accurately recognized [71].

The precision (PR) or positive predictive value (PPV) measures how frequently the model correctly predicts the class or pixel. It provides the precise percentage of positively expected results from models. The most often used statistic that combines SEN and precision is the F1 score [72]. It refers to the two-dimensional harmonic mean.

The Jaccard index (JI), also known as intersection over union (IoU), calculates the percentage of overlap between the model's prediction output and the annotation ground-truth mask.

The spatial overlap between the segmented region of the model and the ground-truth tumor region is measured by the Dice similarity coefficient (DSC). A DSC value of zero means there is no spatial overlap between the annotated model result and the actual tumor location, whereas a value of one means there is complete spatial overlap. The receiver characteristics curve is summarized by the area under the curve (AUC), which compares SEN to the false positive rate as a measure of a classifier's ability to discriminate between classes.

The similarity between the segmentation produced by the model and the expert-annotated ground truth is known as the similarity index (SI). It describes how the identification of the tumor region is comparable to that of the input image [71]. Table 3 summarizes different performance equations.

Table 3. Performance equation.

Parameter	Equation
ACC	$(TP + TN)/(TP + FN + FP + TN)$
SEN	$TP/(TP + FN)$
SPE	$TN/(TN + FP)$
PR	$TP/(TP + FP)$
F1_SCORE	$2 * PR * SEN/(PR + SEN)$
DCS	$2 * TP/(2 * TP + FP + FN$
Jaccard	$TP/(TP + FP + FN)$

5. Literature Review

5.1. Article Selection

The major goal of this study is to review and understand brain tumor classification and detection strategies developed worldwide between 2010 and 2023. This present study aims to review the most popular techniques for detecting brain cancer that have been made available globally, in addition to looking at how successful CAD systems are in this process.

We did not target any one publisher specifically, but we utilized articles from a variety of sources to account for the diversity of knowledge in a particular field. We collected

appropriate articles from several internet scientific research article libraries. We searched the pertinent publications using IEEE Explore, Medline, ScienceDirect, Google Scholar, and ResearchGate.

Each time, the filter choice for the year (2010 to 2023) was chosen so that only papers from the chosen period were presented. Most frequently, we used terms like "detection of MRI images using deep learning," "classification of brain tumor from CT/MRI images using deep learning," "detection and classification of brain tumor using deep learning," "CT brain tumor," "PET brain tumor," etc. This study offers an analysis of 53 chosen publications.

5.2. Publicly Available Datasets

The researchers tested the proposed methods on several publicly accessible datasets. In this part, several significant and difficult datasets are covered. The most difficult MRI datasets are BRATS. Table 4 presents a summary of the dataset names.

Table 4. Summary of the dataset.

Dataset	MRI Sequences	Source
BRATS	T1, T2, FLAIR	[73]
RIDER	T1, T2, FLAIR	[74]
Harvard	T2	[75]
TCGA	T1, T2, FLAIR	[76,77]
Figshare	T1	[78]
IXI	T1, T2	[79]

5.3. Related Work

In addition to the several techniques for segmenting brain tumors that we already highlighted, this section presents a summary of studies that use artificial intelligence to classify brain tumors.

5.3.1. MRI Brain Tumor Segmentation

This section describes the various machine learning, deep learning, region growth, thresholding, and literature-proposed brain tumor segmentation strategies.

To segment brain tumors, Gordillo et al. [80] utilized fuzzy logic structure, which they built utilizing features extracted from MR images and expert knowledge. This system learns unsupervised and is fully automated. With trials conducted on two different forms of brain tumors, glioblastoma multiform and meningioma, the result of segmentation using this approach is shown to be satisfactory, with the lowest accuracy of 71% and a maximum of 93%.

Employing fuzzy c-means clustering on MRI, Rajendran [81] presented logic analyzing for segmenting brain tumors. The region-based technique that iteratively progresses toward the ultimate tumor border was initialized using the tumor type output of fuzzy clustering. Using 15 MR images with manual segmentation ground truth available, tests were conducted on this approach to determine its effectiveness. The overall result was suitable, with a sensitivity of 96.37% and an average Jaccard coefficient value of 83.19%.

An SVM classifier was applied by Kishore et al. to categorize tumor pixels using vectors of features from MR images, such as mean intensity and LBP. Level sets and region-growing techniques were used for the segmentation. The experiments on their suggested methods used MR images with tumor regions manually defined by 11 different participants. Their suggested methods are effective, with a DSC score of 0.69 [82].

A framework for segmenting tumorous MRI 3D images was presented by Abbasi and Tajeripour [38]. The first phase improves the input image's contrast using bias field correction. The data capacity is reduced using the multilevel Otsu technique in the second

phase. LBP in three orthogonal planes and an enhanced histogram of images are employed in the third stage, the feature extraction step. Lastly, the random forest is employed as a classifier for distinguishing tumorous areas since it can work flawlessly with large inputs and has a high level of segmentation accuracy. The overall outcome was acceptable, with a mean Jaccard value of 87% and a DSC of 93%.

By combining two K-means and FCM-clustering approaches, Almahfud et al. [83] suggest a technique for segmenting human brain MRI images to identify brain cancers. Because K-means is more susceptible to color variations, it can rapidly and effectively discover optima and local outliers. So that the cluster results are better and the calculation procedure is simpler, the K-means results are clustered once more with FCM to categorize the convex contour based on the border. To increase accuracy, morphology and noise reduction procedures are also suggested. Sixty-two brain MRI scans were used in the study, and the accuracy rate was 91.94%.

According to Pereira et al. [69], an automated segmentation technique based on CNN architecture was proposed, which explores small three-by-three kernels. Given the smaller number of weights in the network, using small kernels enables the creation of more intricate architectures and helps prevent overfitting. Additionally, they looked at the use of intensity normalizing as an initial processing step, which, when combined with data augmentation, was highly successful in segmenting brain tumors in MRI images. Their suggestion was verified using the BRATS database, yielding Dice similarity coefficient values of 0.88, 0.83, and 0.77 for the Challenge dataset for the whole, core, and enhancing areas.

According to the properties of a separated local square, they suggested a unique approach for segmenting brain tumors [84]. The suggested procedure essentially consists of three parts. An image was divided into homogenous sections with roughly comparable properties and sizes using the super-pixel segmentation technique in the first stage. The second phase was the extraction of gray statistical features and textural information. In the last phase of building the segmentation model, super-pixels were identified as either tumor areas or nontumor regions using SVM. They used 20 images from the BRATS dataset, where a DSC of 86.12% was attained, to test the suggested technique.

The CAD system suggested by Gupta et al. [85] offers a noninvasive method for the accurate tumor segmentation and detection of gliomas. The system takes advantage of the super pixels' combined properties and the FCM-clustering technique. The suggested CAD method recorded 98% accuracy for glioma detection in both low-grade and high-grade tumors.

Brain tumor segmentation using the CNN-based data transfer to SVM classifier approach was proposed by Cui et al. [68]. Two cascaded phases comprise their algorithm. They trained CNN in the initial step to understand the mapping of the image region to the tumor label region. In the testing phase, they passed the testing image and CNN's anticipated label output to an SVM classifier for precise segmentation. Tests and evaluations show that the suggested structure outperforms separate SVM-based or CNN-based segmentation, while DSC achieved 86.12%.

The two-pathway-group CNN architecture described by Razzak et al. is a novel approach for brain tumor segmentation that simultaneously takes advantage of local and global contextual traits. This approach imposes equivariance in the 2PG-CNN model to prevent instability and overfitting parameter sharing. The output of a basic CNN is handled as an extra source and combined at the last layer of the 2PG CNN, where the cascade architecture was included. When a group CNN was embedded into a two-route architecture for model validation using BRATS datasets, the results were DSC 89.2%, PR 88.22%, and SEN 88.32% [86].

A semantic segmentation model for the segmentation of brain tumors from multimodal 3D MRIs for the BRATS dataset was published in [87]. After experimenting with several normalizing techniques, they discovered that group-norm and instance-norm performed equally well. Additionally, they have tested with more advanced methods of data augmentation, such as random histogram pairing, linear image transformations, rotations,

and random image filtering, but these have yet to show any significant benefit. Further, raising the network depth had no positive effect on performance. However, increasing the number of filters consistently produced better results. Their BRATS end testing dataset values were 0.826, 0.882, and 0.837 for overall Dice coefficient or improved tumor core, entire tumor, and tumor center, respectively.

CNN was used by Karayegen and Aksahin [88] to offer a semantic segmentation approach for autonomously segmenting brain tumors on BRATS image datasets that include images from four distinct imaging modalities (T1, T1C, T2, and FLAIR). This technique was effectively used, and images were shown in a variety of planes, including sagittal, coronal, and axial, to determine the precise tumor location and parameters such as height, breadth, and depth. In terms of tumor prediction, evaluation findings of semantic segmentation carried out using networks are incredibly encouraging. The mean IoU and mean prediction ratio were both calculated to be 86.946 and 91.718, respectively.

A novel, completely automatic method for segmenting brain tumor regions was proposed by Ullah et al. [89] using multiscale residual attention CNN (MRA-UNet). To maintain the sequential information, MRA-UNet uses three sequential slices as its input. By employing multiscale learning in a cascade path, it can make use of the adaptable region of interest strategy and precisely segment improved and core tumor regions. In the BRATS-2020 dataset, their method produced novel outcomes with an overall Dice score of 90.18%.

A new technique for segmenting brain tumors using the fuzzy Otsu thresholding morphology (FOTM) approach was presented by Wisaeng and Sa-Ngiamvibool [90]. The values from each single histogram in the original MRI image were modified by using a color normalizing preprocessing method in conjunction with histogram specification. The findings unambiguously demonstrate that image gliomas, image meningiomas, and image pituitary have average accuracy indices of 93.77%, 94.32%, and 94.37%, respectively. A summary of MRI brain tumor segmentation is provided in Table 5.

Table 5. MRI brain tumor segmentation.

Ref.	Scan	Year	Technique	Method	Performance Metrics	Result
[80]	MRI	2010	region-based	FCM	Acc	93.00%
[81]	MRI	2011	region-based	FCM	Jaccard	83.19%
[82]	MRI	2012	NN	LBP with SVM	DSC	69.00%
[69]	MRI	2016	DL	CNN	DSC	88.00%
[84]	MRI	2017	NN	GLCM with SVM	DSC	86.12%
[38]	MRI	2017	NN	LBP with RF	Jaccard and DSC	87% and 93%
[85]	MRI	2018	region-based	FCM	Acc	98.00%
[83]	MRI	2018	region-based	FCM and k-mean	Acc	91.94%
[68]	MRI	2019	DL and NN	CNN with SVM	DSC	88.00%
[86]	MRI	2019	DL	Two-path CNN	DSC	89.20%
[87]	MRI	2019	DL	semantic	Acc	88.20%
[88]	MRI	2021	DL	semantic	IoU	91.72%
[89]	MRI	2022	DL	MRA-UNet	DSC	98.18%
[90]	MRI	2023	region-based	Fuzzy Otsu Threshold	Acc	94.37%

5.3.2. MRI Brain Tumor Classification Using ML

The automated classification of brain cancers using MRI images has been the subject of several studies. Cleaning data, feature extraction, and feature selection are the basic steps in the machine learning (ML) process that have been used for this purpose. Building an ML model based on labeled samples is the last step. A summary of MRI brain tumor classification using ML is provided in Table 6.

An NN-based technique to categorize a given MR brain image as either normal or abnormal is presented in [91]. In this method, features were first extracted from images using the wavelet transform, and then the dimensionality of the features was reduced using PCA methodology. The reduced features were routed to a back-propagation NN that uses a scaled conjugate gradient (SCG) to determine the best weights for the NN. This technique was used on 66 images, 18 of which were normal and 48 abnormal. On training and test images, the classification accuracy was 100%.

An automated and efficient CAD method based on ensemble classifiers was proposed by Arakeri and Reddy [36] for the classification of brain cancers on MRI images as benign or malignant. A tumor's texture, shape, and border properties were extracted and used as a representation. The ICA approach was used to select the most significant features. The ensemble classifier, consisting of SVM, ANN, and kNN classifiers, is trained using these features to describe the tumor. A dataset consisting of 550 patients' T1- and T2-weighted MR images was used for the experiments. With an accuracy of 99.09% (sensitivity 100% and specificity 98.21%), the experimental findings demonstrated that the suggested classification approach achieves strong agreement with the combined classifier and is extremely successful in the identification of brain tumors. Figure 9 illustrates the CAD method based on ensemble classifiers.

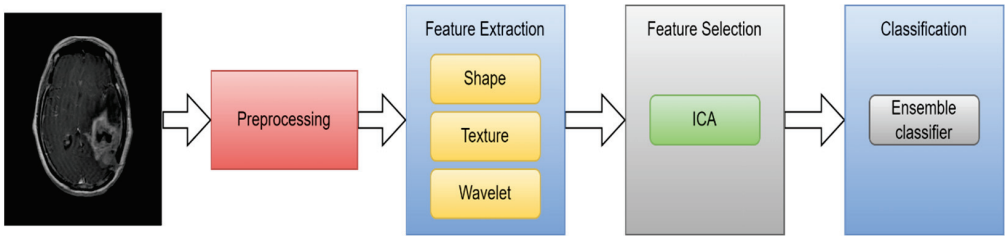


Figure 9. CAD method based on ensemble classifiers.

In [92], the authors suggested a novel, wavelet-energy-based method for automatically classifying MR images of the human brain into normal or abnormal. The classifier was SVM, and biogeography-based optimization (BBO) was utilized to enhance the SVM's weights. They succeeded in achieving 99% precision and 97% accuracy.

Amin et al. [28] suggest an automated technique to distinguish between malignant and benign brain MRI images. The segmentation of potential lesions has used a variety of methodologies. Then, considering shape, texture, and intensity, a feature set was selected for every candidate lesion. The SVM classifier is then used on the collection of features to compare the proposed framework's precision using various cross-validations. Three benchmark datasets, including Harvard, Rider, and Local, are used to verify the suggested technique. For the procedure, the average accuracy was 97.1%, the area under the curve was 0.98, the sensitivity was 91.9%, and the specificity was 98.0%.

A suitable CAD approach toward classifying brain tumors is proposed in [93]. The database includes meningioma, astrocytoma, normal brain areas, and primary brain tumors. The radiologists selected 20 × 20 regions of interest (ROIs) for every image in the dataset. Altogether, these ROI(s) were used to extract 371 intensity and texture features. These three classes were divided using the ANN classifier. Overall classification accuracy was 92.43%.

Four hundred twenty-eight T1 MR images from 55 individuals were used in a varied dataset for multiclass brain tumor classification [94]. A based-on content active contour model extracted 856 ROIs. These ROIs were used to extract 218 intensity and texture

features. PCA was employed in this study to reduce the size of the feature space. The ANN was then used to classify these six categories. The classification accuracy was seen to have reached 85.5%.

A unique strategy for classifying brain tumors in MRI images was proposed in [95] by employing improved structural descriptors and hybrid kernel-SVM. To better classify the image and improve the texture feature extraction process using statistical parameters, they used GLCM and histograms to derive the texture feature from every region. Different kernels were combined to create a hybrid kernel SVM classifier to enhance the classification process. They applied this technique to only axial T1 brain MRI images—93% accuracy for their suggested strategy.

A hybrid system composed of two ML techniques was suggested in [96] for classifying brain tumors. For this, 70 brain MR images overall (60 abnormal, 10 normal) were taken into consideration. DWT was used to extract features from the images. Using PCA, the total number of features was decreased. Following feature extraction, feed-forward back-propagation ANN and KNN were applied individually on the decreased features. The back-propagation learning method for updating weights is covered by FP-ANN. KNN has already been covered. Using KNN and FP-ANN, this technique achieves 97% and 98% accuracy, respectively [96].

A strategy for classifying brain MRI images was presented in [97]. Initially, they used an enhanced image improvement method that comprises two distinct steps: noise removal and contrast enhancement using histogram equalization. Then, using a DWT to extract features from an improved MR brain image, they further decreased these features by mean and standard deviation. Finally, they developed a sophisticated deep neural network (DNN) to classify the brain MRI images as abnormal or normal, and their strategy achieved 95.8%.

Table 6. MRI brain tumor classification using ML.

Ref.	Scan	Year	Feature Extraction	Feature Selection	Classification	Acc.
[96]	MRI	2010	GLCM	PCA	ANN and KNN	98% and 97%
[91]	MRI	2011	Wavelet	PCA	Back-propagation NN	100.00%
[94]	MRI	2013	Intensity and texture	PCA	ANN	85.50%
[95]	MRI	2014	GLCM	-	SVM	93.00%
[36]	MRI	2015	Texture and shape	ICA	SVM	99.09%
[92]	MRI	2015	Wavelet	-	SVM	97.00%
[28]	MRI	2017	Texture and shape	-	SVM	97.10%
[93]	MRI	2017	Intensity and texture	-	ANN	92.43%
[97]	MRI	2020	DWT	Mean and standard deviation	DNN	95.8%

5.3.3. MRI Brain Tumor Classification Using DL

Difficulties remain in categorizing brain cancers from an MRI scan, despite encouraging developments in the field of ML algorithms for the classification of brain tumors into their different types. These difficulties are mostly the result of the ROI detection; typical labor-intensive feature extraction methods could be more effective [98]. Owing to the nature of deep learning, the categorization of brain tumors is now a data-driven problem rather than a challenge based on manually created features [99]. CNN is one of the deep learning models that is frequently utilized in brain tumor classification tasks and has produced a significant result [100].

According to a study [101], the CNN algorithm can be used to divide the severity of gliomas into two categories: low severity or high severity, as well as multiple grades of severity (Grades II, III, and IV). Accuracy rates of 71% and 96% were reached by the classifier.

A DL approach based on a CNN was proposed by Sultan et al. [7] to classify different kinds of brain tumors using two publicly available datasets. The proposed method's block diagram is presented in Figure 10. The first divides cancers into meningioma, pituitary, and glioma tumors. The other one distinguishes among Grade II, III, and IV gliomas. The first and second datasets, which each have 233 and 73 patients, contain a combined total of 3064 and 516 T1 images. The suggested network configuration achieves the best overall accuracy, 96.13% and 98.7%, for the two studies, which results in significant performance [7].

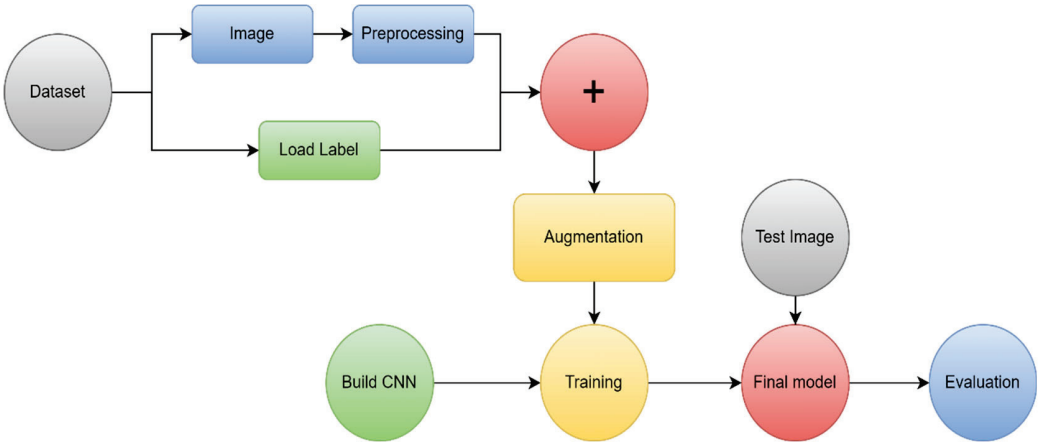


Figure 10. A block schematic showing the suggested approach. Reprinted (adapted) with permission from [7]. Copyright 2019 IEEE.

Similarly, ref. [102] showed how to classify brain MRI scan images into malignant and benign using CNN algorithms in conjunction with augmenting data and image processing. They evaluated the effectiveness of their CNN model with pretrained VGG-16, Inception-v3, and ResNet-50 models using the transfer learning methodology. Even though the experiment was carried out on a relatively small dataset, the results reveal that the model's accuracy result is quite strong and has a very low complexity rate, as it obtained 100% accuracy, compared to VGG-16's 96%, ResNet-50's 89%, and Inception-V3's 75%. The structure of the suggested CNN architecture is shown in Figure 11.

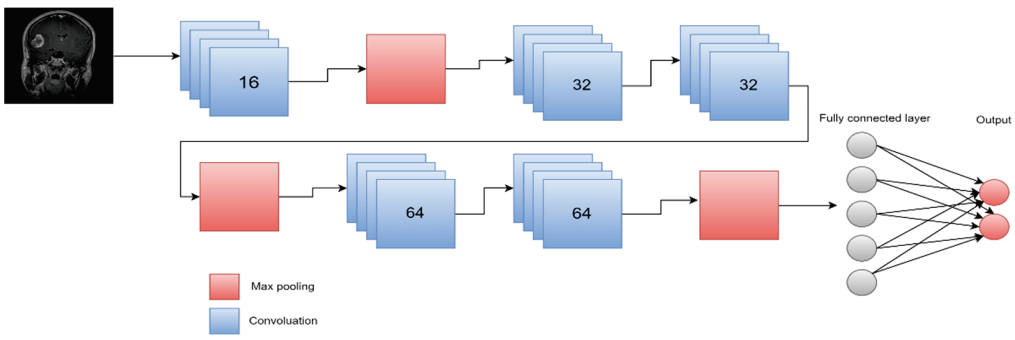


Figure 11. Proposed method. Reprinted (adapted) with permission from [102]. Copyright 2020 Mathematical Biosciences and Engineering.

For accurate glioma grade prediction, researchers developed a customized CNN-based deep learning model [103] and evaluated the performance using AlexNet, GoogleNet, and SqueezeNet by transfer learning. Based on 104 clinical glioma patients with (50 LGGs

and 54 HGGs), they trained and evaluated the models. The training data was expanded using a variety of data augmentation methods. A five-fold cross-validation procedure was used to assess each model's performance. According to the study's findings, their specially created deep CNN model outperformed the pretrained models by an equal or greater percentage. The custom model's accuracy, sensitivity, F1 score, specificity, and AUC values were, respectively, 0.971, 0.980, 0.970, 0.963, and 0.989.

A novel transfer learning-based active learning paradigm for classifying brain tumors was proposed by Ruqian et al. [104]. Figure 12 describes the workflow for active learning. On the MRI training dataset of 203 patients and the baseline validation dataset of 66 patients, they used a 2D slice-based technique to train and fine-tune the model. Their suggested approach allowed the model to obtain an area under the curve (ROC) of 82.89%. The researchers built a balanced dataset and ran the same process on it to further investigate the robustness of their strategy. Compared to the baseline's AUC of 78.48%, the model's AUC was 82%.

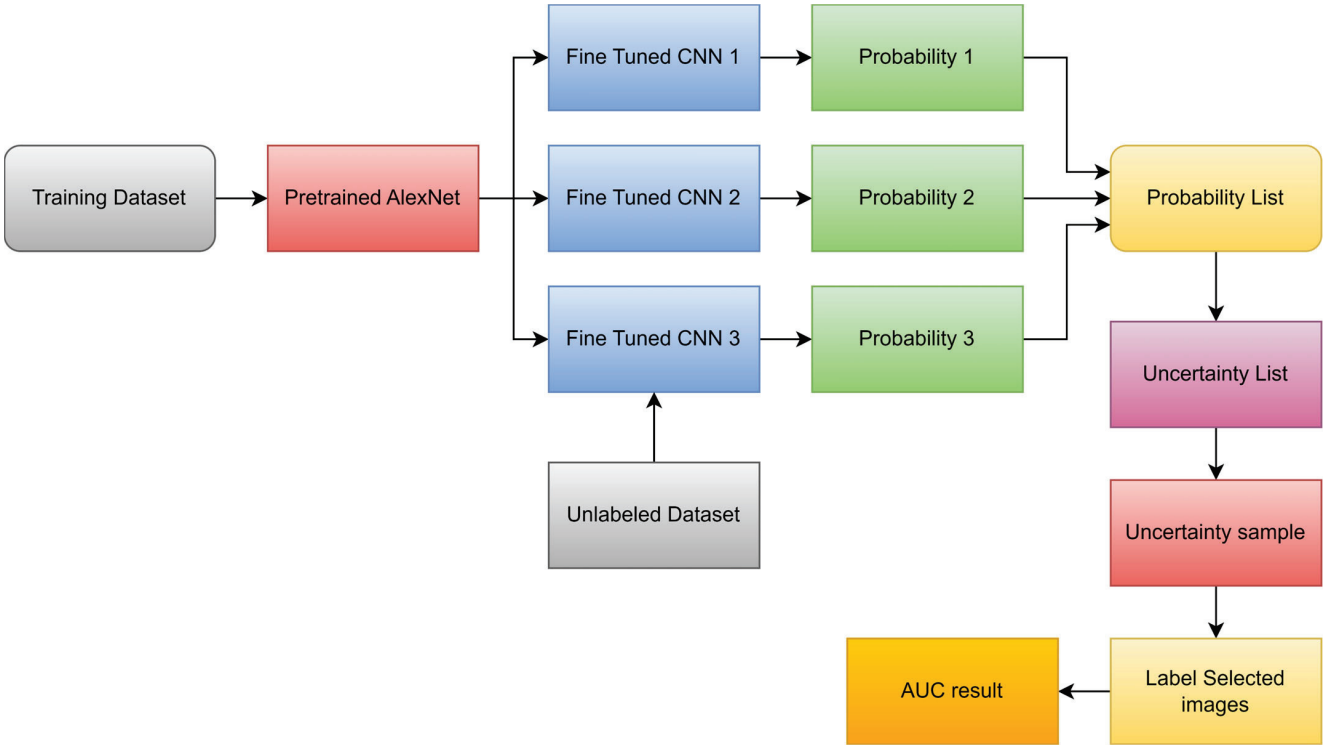


Figure 12. Workflow of the suggested active learning framework based on transfer learning. Reprinted (adapted) with permission from [104]. Copyright 2021 Frontiers in Artificial Intelligence.

A total of 131 patients with glioma were enrolled [105]. A rectangular ROI was used to segment tumor images, and this ROI contained around 80% of the tumor. The test dataset was then created by randomly selecting 20% of the patient-level data. Models previously trained on the expansive natural image database ImageNet were applied to MRI images, and then AlexNet and GoogleNet were developed from scratch and fine-tuned. Five-fold cross-validation (CV) was used on the patient-level split to evaluate the classification task. The averaged performance metrics for validation accuracy, test accuracy, and test AUC from the five-fold CV of GoogleNet were, respectively, 0.867, 0.909, and 0.939.

Hamdaoui et al. [106] proposed an intelligent medical decision-support system for identifying and categorizing brain tumors using images from the risk of malignancy index. They employed deep transfer learning principles to avoid the scarcity of training data required to construct the CNN model. For this, they selected seven CNN architectures that had already been trained on an ImageNet dataset that they carefully fitted on (MRI) data of brain tumors gathered from the BRATS database, as shown in Figure 13. Just the prediction

that received the highest score among the predictions made by the seven pretrained CNNs is produced to increase their model's accuracy. They evaluated the effectiveness of the primary two-class model, which includes LGG and HGG brain cancers, using a ten-way cross-validation method. The test precision, F1 score, test precision, and test sensitivity for their suggested model were 98.67%, 98.06%, 98.33%, and 98.06%, respectively.

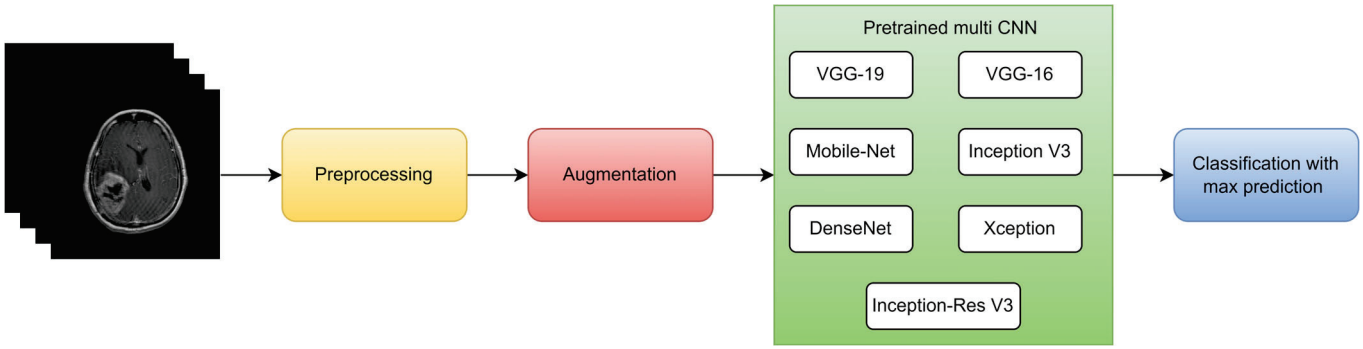


Figure 13. Proposed process for deep transfer learning. Reprinted (adapted) with permission from [106]. Copyright 2021 Indonesian Journal of Electrical Engineering and Computer Science.

A new AI diagnosis model called EfficientNetB0 was created by Khazaee et al. [107] to assess and categorize human brain gliomas utilizing sequences from MR images. They used a common dataset (BRATS-2019) to validate the new AI model, and they showed that the AI components—CNN and transfer learning—provided outstanding performance for categorizing and grading glioma images, with 98.8% accuracy.

In [70], the researchers developed a model using transfer learning and pretrained ResNet18 to identify basal ganglia germinomas more accurately. In this retrospective analysis, 73 patients with basal ganglioma were enrolled. Based on both T1 and T2 data, brain tumors were manually segmented. To create the tumor classification model, the T1 sequence was utilized. Transfer learning and a 2D convolutional network were used. Five-fold cross-validation was used to train the model, and it resulted in a mean AUC of 88%.

Researchers suggested an effective hyperparameter optimization method for CNN based on Bayesian optimization [108]. This method was assessed by categorizing 3064 T1 images into three types of brain cancers (glioma, pituitary, and meningioma). Five popular deep pretrained models are compared to the improved CNN's performance using transfer learning. Their CNN achieved 98.70% validation accuracy after applying Bayesian optimization.

A novel generated transfer DL model was developed by Alanazi et al. [109] for the early diagnosis of brain cancers into their different categories, such as meningioma, pituitary, and glioma. Several layers of the models were first constructed from scratch to test the performance of standalone CNN models performed for brain MRI images. The weights of the neurons were then revised using the transfer learning approach to categorize brain MRI images into tumor subclasses using the 22-layer, isolated CNN model. Consequently, the transfer-learned model that was created had an accuracy rate of 95.75%.

Rizwan et al. [110] suggested a method to identify various BT classes using Gaussian-CNN on two datasets. One of the datasets is employed to categorize lesions into pituitary, glioma, and meningioma. The other distinguishes between the three glioma classes (II, III, and IV). The first and second datasets, respectively, have 233 and 73 victims from a total of 3064 and 516 images on T1 enhanced images. For the two datasets, the suggested method has an accuracy of 99.8% and 97.14%.

A seven-layer CNN was suggested in [111] to assist with the three-class categorization of brain MR images. To decrease computing time, separable convolution was used. The suggested separable CNN model achieved 97.52% accuracy on a publicly available dataset of 3064 images.

Several pretrained CNNs were utilized in [112], including GoogleNet, Alexnet, Resnet50, Resnet101, VGG-16, VGG-19, InceptionResNetV2, and Inceptionv3. To accommodate additional image categories, the final few layers of these networks were modified. Data from the clinical, Harvard, and Figshare repositories were widely used to assess these models. The dataset was divided into training and testing halves in a 60:40 ratio. The validation on the test set demonstrates that, compared to other proposed models, the Alexnet with transfer learning demonstrated the best performance in the shortest time. The suggested method obtained accuracies of 100%, 94%, and 95.92% using three datasets and is more generic because it does not require any manually created features.

The suggested framework [113] describes three experiments that classified brain malignancies such as meningiomas, gliomas, and pituitary tumors using three designs of CNN (AlexNet, VGGNet, and GoogleNet). Using the MRI slices of the brain tumor dataset from Figshare, each study then investigates transfer learning approaches like fine-tuning and freezing. The data augmentation approaches are applied to the MRI slices for results generalization, increasing dataset samples, and minimizing the risk of overfitting. The fine-tuned VGG16 architecture attained the best accuracy at 98.69% in terms of categorization in the proposed studies.

An effective hybrid optimization approach was used in [114] for the segmentation and classification of brain tumors. To improve categorization, the CNN features were extracted. The suggested chronological Jaya honey badger algorithm (CJHBA) was used to train the deep residual network (DRN), which was used to conduct the classification by using the retrieved features as input. The Jaya algorithm, the honey badger algorithm (HBA), and the chronological notion are all combined in the proposed CJHBA. Using BRATS-2018, the performance is assessed. The highest accuracy is 92.10%. A summary of MRI brain tumor classification using DL is provided in Table 7.

Table 7. MRI brain tumor classification using DL.

Ref.	Scan	Year	Technique	Method	Result	Performance Metrics
[101]	MRI	2015	DL	Custom-CNN	96.00%	Acc
[7]	MRI	2019	DL	Custom-CNN	98.70%	Acc
[102]	MRI	2020	DL	VGG-16, Inception-v3, ResNet-50	96% 75% 89%	Acc
[103]	MRI	2021	DL	AlexNet, GoogleNet, SqueezeNet	97.10%	Acc
[104]	MRI	2021	DL	Custom-CNN	82.89%	ROC
[105]	MRI	2018	DL	AlexNet	90.90%	Test acc
[106]	MRI	2021	DL	multi-CNN structure	98.67% 98.06% 98.33% 98.06%	precision, F1 score, precision, sensitivity
[107]	MRI	2022	DL	EfficientNetB0	98.80%	Acc
[70]	MRI	2022	DL	ResNet18	88.00%	AUC
[108]	MRI	2022	DL	Custom-CNN	98.70%	Acc
[109]	MRI	2022	DL	Custom-CNN	95.75%	Acc
[110]	MRI	2022	DL	Gaussian-CNN	99.80%	Acc
[111]	MRI	2020	DL	seven-layer CNN	97.52%	Acc
[112]	MRI	2021	DL	Alexnet	100.00%	Acc
[113]	MRI	2019	DL	VGG16	98.69%	Acc
[114]	MRI	2023	DL	CNN	92.10%	Acc

5.3.4. Hybrid Techniques

Hybrid strategies use multiple approaches to achieve high accuracy, emphasizing each approach's benefits while minimizing the drawbacks. The first method employed a segmentation technique to identify the part of the brain that was infected, and the second method for classification. Hybrid techniques are summarized in Table 8.

The proposed integrated SVM and ANN-based method for classification can be discovered in [115]. The FCM method is used to segment the brain MRI images initially, where the updated membership and k value diverge from the standard method. Two types of characteristics have been retrieved from segmented images to distinguish and categorize tumors. Using SVM, the first category of statistical features was used to differentiate between normal or abnormal brain MRI images. This SVM technique has an accuracy rate of 97.44%. Area, perimeter, orientation, and eccentricity were additional criteria used to distinguish between the tumor and various malignant stages I through IV. The tumor categories and stages of malignant tumors are classified through the ANN back-propagation technique. This suggested strategy has a 97.37% accuracy rate for categorizing tumor stages.

A hybrid segmentation strategy using ANN was suggested in [116] to enhance the brain tumor's classification outcomes. First, the tumor region was segmented using skull stripping and thresholding. The segmented tumor was subsequently recognized using the canny algorithm, and the features of the identified tumor cell region were then used as the input of the ANN for classification; 98.9% accuracy can be attained with the provided strategy.

A system that can identify and categorize the different types of tumors as well as detect them in T1 and T2 image sequences was proposed by Ramdlon et al. [52]. Only the axial section of the MRI results, which are divided into three classifications (Glioblastoma, Astrocytoma, and Oligodendroglioma), are used for the data analysis using this method. Basic image processing techniques were used to identify the tumor region, including image enhancement, binarization, morphology, and watershed. Following the shape extraction feature segmentation, the KNN classifier was used to classify tumors; 89.5% of tumors were correctly classified.

Gurbina et al. [30] described the suggested integrated DWT and SVM classification methods. The initial segmentation of the brain MRI images was performed using Ostu's approach. The DWT features were obtained from segmented images to identify and categorize tumors. Brain MRI images were divided into benign and malignant categories using an SVM classifier. This SVM method has a 99% accuracy rate.

The objective of the study in [117] is multilevel segmentation for effective feature extraction and brain tumor classification from MRI data. The authors used thresholding, the watershed algorithm, and morphological methods for segmentation after preprocessing the MRI image data. Through CNN, features are extracted, and SVM classed the tumor images as malignant or noncancerous. The proposed algorithm has an overall accuracy of 87.4%.

The classification of brain tumors into three types—glioblastoma, sarcoma, and metastatic—has been proposed by the authors of [118]. The authors first used FCM clustering to segment the brain tumor and then DWT to extract the features. PCA was then used to minimize the characteristics. Using six layers of DNN, categorization was completed. The suggested method displays 98% accuracy.

The method presented by Babu et al. [119] focused on categorizing and segmenting brain cancers from MRI images. Four processes compose the procedure: image denoising, segmentation of tumor, extracting features, and hybrid classification. They used the wavelet-based method to extract features after employing the thresholding process to remove tumors from brain MRI images. The final hybrid categorization was performed using CNN. The experiment's findings showed that the approach had a segmentation accuracy of 95.23%, but the suggested optimized CNN had a classification accuracy of 99%.

Improved SVM was suggested as a novel algorithm by Ansari [120]. They recommended four steps for identifying and classifying brain tumors using MRI data: preprocess-

ing, segmentation of images, extracting features, and image categorization. They segmented tumors using a fuzzy clustering approach and extracted key features using GLCM. In the classification stage, improved SVM was finally used. The suggested approach has an 88% accuracy rate.

A fully automated system for segmenting and diagnosing brain tumors was proposed by Farajzadeh et al. [121]. This is accomplished by first applying five distinct preprocessing techniques to an MR image, passing the images through a DWT, and then extracting six local attributes from the image. The processed images are then delivered to an NN, which subsequently extracts higher-order attributes from them. Another NN then weighs the features and concatenates them with the initial MR image. The hybrid U-Net is then fed with the concatenated data to segment the tumor and classify the image. For segmenting and categorizing brain tumors, they attained accuracy rates of 98.93% and 98.81%, respectively.

Table 8. Hybrid techniques.

Ref.	Year	Segmentation Method	Feature Extraction	Classifier	Accuracy
[115]	2017	FCM	shape and statistical	SVM and ANN	97.44% and 97.37%
[118]	2017	FCM	DWT and PCA	CNN	98.00%
[52]	2019	watershed	shape	KNN	89.50%
[30]	2019	Ostu's	DWT	SVM	99.00%
[117]	2020	thresholding and watershed	CNN	SVM	87.4%.
[116]	2020	canny	GLCM and Gabor	ANN	98.90%
[119]	2023	thresholding	wavelet	CNN	99.00%
[120]	2023	fuzzy clustering	GLCM	Improved SVM	88.00%
[121]	2023	U-Net	DWT	CNN	98.93%

5.3.5. Various Segmentation and Classification Methods Employing CT Images

Wavelet statistical texture features (WST) and wavelet co-occurrence texture features (WCT) were combined to segment brain tumors in CT images [122] automatically. After utilizing GA to choose the best texture features, two different NN classifiers were tested to segment the region of a tumor. This approach is shown to provide good outcomes with an accuracy rate of above 97%. Architecture of NN is shown in Figure 14.

For the segmentation and classification of cancers in brain CT images utilizing SVM with GA feature selection, a novel dominating feature extraction methodology was presented in [123]. They used FCM and K-means during the segmentation step and GLCM and WCT during the feature extraction stage. This approach is shown to provide positive results with an accuracy rate of above 98%.

An improved semantic segmentation model for CT images was suggested in [124]. Additionally, classification is used in the suggested work. In the suggested architecture, the semantic segmentation network, which has several convolutional layers and pooling layers, was used to first segment the brain image. Then, using the GoogleNet model, the tumor was divided into three groups: meningioma, glioma, and pituitary tumor. The overall accuracy achieved with this strategy was 99.6%.

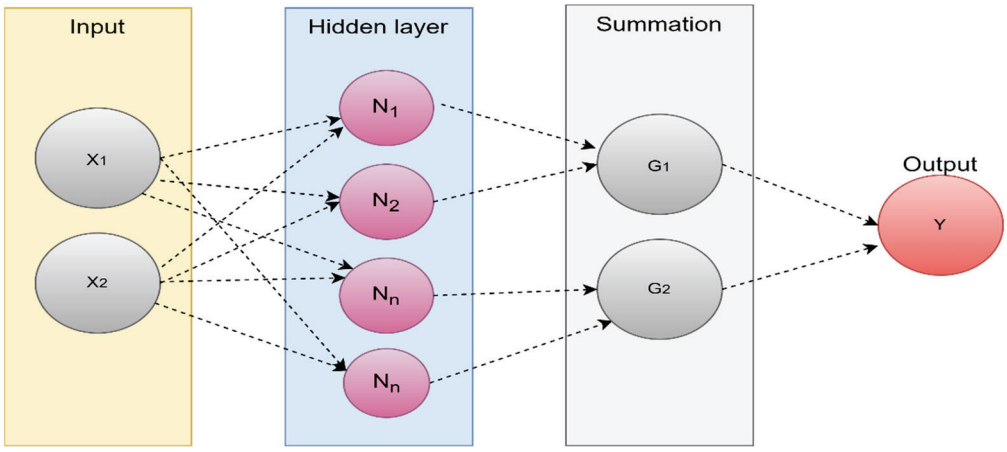


Figure 14. Architecture of NN.

A unique correlation learning technique utilizing CNN and ANN was proposed by Woniak et al. [125]. CNN used the support neural network to determine the best filters for the convolution and pooling layers. Consequently, the main neural classification improved efficiency and learns more quickly. Results indicated that the CLM model can achieve 96% accuracy, 95% precision, and 95% recall.

The contribution of image fusion to an enhanced brain tumor classification framework was examined by Nanmaran et al. [126], and this new fusion-based tumor categorization model can be more successfully applied to personalized therapy. A distinct cosine transform-based (DCT) fusion technique is utilized to combine MRI and SPECT images of benign and malignant class brain tumors. With the help of the features extracted from fused images, SVM, KNN, and decision trees were set to test. When using features extracted from fused images, the SVM classifier outperformed KNN and decision tree classifiers with an overall accuracy of 96.8%, specificity of 93%, recall of 94%, precision of 95%, and F1 score of 91%. Table 9 provides different segmentation and classification methods employing CT images.

Table 9. Various segmentation and classification methods employing CT images.

Ref.	Year	Type	Segmentation	Feature Extraction	Feature Selection	Classification	Result
[122]	2011	CT	NN	WCT and WST	GA	-	97.00%
[123]	2011	CT	FCM and k-mean	GLCM and WCT	GA	SVM	98.00%
[124]	2020	CT	Semantic	-	-	GoogleNet	99.60%
[125]	2021	CT	-	-	-	CNN	96.00%
[126]	2022	SPECT/MRI	-	DCT	-	SVM	96.80%

6. Discussion

Most brain tumor segmentation and classification strategies are presented in this review. The quantitative efficiency of numerous conventional ML- and DL-based algorithms is covered in this article. Figure 15 displays the total number of publications published between 2010 and 2022 used in this review. Figure 16 displays the total number of articles published that perform classification, segmentation, or both.

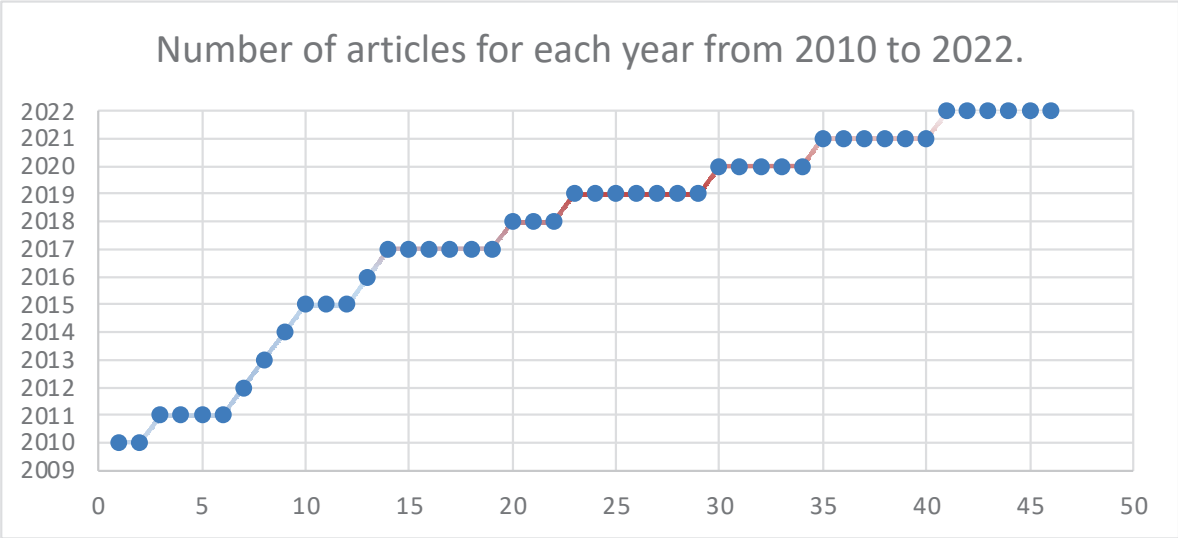


Figure 15. Number of articles published from 2010 to 2022.

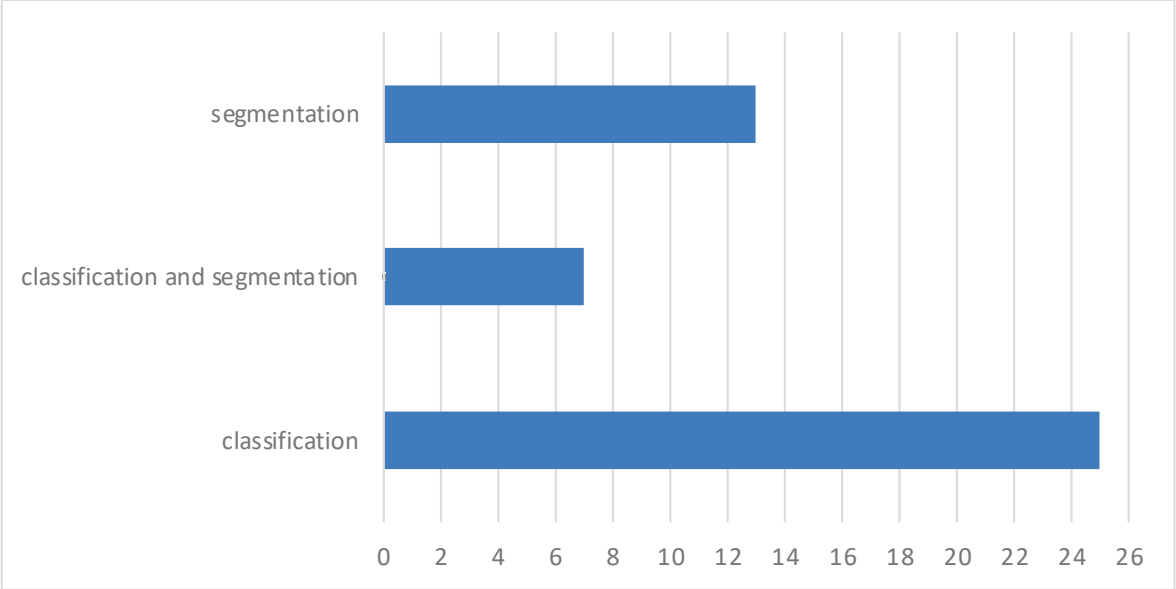


Figure 16. Number of articles published that perform classification, segmentation, or both.

Brain tumor segmentation uses traditional image segmentation methods like region growth and unsupervised machine learning. Noise, low image quality, and the initial seed point are its biggest challenges. The classification of pixels into multiple classes has been accomplished in the second generation of segmentation methods using unsupervised ML, such as FCM and K-means. These techniques are, nevertheless, quite noise sensitive. Pixel-level classification-based segmentation approaches utilizing conventional supervised ML have been presented to overcome this difficulty. Feature engineering, which extracts the tumor-descriptive pieces of information for the model's training, is frequently used in conjunction with these techniques. Additionally, postprocessing helps further improve the results of supervised machine learning segmentation. Through the pipeline of its component parts, the deep learning-based approach accomplishes an end-to-end segmentation of tumors using an MRI image. These models frequently eliminate the requirement for manually built features by automatically extracting tumor descriptive information. However, their application in the medical domains is limited by the need for a big dataset for training the models and the complexity of understanding them.

In addition to the segmentation of the brain cancer region from the MRI scan, the classification of the tumor into its appropriate type is crucial for diagnosis and treatment planning, which in today's medical practice necessitates a biopsy process. Several approaches that use shallow ML and DL have been put forth for classifying brain tumors. Type shallow ML techniques frequently include preprocessing, ROI identification, and feature extraction steps. Extracting descriptive information is a difficult task because of the inherent noise sensitivity associated with MRI image collection as well as differences in the shape, size, and position of tumor tissue cells. As a result, deep learning algorithms are currently the most advanced method for classifying many types of brain cancers, including astrocytomas, gliomas, meningiomas, and pituitary tumors. This review has covered several classifications of brain tumors.

The noisy nature of an MRI image is one of the most frequent difficulties in ML-based segmentation and classification of brain tumors. To increase the precision of brain tumor segmentation and classification models, noise estimation and denoising MRI images is a vital preprocessing operation. As a result, several methods, including the median filter [115], Wiener filter and DWT [30], and DL-based methods [117], have been suggested for denoising MRI images.

Large amounts of data are needed for DL models to operate effectively, but there need to be more datasets available. Data augmentation aids in expanding small datasets and creating a powerful generalized model. A common augmentation method for MRI images has yet to be developed. Although many methods have been presented by researchers, their primary goal is to increase the number of images. Most of the time, they ignore the connections between space and texture. An identical augmentation technique is required for comparative analysis to be conducted on its foundation.

7. General Problems and Challenges

Features are first manually extracted for ML, and are then fed into the ML-based differentiation system. Continuous variation within image classes makes utilizing ML-based algorithms for image classification challenging. Furthermore, the feature extraction methods' usage of modern distance metrics makes it impossible to determine the similarity between two images.

Deep learning analyzes several parameters and optimizes them to extract and select features on its own. However, the system lacks intelligence in feature selection and typically pools, which reduces parameters and eliminates features that could be useful to the entire system.

Furthermore, DL models need data, and those data are coupled with millions or trillions of parameters. Therefore, enormous amounts of memory and GPU-based computers are required in the current environment. However, because of their high cost, these devices are not available to everyone. Consequently, many researchers need to create models that fit within their available budgets, which significantly impacts the quality of their study.

The noisy nature of an MRI image is one of the most frequent difficulties in ML-based brain tumor detection and classification. Preprocessing is necessary to remove all forms of noise from data and make it more suitable for the task at hand. Preprocessing difficulties exist in all the available datasets. However, the BRATS datasets have problems, such as motion artifacts and noise. There is no established preprocessing standard currently. People employ subpar application software, causing the image quality to decrease rather than improve.

7.1. Brain Cancer and Other Brain Disorders

7.1.1. Stroke

Hemorrhagic strokes come from blood vessel injury or aberrant vascular structure, while ischemic strokes occur when the brain's blood supply is cut off. Although the fact that strokes and brain tumors are two distinct illnesses, the connections associated with them have been studied [127].

They discovered that stroke patients are more likely than other cancer types to acquire brain cancer. Another intriguing conclusion of the study is that women between the ages of 40 and 60 and elderly stroke patients are more likely to acquire brain cancer.

7.1.2. Alzheimer's Disease

Short-term loss of memory is an initial symptom of Alzheimer's disease (AD), a chronic neurodegenerative illness that may become worse over time as the disease progresses [108]. Despite AD and cancer being two distinct diseases, several studies have found a connection between them. According to the research, there is an inverse association between cancer and Alzheimer's disease. They discovered that patients who had cancer had a 33% lower risk of Alzheimer's disease than individuals who had not had cancer throughout the course of a mean follow-up of 10 years. Another intriguing finding of the study was that people with AD had a 61% lower risk of developing cancer.

8. Future Directions

The main applications of CADx systems are in educating and training; clinical practice is not one of them. CADx-based systems still need to be widely used in clinics. The absence of established techniques for assessing CADx systems in a practical environment is one cause of this. The performance metrics outlined in this study provide a helpful and necessary baseline for comparing algorithms, but because they are all so dependent on the training set, more advanced tools are required.

The fact that the image formats utilized to train the models were those characteristics of the AI research field (PNG) rather than those of the radiology field (DICOM, NIfTI) is noteworthy. Many of the articles analyzed needed authors with clinical backgrounds.

A different but related technical issue that may affect the performance of CADx systems in practice is the need for physician training on interacting with and interpreting the results of such systems for diagnostic decisions. This issue must be dealt with in all the papers included in the review. In terms of research project relevance and the acceptance of its findings, greater participation by doctors in the process may be advantageous.

9. Conclusions

A brain tumor is an abnormal growth of brain tissue that affects the brain's ability to function normally. The primary objective in medical image processing is to find accurate and helpful information with the minimum possible errors by using algorithms. The four steps involved in segmenting and categorizing brain tumors using MRI data are preprocessing, picture segmentation, extracting features, and image classification. The diagnosis, treatment strategy, and patient follow-up can all be greatly enhanced by automating the segmentation and categorization of brain tumors. It is still difficult to create a fully autonomous system that can be deployed on clinical floors due to the appearance of the tumor and its irregular size, form, and nature. The review's primary goal is to present the state-of-the-art in the field of brain cancer, which includes the pathophysiology of the disease, imaging technologies, WHO classification standards for tumors, primary methods of diagnosis, and CAD algorithms for brain tumor classifications using ML and DL techniques. Automating the segmentation and categorization of brain tumors using deep learning techniques has many advantages over region-growing and shallow ML systems. DL algorithms' powerful feature learning capabilities are primarily to blame for this. Although DL techniques have made a substantial contribution, a general technique is still needed. This study reviewed 53 studies that used ML and DL to classify brain tumors based on MRI, and it examined the challenges and obstacles that CAD brain tumor classification techniques now face in practical application and advancement—a thorough examination of the variables that might impact classification accuracy. The MRI sequences and web address of the online repository for the dataset are among the publicly available databases that have been briefly listed in Table 4 and used in the experiments evaluated in this paper.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

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

References

1. Watson, C.; Kirkcaldie, M.; Paxinos, G. The Brain: An Introduction to Functional Neuroanatomy. 2010. Available online: http://ci.nii.ac.jp/ncid/BB04049625 (accessed on 22 May 2023).
2. Jellinger, K.A. The Human Nervous System Structure and Function, 6th edn. *Eur. J. Neurol.* **2009**, *16*, e136. [CrossRef]
3. DeAngelis, L.M. Brain tumors. *N. Engl. J. Med.* **2001**, *344*, 114–123. [CrossRef]
4. Louis, D.N.; Perry, A.; Wesseling, P.; Brat, D.J.; Cree, I.A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.K.; Pfister, S.M.; Reifenberger, G.; et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A summary. *Neuro-Oncology* **2021**, *23*, 1231–1251. [CrossRef]
5. Hayward, R.M.; Patronas, N.; Baker, E.H.; Vézina, G.; Albert, P.S.; Warren, K.E. Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. *J. Neuro-Oncol.* **2008**, *90*, 57–61. [CrossRef]
6. Mahaley, M.S., Jr.; Mettlin, C.; Natarajan, N.; Laws, E.R., Jr.; Peace, B.B. National survey of patterns of care for brain-tumor patients. *J. Neurosurg.* **1989**, *71*, 826–836. [CrossRef] [PubMed]
7. 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]
8. Johnson, D.R.; Guerin, J.B.; Giannini, C.; Morris, J.M.; Eckel, L.J.; Kaufmann, T.J. 2016 Updates to the WHO Brain Tumor Classification System: What the Radiologist Needs to Know. *RadioGraphics* **2017**, *37*, 2164–2180. [CrossRef] [PubMed]
9. Buckner, J.C.; Brown, P.D.; O'Neill, B.P.; Meyer, F.B.; Wetmore, C.J.; Uhm, J.H. Central Nervous System Tumors. *Mayo Clin. Proc.* **2007**, *82*, 1271–1286. [CrossRef] [PubMed]
10. World Health Organization: WHO, "Cancer". July 2019. Available online: https://www.who.int/health-topics/cancer (accessed on 30 March 2022).
11. Amyot, F.; Arciniegas, D.B.; Brazaitis, M.P.; Curley, K.C.; Diaz-Arrastia, R.; Gandjbakhche, A.; Herscovitch, P.; Hinds, S.R.; Manley, G.T.; Pacifico, A.; et al. A Review of the Effectiveness of Neuroimaging Modalities for the Detection of Traumatic Brain Injury. *J. Neurotrauma* **2015**, *32*, 1693–1721. [CrossRef]
12. Pope, W.B. Brain metastases: Neuroimaging. *Handb. Clin. Neurol.* **2018**, *149*, 89–112. [CrossRef]
13. Abd-Ellah, M.K.; Awad, A.I.; Khalaf, A.A.; Hamed, H.F. A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned. *Magn. Reson. Imaging* **2019**, *61*, 300–318. [CrossRef] [PubMed]
14. Ammari, S.; Pitre-Champagnat, S.; Dercle, L.; Chouzenoux, E.; Moalla, S.; Reuze, S.; Talbot, H.; Mokoyoko, T.; Hadchiti, J.; Diffetocq, S.; et al. Influence of Magnetic Field Strength on Magnetic Resonance Imaging Radiomics Features in Brain Imaging, an In Vitro and In Vivo Study. *Front. Oncol.* **2021**, *10*, 541663. [CrossRef] [PubMed]
15. Sahoo, L.; Sarangi, L.; Dash, B.R.; Palo, H.K. Detection and Classification of Brain Tumor Using Magnetic Resonance Images. In *Advances in Electrical Control and Signal Systems: Select Proceedings of AECSS, Bhubaneswar, India, 8–9 November 2019*; Springer: Singapore, 2020; Volume 665, pp. 429–441. [CrossRef]
16. Kaur, R.; Doegar, A. Localization and Classification of Brain Tumor using Machine Learning & Deep Learning Techniques. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 59–66.
17. The Radiology Assistant: Multiple Sclerosis 2.0. 1 December 2021. Available online: https://radiologyassistant.nl/neuroradiology/multiple-sclerosis/diagnosis-and-differential-diagnosis-3#mri-protocol-ms-brain-protocol (accessed on 22 May 2023).
18. Savoy, R.L. Functional magnetic resonance imaging (fMRI). In *Encyclopedia of Neuroscience*; Elsevier: Charlestown, MA, USA, 1999.
19. Luo, Q.; Li, Y.; Luo, L.; Diao, W. Comparisons of the accuracy of radiation diagnostic modalities in brain tumor. *Medicine* **2018**, *97*, e11256. [CrossRef]
20. Positron Emission Tomography (PET). Johns Hopkins Medicine. 20 August 2021. Available online: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/positron-emission-tomography-pet (accessed on 20 May 2023).
21. Mayfield Brain and Spine. SPECT Scan. 2022. Available online: https://mayfieldclinic.com/pe-spect.htm (accessed on 22 May 2023).
22. Sastry, R.; Bi, W.L.; Pieper, S.; Frisken, S.; Kapur, T.; Wells, W.; Golby, A.J. Applications of Ultrasound in the Resection of Brain Tumors. *J. Neuroimaging* **2016**, *27*, 5–15. [CrossRef]
23. Nasrabadi, N.M. Pattern recognition and machine learning. *J. Electron. Imaging* **2007**, *16*, 49901.
24. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine learning for medical imaging. *Radiographics* **2017**, *37*, 505–515. [CrossRef]
25. Mohan, M.R.M.; Sulochana, C.H.; Latha, T. Medical image denoising using multistage directional median filter. In Proceedings of the 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], Nagercoil, India, 9–20 March 2015.
26. Borole, V.Y.; Nimbhore, S.S.; Kawthekar, S.S. Image processing techniques for brain tumor detection: A review. *Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS)* **2015**, *4*, 2.

27. Ziedan, R.H.; Mead, M.A.; Eltawel, G.S. Selecting the Appropriate Feature Extraction Techniques for Automatic Medical Images Classification. *Int. J.* **2016**, *4*, 1–9.
28. Amin, J.; Sharif, M.; Yasmin, M.; Fernandes, S.L. A distinctive approach in brain tumor detection and classification using MRI. *Pattern Recognit. Lett.* **2017**, *139*, 118–127. [CrossRef]
29. Islam, A.; Reza, S.M.; Iftekharuddin, K.M. Multifractal texture estimation for detection and segmentation of brain tumors. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 3204–3215. [CrossRef]
30. Gurbină, M.; Lascu, M.; Lascu, D. Tumor detection and classification of MRI brain image using different wavelet transforms and support vector machines. In Proceedings of the 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, 1–3 July 2019; pp. 505–508.
31. Xu, X.; Zhang, X.; Tian, Q.; Zhang, G.; Liu, Y.; Cui, G.; Meng, J.; Wu, Y.; Liu, T.; Yang, Z.; et al. Three-dimensional texture features from intensity and high-order derivative maps for the discrimination between bladder tumors and wall tissues via MRI. *Int. J. Comput. Assist. Radiol. Surg.* **2017**, *12*, 645–656. [CrossRef]
32. Kaplan, K.; Kaya, Y.; Kuncan, M.; Ertunç, H.M. Brain tumor classification using modified local binary patterns (LBP) feature extraction methods. *Med. Hypotheses* **2020**, *139*, 109696. [CrossRef]
33. Afza, F.; Khan, M.S.; Sharif, M.; Saba, T. Microscopic skin laceration segmentation and classification: A framework of statistical normal distribution and optimal feature selection. *Microsc. Res. Tech.* **2019**, *82*, 1471–1488. [CrossRef]
34. Lakshmi, A.; Arivoli, T.; Rajasekaran, M.P. A Novel M-ACA-Based Tumor Segmentation and DAPP Feature Extraction with PPCSO-PKC-Based MRI Classification. *Arab. J. Sci. Eng.* **2017**, *43*, 7095–7111. [CrossRef]
35. Adair, J.; Brownlee, A.; Ochoa, G. Evolutionary Algorithms with Linkage Information for Feature Selection in Brain Computer Interfaces. In *Advances in Computational Intelligence Systems*; Springer Nature: Cham, Switzerland, 2016; pp. 287–307.
36. Arakeri, M.P.; Reddy, G.R.M. Computeraided diagnosis system for tissue characterization of brain tumor on magnetic resonance images. *Signal Image Video Process.* **2015**, *9*, 409–425. [CrossRef]
37. Wang, S.; Zhang, Y.; Dong, Z.; Du, S.; Ji, G.; Yan, J.; Phillips, P. Feed-forward neural network optimized by hybridization of PSO and ABC for abnormal brain detection. *Int. J. Imaging Syst. Technol.* **2015**, *25*, 153–164. [CrossRef]
38. Abbasi, S.; Tajeripour, F. Detection of brain tumor in 3D MRI images using local binary patterns and histogram orientation gradient. *Neurocomputing* **2017**, *219*, 526–535. [CrossRef]
39. Zöllner, F.G.; Emblem, K.E.; Schad, L.R. SVM-based glioma grading: Optimization by feature reduction analysis. *Z. Med. Phys.* **2012**, *22*, 205–214. [CrossRef]
40. Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
41. Bhatele, K.R.; Bhadauria, S.S. Brain structural disorders detection and classification approaches: A review. *Artif. Intell. Rev.* **2019**, *53*, 3349–3401. [CrossRef]
42. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
43. Hu, A.; Razmjooy, N. Brain tumor diagnosis based on metaheuristics and deep learning. *Int. J. Imaging Syst. Technol.* **2020**, *31*, 657–669. [CrossRef]
44. Tandel, G.S.; Balestrieri, A.; Jujaray, T.; Khanna, N.N.; Saba, L.; Suri, J.S. Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm. *Comput. Biol. Med.* **2020**, *122*, 103804. [CrossRef] [PubMed]
45. Sahaai, M.B. Brain tumor detection using DNN algorithm. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2021**, *12*, 3338–3345.
46. Hashemi, M. Enlarging smaller images before inputting into convolutional neural network: Zero-padding vs. interpolation. *J. Big Data* **2019**, *6*, 98. [CrossRef]
47. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. *Briefings Bioinform.* **2017**, *19*, 1236–1246. [CrossRef]
48. Gorach, T. Deep convolutional neural networks—A review. *Int. Res. J. Eng. Technol. (IRJET)* **2018**, *5*, 439.
49. Ogundokun, R.O.; Maskeliunas, R.; Misra, S.; Damaševičius, R. Improved CNN Based on Batch Normalization and Adam Optimizer. In Proceedings of the Computational Science and Its Applications–ICCSA 2022 Workshops, Malaga, Spain, 4–7 July 2022; Part V. pp. 593–604.
50. Ismael SA, A.; Mohammed, A.; Hefny, H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks. *Artif. Intell. Med.* **2020**, *102*, 101779. [CrossRef]
51. Baheti, P. A Comprehensive Guide to Convolutional Neural Networks. V7. Available online: https://www.v7labs.com/blog/convolutional-neural-networks-guide (accessed on 24 April 2023).
52. Ramdlon, R.H.; Kusumaningtyas, E.M.; Karlita, T. Brain Tumor Classification Using MRI Images with K-Nearest Neighbor Method. In Proceedings of the 2019 International Electronics Symposium (IES), Surabaya, Indonesia, 27–28 September 2019; pp. 660–667. [CrossRef]
53. Gurusamy, R.; Subramaniam, V. A machine learning approach for MRI brain tumor classification. *Comput. Mater. Contin.* **2017**, *53*, 91–109.
54. Pohle, R.; Toennies, K.D. Segmentation of medical images using adaptive region growing. In Proceedings of the Medical Imaging 2001: Image Processing, San Diego, CA, USA, 4–10 November 2001; Volume 4322, pp. 1337–1346. [CrossRef]
55. Dey, N.; Ashour, A.S. Computing in medical image analysis. In *Soft Computing Based Medical Image Analysis*; Academic Press: Cambridge, MA, USA, 2018; pp. 3–11.

56. Hooda, H.; Verma, O.P.; Singhal, T. Brain tumor segmentation: A performance analysis using K-Means, Fuzzy C-Means and Region growing algorithm. In Proceedings of the 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, India, 8–10 May 2014; pp. 1621–1626.
57. Sharif, M.; Tanvir, U.; Munir, E.U.; Khan, M.A.; Yasmin, M. Brain tumor segmentation and classification by improved binomial thresholding and multi-features selection. *J. Ambient. Intell. Humaniz. Comput.* **2018**, 1–20. [CrossRef]
58. Shanthi, K.J.; Kumar, M.S. Skull stripping and automatic segmentation of brain MRI using seed growth and threshold techniques. In Proceedings of the 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, Malaysia, 25–28 November 2007; pp. 422–426. [CrossRef]
59. Zhang, F.; Hancock, E.R. New Riemannian techniques for directional and tensorial image data. *Pattern Recognit.* **2010**, *43*, 1590–1606. [CrossRef]
60. Singh, N.P.; Dixit, S.; Akshaya, A.S.; Khodanpur, B.I. Gradient Magnitude Based Watershed Segmentation for Brain Tumor Segmentation and Classification. In *Advances in Intelligent Systems and Computing*; Springer Nature: Cham, Switzerland, 2017; pp. 611–619. [CrossRef]
61. Couprie, M.; Bertrand, G. Topological gray-scale watershed transformation. *Vis. Geom. VI* **1997**, *3168*, 136–146. [CrossRef]
62. Khan, M.S.; Lali, M.I.U.; Saba, T.; Ishaq, M.; Sharif, M.; Saba, T.; Zahoor, S.; Akram, T. Brain tumor detection and classification: A framework of marker-based watershed algorithm and multilevel priority features selection. *Microsc. Res. Tech.* **2019**, *82*, 909–922. [CrossRef]
63. Lotufo, R.; Falcao, A.; Zampirolli, F. IFT-Watershed from gray-scale marker. In Proceedings of the XV Brazilian Symposium on Computer Graphics and Image Processing, Fortaleza, Brazil, 10 October 2003. [CrossRef]
64. Dougherty, E.R. *An Introduction to Morphological Image Processing*; SPIE Optical Engineering Press: Bellingham, WA, USA, 1992.
65. Kaur, D.; Kaur, Y. Various image segmentation techniques: A review. *Int. J. Comput. Sci. Mob. Comput.* **2014**, *3*, 809–814.
66. Aslam, A.; Khan, E.; Beg, M.S. Improved Edge Detection Algorithm for Brain Tumor Segmentation. *Procedia Comput. Sci.* **2015**, *58*, 430–437. [CrossRef]
67. Egmont-Petersen, M.; de Ridder, D.; Handels, H. Image processing with neural networks—A review. *Pattern Recognit.* **2002**, *35*, 2279–2301. [CrossRef]
68. Cui, B.; Xie, M.; Wang, C. A Deep Convolutional Neural Network Learning Transfer to SVM-Based Segmentation Method for Brain Tumor. In Proceedings of the 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), Jinan, China, 18–20 October 2019; pp. 1–5. [CrossRef]
69. Pereira, S.; Pinto, A.; Alves, V.; Silva, C.A. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1240–1251. [CrossRef]
70. Ye, N.; Yu, H.; Chen, Z.; Teng, C.; Liu, P.; Liu, X.; Xiong, Y.; Lin, X.; Li, S.; Li, X. Classification of Gliomas and Germinomas of the Basal Ganglia by Transfer Learning. *Front. Oncol.* **2022**, *12*, 844197. [CrossRef]
71. Biratu, E.S.; Schwenker, F.; Ayano, Y.M.; Debelee, T.G. A survey of brain tumor segmentation and classification algorithms. *J. Imaging* **2021**, *7*, 179. [CrossRef]
72. Wikipedia Contributors. F Score. Wikipedia. 2023. Available online: https://en.wikipedia.org/wiki/F-score (accessed on 22 May 2023).
73. Brain Tumor Segmentation (BraTS) Challenge. Available online: http://www.braintumorsegmentation.org/ (accessed on 22 May 2023).
74. RIDER NEURO MRI—The Cancer Imaging Archive (TCIA) Public Access—Cancer Imaging Archive Wiki. Available online: https://wiki.cancerimagingarchive.net/display/Public/RIDER+NEURO+MRI (accessed on 22 May 2023).
75. Harvard Medical School Data. Available online: http://www.med.harvard.edu/AANLIB/ (accessed on 16 March 2021).
76. The Cancer Genome Atlas. TCGA. Available online: https://wiki.cancerimagingarchive.net/display/Public/TCGA-GBM (accessed on 22 May 2023).
77. The Cancer Genome Atlas. TCGA-LGG. Available online: https://wiki.cancerimagingarchive.net/display/Public/TCGA-LGG (accessed on 22 May 2023).
78. Cheng, J. Figshare Brain Tumor Dataset. 2017. Available online: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427/5 (accessed on 13 May 2022).
79. IXI Dataset—Brain Development. Available online: https://brain-development.org/ixi-dataset/ (accessed on 22 May 2023).
80. Gordillo, N.; Montseny, E.; Sobrevilla, P. A new fuzzy approach to brain tumor segmentation. In Proceedings of the 2010 IEEE International Conference, Barcelona, Spain, 18–23 July 2010; pp. 1–8. [CrossRef]
81. Rajendran; Dhanasekaran, R. A hybrid Method Based on Fuzzy Clustering and Active Contour Using GGVF for Brain Tumor Segmentation on MRI Images. *Eur. J. Sci. Res.* **2011**, *61*, 305–313.
82. Reddy, K.K.; Solmaz, B.; Yan, P.; Avgeropoulos, N.G.; Rippe, D.J.; Shah, M. Confidence guided enhancing brain tumor segmentation in multi-parametric MRI. In Proceedings of the 9th IEEE International Symposium on Biomedical Imaging, Barcelona, Spain, 2–5 May 2012; pp. 366–369. [CrossRef]
83. Almahfud, M.A.; Setyawan, R.; Sari, C.A.; Setiadi, D.R.I.M.; Rachmawanto, E.H. An Effective MRI Brain Image Segmentation using Joint Clustering (K-Means and Fuzzy C-Means). In Proceedings of the 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 21–22 November 2018; pp. 11–16.

84. Chen, W.; Qiao, X.; Liu, B.; Qi, X.; Wang, R.; Wang, X. Automatic brain tumor segmentation based on features of separated local square. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017.
85. Gupta, N.; Mishra, S.; Khanna, P. Glioma identification from brain MRI using superpixels and FCM clustering. In Proceedings of the 2018 Conference on Information and Communication Technology (CICT), Jabalpur, India, 26–28 October 2018. [CrossRef]
86. Razzak, M.I.; Imran, M.; Xu, G. Efficient Brain Tumor Segmentation with Multiscale Two-Pathway-Group Conventional Neural Networks. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 1911–1919. [CrossRef] [PubMed]
87. Myronenko, A.; Hatamizadeh, A. Robust Semantic Segmentation of Brain Tumor Regions from 3D MRIs. In Proceedings of the International MICCAI Brainlesion Workshop, Singapore, 18 September 2020; pp. 82–89. [CrossRef]
88. Karayegen, G.; Aksahin, M.F. Brain tumor prediction on MR images with semantic segmentation by using deep learning network and 3D imaging of tumor region. *Biomed. Signal Process. Control.* **2021**, *66*, 102458. [CrossRef]
89. Ullah, Z.; Usman, M.; Jeon, M.; Gwak, J. Cascade multiscale residual attention CNNs with adaptive ROI for automatic brain tumor segmentation. *Inf. Sci.* **2022**, *608*, 1541–1556. [CrossRef]
90. Wisaeng, K.; Sa-Ngiamvibool, W. Brain Tumor Segmentation Using Fuzzy Otsu Threshold Morphological Algorithm. *IAENG Int. J. Appl. Math.* **2023**, *53*, 1–12.
91. Zhang, Y.; Dong, Z.; Wu, L.; Wang, S. A hybrid method for MRI brain image classification. *Expert Syst. Appl.* **2011**, *38*, 10049–10053. [CrossRef]
92. Yang, G.; Zhang, Y.; Yang, J.; Ji, G.; Dong, Z.; Wang, S.; Feng, C.; Wang, Q. Automated classification of brain images using wavelet-energy and biogeography-based optimization. *Multimed. Tools Appl.* **2015**, *75*, 15601–15617. [CrossRef]
93. Tiwari, P.; Sachdeva, J.; Ahuja, C.K.; Khandelwal, N. Computer Aided Diagnosis System—A Decision Support System for Clinical Diagnosis of Brain Tumours. *Int. J. Comput. Intell. Syst.* **2017**, *10*, 104–119. [CrossRef]
94. Sachdeva, J.; Kumar, V.; Gupta, I.; Khandelwal, N.; Ahuja, C.K. Segmentation, Feature Extraction, and Multiclass Brain Tumor Classification. *J. Digit. Imaging* **2013**, *26*, 1141–1150. [CrossRef]
95. Jayachandran, A.; Dhanasekaran, R. Severity Analysis of Brain Tumor in MRI Images Using Modified Multitexton Structure Descriptor and Kernel-SVM. *Arab. J. Sci. Eng.* **2014**, *39*, 7073–7086. [CrossRef]
96. El-Dahshan, E.-S.A.; Hosny, T.; Salem, A.-B.M. Hybrid intelligent techniques for MRI brain images classification. *Digit. Signal Process.* **2010**, *20*, 433–441. [CrossRef]
97. Ullah, Z.; Farooq, M.U.; Lee, S.-H.; An, D. A hybrid image enhancement based brain MRI images classification technique. *Med. Hypotheses* **2020**, *143*, 109922. [CrossRef] [PubMed]
98. Kang, J.; Ullah, Z.; Gwak, J. MRI-Based Brain Tumor Classification Using Ensemble of Deep Features and Machine Learning Classifiers. *Sensors* **2021**, *21*, 2222. [CrossRef]
99. Díaz-Pernas, F.; Martínez-Zarzuela, M.; Antón-Rodríguez, M.; González-Ortega, D. A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network. *Healthcare* **2021**, *9*, 153. [CrossRef] [PubMed]
100. Badža, M.M.; Barjaktarović, M. Classification of Brain Tumors from MRI Images Using a Convolutional Neural Network. *Appl. Sci.* **2020**, *10*, 1999. [CrossRef]
101. 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. In Proceedings of the AMIA Annual Symposium, San Francisco, CA, USA, 14–18 November 2015; Volume 2015, pp. 1899–1908.
102. Khan, H.A.; Jue, W.; Mushtaq, M.; Mushtaq, M.U. Brain tumor classification in MRI image using convolutional neural network. *Math. Biosci. Eng.* **2020**, *17*, 6203–6216. [CrossRef]
103. Özcan, H.; Emiroğlu, B.G.; Sabuncuoğlu, H.; Özdoğan, S.; Soyer, A.; Saygı, T. A comparative study for glioma classification using deep convolutional neural networks. *Math. Biosci. Eng. MBE* **2021**, *18*, 1550–1572. [CrossRef]
104. Hao, R.; Namdar, K.; Liu, L.; Khalvati, F. A Transfer Learning–Based Active Learning Framework for Brain Tumor Classification. *Front. Artif. Intell.* **2021**, *4*, 635766. [CrossRef]
105. Yang, Y.; Yan, L.-F.; Zhang, X.; Han, Y.; Nan, H.-Y.; Hu, Y.-C.; Hu, B.; Yan, S.-L.; Zhang, J.; Cheng, D.-L.; et al. Glioma Grading on Conventional MR Images: A Deep Learning Study with Transfer Learning. *Front. Neurosci.* **2018**, *12*, 804. [CrossRef]
106. El Hamdaoui, H.; Benfares, A.; Boujraf, S.; Chaoui, N.E.H.; Alami, B.; Maaroufi, M.; Qjidaa, H. High precision brain tumor classification model based on deep transfer learning and stacking concepts. *Indones. J. Electr. Eng. Comput. Sci.* **2021**, *24*, 167–177. [CrossRef]
107. Khazaee, Z.; Langarizadeh, M.; Ahmadabadi, M.E.S. Developing an Artificial Intelligence Model for Tumor Grading and Classification, Based on MRI Sequences of Human Brain Gliomas. *Int. J. Cancer Manag.* **2022**, *15*, e120638. [CrossRef]
108. Amou, M.A.; Xia, K.; Kamhi, S.; Mouhafid, M. A Novel MRI Diagnosis Method for Brain Tumor Classification Based on CNN and Bayesian Optimization. *Healthcare* **2022**, *10*, 494. [CrossRef] [PubMed]
109. Alanazi, M.; Ali, M.; Hussain, J.; Zafar, A.; Mohatram, M.; Irfan, M.; AlRuwaili, R.; Alruwaili, M.; Ali, N.T.; Albarrak, A.M. Brain Tumor/Mass Classification Framework Using Magnetic-Resonance-Imaging-Based Isolated and Developed Transfer Deep-Learning Model. *Sensors* **2022**, *22*, 372. [CrossRef] [PubMed]
110. Rizwan, M.; Shabbir, A.; Javed, A.R.; Shabbr, M.; Baker, T.; Al-Jumeily, D. Brain Tumor and Glioma Grade Classification Using Gaussian Convolutional Neural Network. *IEEE Access* **2022**, *10*, 29731–29740. [CrossRef]
111. Isunuri, B.V.; Kakarla, J. Three-class brain tumor classification from magnetic resonance images using separable convolution based neural network. *Concurr. Comput. Pract. Exp.* **2021**, *34*, e6541. [CrossRef]

112. Kaur, T.; Gandhi, T.K. Deep convolutional neural networks with transfer learning for automated brain image classification. *J. Mach. Vis. Appl.* **2020**, *31*, 20. [CrossRef]
113. Rehman, A.; Naz, S.; Razzak, M.I.; Akram, F.; Imran, M. A Deep Learning-Based Framework for Automatic Brain Tumors Classification Using Transfer Learning. *Circuits Syst. Signal Process.* **2019**, *39*, 757–775. [CrossRef]
114. Deepa, S.; Janet, J.; Sumathi, S.; Ananth, J.P. Hybrid Optimization Algorithm Enabled Deep Learning Approach Brain Tumor Segmentation and Classification Using MRI. *J. Digit. Imaging* **2023**, *36*, 847–868. [CrossRef]
115. Ahmmed, R.; Swakshar, A.S.; Hossain, M.F.; Rafiq, M.A. Classification of tumors and it stages in brain MRI using support vector machine and artificial neural network. In Proceedings of the 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 16–18 February 2017.
116. Sathi, K.A.; Islam, S. Hybrid Feature Extraction Based Brain Tumor Classification using an Artificial Neural Network. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 155–160. [CrossRef]
117. Islam, R.; Imran, S.; Ashikuzzaman; Khan, M.A. Detection and Classification of Brain Tumor Based on Multilevel Segmentation with Convolutional Neural Network. *J. Biomed. Sci. Eng.* **2020**, *13*, 45–53. [CrossRef]
118. Mohsen, H.; El-Dahshan, E.A.; El-Horbaty, E.M.; Salem, A.M. Classification using deep learning neural networks for brain tumors. *Future Comput. Inform. J.* **2017**, *3*, 68–71. [CrossRef]
119. Babu, P.A.; Rao, B.S.; Reddy, Y.V.B.; Kumar, G.R.; Rao, J.N.; Koduru, S.K.R. Optimized CNN-based Brain Tumor Segmentation and Classification using Artificial Bee Colony and Thresholding. *Int. J. Comput. Commun. Control.* **2023**, *18*, 577. [CrossRef]
120. Ansari, A.S. Numerical Simulation and Development of Brain Tumor Segmentation and Classification of Brain Tumor Using Improved Support Vector Machine. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 35–44.
121. Farajzadeh, N.; Sadeghzadeh, N.; Hashemzadeh, M. Brain tumor segmentation and classification on MRI via deep hybrid representation learning. *Expert Syst. Appl.* **2023**, *224*, 119963. [CrossRef]
122. Padma, A.; Sukanesh, R. A wavelet based automatic segmentation of brain tumor in CT images using optimal statistical texture features. *Int. J. Image Process.* **2011**, *5*, 552–563.
123. Padma, A.; Sukanesh, R. Automatic Classification and Segmentation of Brain Tumor in CT Images using Optimal Dominant Gray level Run length Texture Features. *Int. J. Adv. Comput. Sci. Appl.* **2011**, *2*, 53–121. [CrossRef]
124. Ruba, T.; Tamilselvi, R.; Beham, M.P.; Aparna, N. Accurate Classification and Detection of Brain Cancer Cells in MRI and CT Images using Nano Contrast Agents. *Biomed. Pharmacol. J.* **2020**, *13*, 1227–1237. [CrossRef]
125. Woźniak, M.; Siłka, J.; Wieczorek, M.W. Deep neural network correlation learning mechanism for CT brain tumor detection. *Neural Comput. Appl.* **2021**, *35*, 14611–14626. [CrossRef]
126. Nanmaran, R.; Srimathi, S.; Yamuna, G.; Thanigaivel, S.; Vickram, A.S.; Priya, A.K.; Karthick, A.; Karpagam, J.; Mohanavel, V.; Muhibbullah, M. Investigating the Role of Image Fusion in Brain Tumor Classification Models Based on Machine Learning Algorithm for Personalized Medicine. *Comput. Math. Methods Med.* **2022**, *2022*, 7137524. [CrossRef]
127. Burns, A.; Iliffe, S. Alzheimer's disease. *BMJ* **2009**, *338*, b158. [CrossRef]

diagnostics

MDPI

Review

Cervical Cancer Detection Techniques: A Chronological Review

Wan Azani Mustafa [1,2,*], Shahrina Ismail [3], Fahirah Syaliza Mokhtar [4], Hiam Alquran [5] and Yazan Al-Issa [6]

[1] Faculty of Electrical Engineering Technology, Campus Pauh Putra, Universiti Malaysia Perlis, Arau 02600, Perlis, Malaysia
[2] Advanced Computing (AdvComp), Centre of Excellence (CoE), Universiti Malaysia Perlis, Arau 02600, Perlis, Malaysia
[3] Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai 71800, Negeri Sembilan, Malaysia
[4] Faculty of Business, Economy and Social Development, Universiti Malaysia Terengganu, Kuala Nerus 21300, Terengganu, Malaysia
[5] Department of Biomedical Systems and Informatics Engineering, Yarmouk University, 556, Irbid 21163, Jordan
[6] Department of Computer Engineering, Yarmouk University, Irbid 22110, Jordan
* Correspondence: wanazani@unimap.edu.my

Abstract: Cervical cancer is known as a major health problem globally, with high mortality as well as incidence rates. Over the years, there have been significant advancements in cervical cancer detection techniques, leading to improved accuracy, sensitivity, and specificity. This article provides a chronological review of cervical cancer detection techniques, from the traditional Pap smear test to the latest computer-aided detection (CAD) systems. The traditional method for cervical cancer screening is the Pap smear test. It consists of examining cervical cells under a microscope for abnormalities. However, this method is subjective and may miss precancerous lesions, leading to false negatives and a delayed diagnosis. Therefore, a growing interest has been in shown developing CAD methods to enhance cervical cancer screening. However, the effectiveness and reliability of CAD systems are still being evaluated. A systematic review of the literature was performed using the Scopus database to identify relevant studies on cervical cancer detection techniques published between 1996 and 2022. The search terms used included "(cervix OR cervical) AND (cancer OR tumor) AND (detect* OR diagnosis)". Studies were included if they reported on the development or evaluation of cervical cancer detection techniques, including traditional methods and CAD systems. The results of the review showed that CAD technology for cervical cancer detection has come a long way since it was introduced in the 1990s. Early CAD systems utilized image processing and pattern recognition techniques to analyze digital images of cervical cells, with limited success due to low sensitivity and specificity. In the early 2000s, machine learning (ML) algorithms were introduced to the CAD field for cervical cancer detection, allowing for more accurate and automated analysis of digital images of cervical cells. ML-based CAD systems have shown promise in several studies, with improved sensitivity and specificity reported compared to traditional screening methods. In summary, this chronological review of cervical cancer detection techniques highlights the significant advancements made in this field over the past few decades. ML-based CAD systems have shown promise for improving the accuracy and sensitivity of cervical cancer detection. The Hybrid Intelligent System for Cervical Cancer Diagnosis (HISCCD) and the Automated Cervical Screening System (ACSS) are two of the most promising CAD systems. Still, deeper validation and research are required before being broadly accepted. Continued innovation and collaboration in this field may help enhance cervical cancer detection as well as ultimately reduce the disease's burden on women worldwide.

Keywords: cervix; tumor; review; CAD

Citation: Mustafa, W.A.; Ismail, S.; Mokhtar, F.S.; Alquran, H.; Al-Issa, Y. Cervical Cancer Detection Techniques: A Chronological Review. *Diagnostics* **2023**, *13*, 1763. https://doi.org/10.3390/diagnostics13101763

Academic Editor: Dechang Chen

Received: 3 May 2023
Revised: 12 May 2023
Accepted: 15 May 2023
Published: 17 May 2023

1. Introduction

In 2020, cervical cancer recorded 604,127 new cases and death in 341,831 cases, according to the Global Cancer Observatory (GCO) [1]. In Malaysia, cervical cancer is the fourth most common cancer among women, accounting for around 1740 newly diagnosed cases and 991 yearly fatalities in 2020 [2]. Every year, between 2000 and 3000 cases of cervical cancer are hospitalized in Malaysia, according to the Ministry of Health (MoH). The majority of these cases come late in the course of the disease. Malaysia's mortality rate from cervical cancer is more than twice as high as that of the United Kingdom, the Netherlands, and Finland. The mortality rate has not decreased despite the implementation of screening programs and immunization campaigns against cervical cancer. The economic burden of cervical cancer is significant. In Malaysia, managing cervical cancer (from prevention to handling invasive diseases) costs around RM 312 million (USD 76 million). The majority of this (67%) goes towards treating aggressive cancer patients [3]. Pap smear screening is employed for early cervical cancer detection. The most crucial step is analyzing the Pap smear slide, and the identification of any condition or disease is crucial in order to administer the appropriate treatment [4,5]. Additionally, the Pap smear diagnostic reaction to a medication or treatment must be viewed or measured for clinical research. Clinically, microscope images are frequently utilized to diagnose Pap smear results. The sample images in the traditional approach, which involves taking a sample image under a microscope, run the risk of blurring effects, noise, shadows, lighting issues, as well as artifact issues on the images of thin smears [6,7]. Images from a Pap smear may have noise or other artifacts. Images from Pap smears may have poorer quality owing to noise or low contrast. Since the diagnosis relies on an individual, there are risks associated with the conventional procedure that might result in incorrect findings. A woman's cervix is where cervical cancer first develops. The female reproductive system is depicted in Figure 1 [8]. It happens as a result of abnormal cervix cell growth [9]. The cervix and tissues nearby, as well as organs consisting of the liver or lungs, will be invaded by this. Human papillomavirus (HPV) infection is linked to an increased risk of generating abnormal cells. Abnormal menstruation, irregular menstruation, heavy menstruation, weight loss, pelvic pain, and vaginal discomfort are the initial indications of cervical cancer.

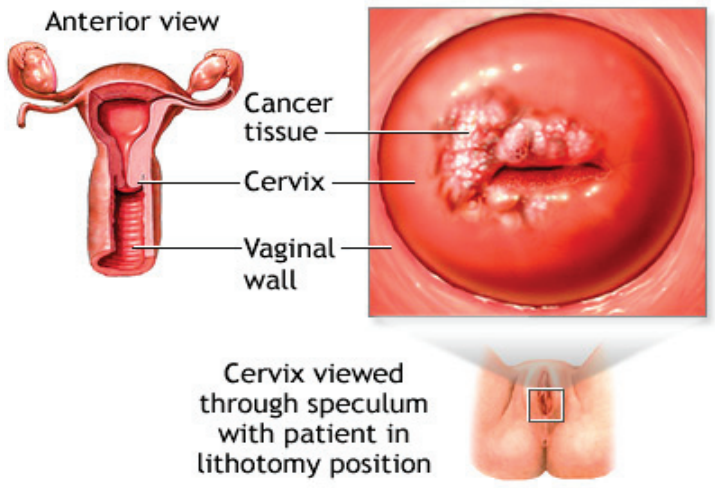


Figure 1. Female reproductive system [10].

Cervical cancer is caused by a group of viruses called HPV. Having sexual activity with another person may transmit HPV. There is evidence that HPV plays a role in the occurrence of penis, vagina, vulva, and anus cancers. There are more than 100 types of HPV, and HPV types 16 and 18 account for approximately 70% of all cervical cancer cases globally [11]. All women ranging in age from 25 to 74 are invited to screening tests. There are various methods to screen the cervical lining using a colposcopy, which is used to

magnify the area that the doctor wants to check after inserting the speculum into the vagina to check both the vagina and the cervix [12].

Early detection of cervical cancer is crucial since late diagnosis reduces the chance of survival in the entire world's female population [13]. According to Logeswaran (2020), 90% of women with cervical cancer diagnoses in low- and middle-income countries such as India may die unexpectedly as a consequence of inadequate detection, early diagnosis, effective screening, and treatment [14]. J. Lu et al. (2020) conducted a similar study and discovered that early screening is the most successful strategy for reducing the worldwide cervical cancer burden. Nonetheless, because of a lack of information, limited access to medical facilities, and prohibitively costly processes in developing countries, vulnerable patient populations are unable to bear routine examinations [15].

It may be diagnosed using a variety of screening tests, but the Papanicolaou smear test, which employs cell cytology, is the most common. It is a reliable method for detecting cervical cancer, although there is always a possibility of misinterpretation owing to human observational mistakes [16,17]. According to a study conducted in the medical field by Jaya and Latha (2019), image processing plays a crucial role in making the correct choice by utilizing a variety of techniques and algorithms. However, it is difficult to detect Pap smear images through microscopes. Traditional cervical cell screening also relies heavily on the pathologists' experience, which has the disadvantages of poor efficiency as well as low accuracy. Cervical cancer cells do not differ much in texture or color from normal cells, making their detection with smear tests very difficult [18].

However, cone biopsy screening is used when an abnormal cell is suspected in the cervix in order to detect it early. The most common screen test, as well as the Pap smear, also called the Papanicolaou test, is based mainly on using a brush to remove a small part of the lining tissue and checking it under microscopic levels to see if there are changes in the cell. This type of test can be used to discover if there is an infection or inflammation in the cervix or the presence of the HPV virus. The resultant images that have been obtained are called Pap smear images, which form a huge factor in early cervical cancer detection as well as classification. The new method for screening is based on the detection of HPV absence or presence [19]. Much research is carried out on the detection and classification of this type of cancer utilizing nanotechnology and building a biosensor to detect HPV, as well as using Pap smear images to detect and classify abnormal cells utilizing the benefits of deep and machine learning (ML) techniques. Other research focused on electrical impedance matching of affected signals with a 3D finite element model for cancer and non-cancerous cells. Cervical cancer affects the female reproductive system and is strongly associated with HPV infection, obesity, smoking, and sexually transmitted diseases (STDs). Manual Pap tests (Papanicolaou tests) are widely used for the early detection of cancer, but they are costly, stagnant, and highly dependent on the pathologist's expertise. Several computer aided diagnostic (CAD) systems were developed to automatically detect cervical cancer. Developing automatic prediction models to identify vulnerable patients can improve the efficacy of screening programs and eliminate inconsistencies and subjectivity resulting from cytopathologists' lack of expertise.

2. Materials and Methods

The primary goal of this study is to explore and understand the methodology of cervical cancer detection around the world between 1996 and 2022. The purpose of the current narrative analysis is to respond to the primary research question: (1) What types of cervical cancer detection have been proposed around the world? (2) How effective were computer-aided diagnostics for the Pap smear screening process? Contrary to that, cervical cancer detection has evolved significantly over the years, with several different techniques now available. The Pap smear test remains the most frequently employed method. Still, newer techniques such as visual inspection with acetic acid (VIA) and HPV testing, as well as Lugol's iodine, are becoming more widely used. Early detection is the key to successful treatment and improved outcomes, and women should undergo regular cervical

cancer screening according to recommended guidelines. In addition, this part discusses the requirement for a comprehensive evaluation of the cervical cancer situation. The outline of this review paper consists of three sections: Section 1 discusses an introduction and related research, and Section 2 describes the review data. The conclusions of this research are discussed in Section 3.

A method for obtaining the literature is shown in Table 1. The systematic review approach comprises three primary phases that were employed to determine the many relevant publications for this study. The initial phase is keyword recognition and the search for connected, related phrases utilizing the encyclopedia, dictionaries, and thesaurus, as well as prior research. Therefore, search strings were developed for the Scopus database once all pertinent terms were chosen. Considering literature (research papers) is the main source of pertinent information, it was the initial criterion. It also covers the exclusion of conference proceedings, chapters, books, book series, meta-synthesis, meta-analysis, reviews, and systematic reviews from the present research. Additionally, the review was limited to English-language studies only. A total of 108 publications were chosen in accordance with particular parameters.

Table 1. The specification for primary data searching.

Keyword	**Cervix, Cervical, Cancer, Tumor, Detect, Diagnosis**
Inclusion	Article, Journal, English, computer science, and engineering
Exclusion	Pure medicine, review article, other languages
Final Search String (Scopus)	TITLE ((cervix OR cervical) AND (cancer OR tumor) AND (detect* OR diagnosis)) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "COMP")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (SRCTYPE, "j"))
Number of Primary Article	108

Figure 2 represents the number of documents about cervical cancer per year. Obviously, interest in this topic started in 1996 with only one paper, and no production from 1997 to 2001 appeared in other documents. The settlement of ignorance was shown from 2002 to 2007, and two documents appeared in 2008. The steady increasing pattern appeared from 2009 to 2011. The sharp growth appeared from 2009 to 2015. In 2015–2018, there were swings between increasing and decreasing, but the average number was around eight documents per year. The total number sharply increased from 2018 to 2022 to be the mean of around 15 documents per year as well. That reflects people's consideration of the danger of cervical cancer as well as the significance of research to build a solid understanding of the nature of the disease and the tools to overcome or reduce its impacts on women.

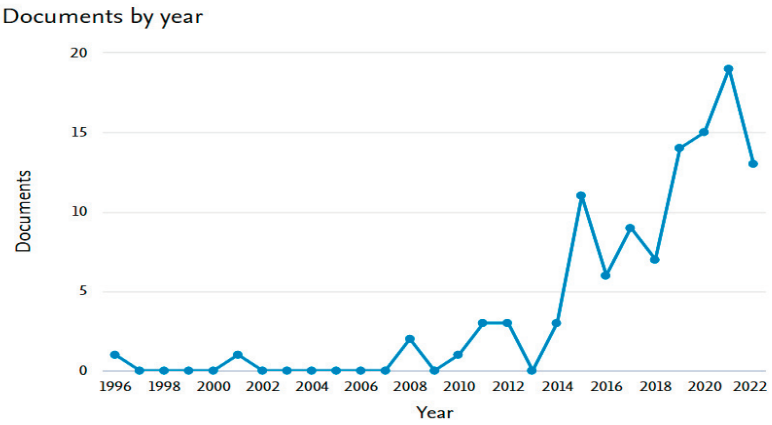


Figure 2. Number of documents per year.

3. Review of the Study

3.1. 1996–2015

Many studies have been conducted in the past to investigate cervical cancer diagnosis. Worldwide research is being conducted by doctors to better understand cervical cancer, how to prevent it, how to cure it, and how to provide treatment for those who have been diagnosed with the disease. For example, in 1996, an innovative method for the creation of segmentation and diagnostic algorithms for biomedical image analysis was given by [20]. In this case, a prototype expert system was created to give gynecologists a reliable and objective tool. Moreover, a collection of knowledge sources was created using specialized image-analysis methods. The robust control method employed by the expert system reduces the need for domain-specific control knowledge and has been shown to efficiently identify cervical cancer. The composition of segmentation and diagnostic methods for biomedical image analysis was also discussed in this paper, employing a new technique.

After many years, cervical cancer diagnosis evolved due to technological development. Following that, in 2001 [21], it was stated that the principal component analysis (PCA) in the wavelet domain delivers robust novel features with regard to the non-invasive detection of cervical intraepithelial neoplasia (CIN) employing fluorescence imaging spectroscopy. The term "principal wavelet components" (PWCs) refers to these characteristics. Average accurate classification rates for five cervical tissue classes—low-grade dysplasia (CIN 1), squamous, columnar, and metaplasia—as well as a fifth class for other unidentified tissue types, blood, and mucus—were 95% when PWC characteristics were employed as inputs to a 5-class NN. Apart from these [22], we presented a new technique to determine cervical cancer employing microwaves to measure the dielectric properties of the smear at microwave frequencies. This measuring approach is easy, and the smear collection is non-surgical and painless. The findings propose another option to the Papanikolaou or Papanicolaou tests and demonstrate a new technique for detecting cervical cancer using microwave measurement that may offer a less invasive alternative to these surgical procedures for detecting the disease.

On the other hand, in vivo, cervical dysplasia and cancer detection utilizing model-based analysis of reflectance and fluorescence spectra have been proven [23]. Here, a theory-based diffusion model is employed along with two analytical methods for calculating reflectance spectra that are contrasted with Monte Carlo simulations. A diagnostic algorithm is also created and tested utilizing cross-validation based on these obtained parameters. This algorithm's sensitivity/specificity for each measurement in comparison to the gold standard of histopathology are 85/51%. The accuracy described in previous research using optical technology to identify cervical cancer and its precursors corresponds to this.

Meanwhile, in [24], a quantitative colposcopic imaging system for early cervical cancer diagnosis is assessed in a clinical study. The cervix of living human beings is employed to assess the kinetics of the acetowhitening process in order to obtain diagnostic information. The imaging method relies on 3D active stereo vision as well as motion tracking. It was possible to distinguish between normal tissue and HPV-infected tissue, as well as low-grade and high-grade CIN lesions, utilizing a diagnostic algorithm with 91% SE and 90% SP. The findings show that the quantitative colposcopic imaging system may be able to deliver unbiased screening and diagnostic information for the early detection of cervical cancer.

Additionally, [25] immobilized anti-HPV18 and *E. coli* O157: H7 antibodies on magnetic silica-coated Fe_3O_4 for early diagnosis of cervical cancer as well as diarrhea. Uncoated Fe_3O_4 nanoparticles having a 9–16 nm average diameter as well as a saturation magnetization of around 66 emu/g were first prepared using the co-precipitation method. The findings revealed that magnetic SiO_2-coated Fe_3O_4 nanoparticles could be an auspicious contender for diagnosing cervical cancer at an early stage, specifically with high accuracy.

In 2011, [26] employed an optoelectronic method to detect CIN as well as cervical cancer. The pNOR number and the sensitivity/specificity of the optoelectronic approach were shown by the authors to be correlated. The specificity of the optoelectronic approach

was calculated to be 65.70% for LGSIL and 90.38% for HGSIL and cervix squamous cell carcinoma. The optoelectronic technique utilized to validate the absence of cervical pathology was assessed to have a 78.89% specificity. Here, CIN, which exists in the squamous epithelium as well as squamous cell carcinoma of the cervix, is easily detected using the optoelectronic approach.

In the same year, [27] investigated the hWAPL histological expression value assessment in the cytological as well as histological diagnosis with regard to cervical intraepithelial neoplasia and cervical cancer. The expression intensity of hWAPL protein in the HSIL group, LSIL group, ASCUS group, and ASC-H group was obviously greater than that in the NILM group ($p < 0.05$), and the expression intensity in the ASCUS group and ASC-H group was higher than that in the LSIL group ($p < 0.05$). Furthermore, in the ASCUS and ASC-H groups, the frequency of SCC + CIN III was above 50%. Therefore, hWAPL may be a promising candidate for diagnosing low-grade CIN. Furthermore, the histological expression of hWAPL is consistent with the cervical lesions' cytological type.

A year later, in 2012, in order to enhance cervical cancer risk classification, [28] investigated the automated detection of dual p16/Ki67 nuclear immunoreactivity in liquid-based Pap tests. Algorithms were created to digitize and examine smears stained with p16 as well as Ki67 antibodies. The nuclear mask was produced employing a gradient-based radial symmetry operator along with adaptive symmetry image processing. This was subsequently followed by the extraction of attributes from each nucleus, such as pixel data as well as immunoreactivity signatures. The quantitative analysis of immunoreactivity offered by the further emphasis on classified nuclei, according to the authors, may have a positive influence on the effectiveness and screening results of the Pap test.

In the same year, which is 2012 [29], a new technique was proposed to construct a tumor probability map while gradually determining the boundaries of an organ of interest on the basis of the accomplished nonrigid transformation. The technique dealt with the difficulties of considerable tumor regression and its impact on nearby tissues. Findings indicate that the suggested technique greatly surpasses the current registration algorithms and reaches a precision equivalent to manual segmentation. Additionally, there is excellent agreement between the suggested method's tumor detection results and manual delineation by an experienced doctor.

Moreover, in [30], blood and urine samples from cervical cancer patients were collected, and their fluorescence emission spectra (FES) as well as Stokes shift spectra (SSS) were contrasted to those of normal controls. Both spectra demonstrated that in cervical cancer patients, the relative levels of biomolecules, which include flavin, nicotinamide, adenine dinucleotide, collagen, and porphyrin, were out of balance. The author also stated that this is the first study on FES and SSS of blood and urine samples from patients with cervical cancer that provides a sensitivity of 80% as well as a specificity of 78%.

A total of 2 years later, in 2014 [31], it was proposed to use time-resolved blood component spectra to identify cervical cancer. Porphyrin served as the biomarker indicative of cancer in this instance, with samples from cancer patients having fluorescence decay times that are 60% greater than those from normal controls. A randomized set of samples from cancer patients and controls ($n = 27$ in total) could be categorized with sensitivity (92%) and specificity (86%) using these parameters.

Utilizing reduced graphene oxide–tetraethylene pentamine as electrode materials and distinct redox probes as labels [32], this suggested simultaneous electrochemical detection of cervical cancer indicators in the same year. In accordance with the peak current change of neutral red and thionine prior to and following the antigen-antibody reaction, the immunosensor was constructed with a sandwich structure. According to the findings, the immunosensor exhibited a broad linear range, a small detection limit, high reproducibility, and stability. Furthermore, the technique has been employed successfully to examine serum samples.

Moreover, [33] utilized extracted intrinsic fluorescence as well as PCA to identify the advancement of cervical cancer. Here, along with the intrinsic fluorescence, the effective-

ness of PCA in separating the aggregate behavior from smaller associated clusters in a dimensionally diminished space is tested. By closely observing the sectorial behavior of the dominant eigenvectors of PCA, it is possible to determine the various activities of the dominant fluorophores, flavins, nicotinamide adenine dinucleotide, collagen, and porphyrin of various classes of precancers. The Mahalanobis distance was also computed utilizing the scores of the chosen major components in order to better categorize the various grades.

A year later, [34] presented a method for colposcopic images-based automated cervical cancer diagnosis. Here, abnormal and normal tissue are distinguished using wavelet and statistically based attributes. The wavelet-decomposed image is employed to obtain the wavelet energies. The feature vector produced from the combination of these features is then applied to the detection. The segmented cancer region demonstrates that the suggested fusion technique is capable of identifying the cancer-affected region with more accuracy over the wavelet, along with statistical features-based approaches.

In addition, a hybrid classifier-based computer-aided detection (CAD) of cervical cancer utilizing Pap smear images was also suggested by [35] in 2015. It is utilized to divide the cell image from the test Pap smear into normal and dysplastic cell images. Following that, morphological techniques are employed to identify and segment the abnormal cell region. On images from databases with free access to the public, the suggested technique is evaluated. A unique illumination correction and intensity normalization approach on cervigrams was put out by [36] in the same year in order to aid in the early detection of uterine cervical cancer. In light of our study's results, we draw the conclusion that the peak of the squamous epithelium (SE) region's intensity distribution and the peak of the entire cervix region are significantly associated.

Furthermore, by using the nested structure of its data to extract patient-level features from the cell-level data, utilizing a statistical model that takes advantage of the hierarchical data structure, and classifying the cellular level [37], it executed comparative research on three primary methods for solving problems. With an estimated 61% sensitivity and 89% specificity on independent data, the optimal method was to classify at the cellular level and count the number of cells with a posterior probability larger than a threshold value. In addition, recent advancements in statistical learning make it feasible to reach great accuracy. Apart from that, new clinical studies that support the use of HPV E6/E7 mRNA as a marker in advanced cervical cancer screening programs were reported in 2015 [38]. The authors give a general review of the research study sample size, age, recruitment setting, HPV mRNA, and HPV DNA tests. It was demonstrated by the pooled evaluation of clinical research that HPV mRNA may be a useful diagnostic biomarker. To draw a firm conclusion, however, further research must be conducted.

On the other hand, in the same year, [39] investigated the degree of squaraine dye aggregation that affects the strength of surface-enhanced Raman signal scattering (SERS) after adsorption on a gold surface that has been nano-roughened. When chemisorbed on spherical gold nanoparticles, the SQ2 (mono lipoic acid appended), SQ5 (conjugated with hexyl and dodecyl side chains), and SQ6 (conjugated with hexyl and dodecyl side chains) squaraine derivatives demonstrated a substantial rise in Raman scattering in the fingerprint region. HeLa cells demonstrated pronounced SERS mapping intensity and selectivity towards the cell surface and nucleus after further conjugating this nanotag with monoclonal antibodies that targeted overexpressed receptors, EGFR and p16/Ki-67, in cervical cancer cells.

Subsequently, [40] proposes a system for automatically classifying and segmenting cervical cells. Radiating Gradient Vector Flow (RGVF) Snake is employed to separate the cytoplasm, nucleus, and background of a single cervical cell image. For system training, several cellular and nuclear properties are retrieved. Artificial neural networks (ANN) are employed to examine the dataset's ability to categorize seven distinct cell types and distinguish between abnormal and normal cells. The clinical research on styping identification of HPV infection using microarrays from paraffin-embedded tissues of precursor lesions as well as cervical cancer was also explored by [41]. This led to the identification of the

prevalence and type distribution of HPV in cervical cancer and CIN in Jiangsu, China. The findings indicate that Jiangsu's (China's) high rate of HPV 16, 18, 33, 31, and 58 warrants further notice. It has significant repercussions for the effective administration of the HPV vaccination and the selection of testing techniques.

Apart from that, [42] examined the fractal dimension of AFM images of human cervical epithelial cells at various stages of cancer growth to evaluate the early detection of cervical cancer. Individual human cervical epithelial cells at three phases of cancer progression—normal, immortal (pre-malignant), and carcinoma cells—were examined using the AFM HarmoniX modality by the author. The authors were successful in distinguishing between abnormal and normal cells by utilizing AFM to examine the surface characteristics of human cervical epithelial cells. This technique could supplement current techniques to improve the accuracy of diagnosis.

Moreover, [43] proposed using nanotechnology and biomarkers for cervical cancer's early detection and treatment. Nanomaterials are special in their optical, physical, and electrical characteristics, which has made them particularly advantageous for sensing. Cancer biomarkers, which are employed as targets in the detection and monitoring of cancer, are mostly composed of RNA fragments, DNA fragments, antibody fragments, and proteins. In a few decades, it is expected to be feasible to identify cancer at a very early stage, giving a significantly greater probability of treatment.

Subsequently, [44] describes an ultrasensitive electrochemical immunosensor for accurate detection of p16 and shows how effectively it performs when used with patient cell lysates to detect solubilized p16 protein. Furthermore, the authors also reported that the suggested immunosensor successfully detected raised p16 levels in cervical swab samples taken from 10 patients who had received positive results from a standard Pap smear test, demonstrating that electrochemical immunosensors hold great potential for the early detection of cervical cancer in a clinical setting.

3.2. 2016–2018

Several studies tried to diagnose cervical cancer using various techniques. For instance, in 2015, Yulan Wang et al. recommended the use of fluorescence lifetime imaging microscopy (FLIM) for the early detection of cervical cancer. They discovered that the lifetime of cancerous cells was shorter compared to normal cells. They recommend FLIM as a highly precise and specific method that can detect the occurrence of precancerous as well as cancerous cells quickly [45].

In 2016, S. Athinarayanan et al. [46] suggested an automatic multistage cervical cancer diagnostic system using Pap smear images (obtained from the Herlev dataset described in Table 2) and machine learning (ML) methods. In the preprocessing stage, images were denoised, intensity and texture features were extracted, and finally, images were differentiated using SVM into normal and abnormal classes. They succeeded in detecting cervical cancer with 94% accuracy. Moreover, Anousouya Devi et al. [47] developed an image analysis algorithm to replace time-consuming Pap smear screening tests. The authors discussed a variety of segmentation algorithms and feature extraction techniques with regard to the efficient segmentation of Pap smear slides.

Furthermore, Xianfeng Xu et al. [48] investigated the value of PET/CT scanning in detecting cervical carcinoma in 51 patients. Note that PET/CT diagnosis capability is superior to the classical FIGO discrimination technique. For example, PET/CT detected primary tumors with 84.31% accuracy, 80.77% specificity, and 88% sensitivity. On the other hand, it detected lymph nodes with 76.47% accuracy, 71.43% specificity, and 82.61% sensitivity. Subsequently, Jose Amaya et al. [49] designed a high-stability voltage current source for the recognition of cervical cancer using electrical bio-impedance spectroscopy. Here, the medical kit they designed was compatible with international standards. Finally, Rizanda Sobar et al. [50] determined seven behavior features and surveyed 72 respondents (including 22 cancer patients) in Indonesia. They used two machine learning (ML) techniques, particularly logistic regression (LR) and Naïve Bayes, to forecast the risk of

becoming a cervical cancer patient. With respect to accuracy, Naïve Bayes outperformed LR (91.67% compared to 87.5%), and with respect to AUC, LR outperformed Naïve Bayes (0.97 compared to 0.96).

Table 2. Pap smear image classification in the Herlev dataset [9].

Cell	Class Name	Cell Count	Sub-Total
Normal	Normal Superficial Squamous	74	242
	Normal Intermediate Squamous	70	
	Normal Columnar	98	
Abnormal	Carcinoma In Situ	150	675
	Light Dysplastic	182	
	Moderate Dysplastic	146	
	Severe Dysplastic	197	
	Total	917	917

One year later, Irvin Sitompul et al. [51] conducted a descriptive qualitative study using a questionnaire to evaluate the knowledge of aged women in the Cakung health center regarding the early detection and prevention of cervical cancer. They concluded that knowledge of the Human Papilloma Virus (HPV) vaccine is weak. Meanwhile, Branislava Jeftic et al. [52] presented a cervical cancer detection method relying on optomagnetic imaging spectroscopy (OMIS) and compared the findings utilizing unstained and stained Papanicolaou smears. Using the Naïve Bayes classifier, they separated the samples into four groups: the II Pap group (normal cells), the III Pap group (abnormal cells), and the IV and V Pap group (cancerous cells). Unstained sample classification with Naïve Bayes achieved 96% accuracy, whereas stained sample classification achieved 85.18% accuracy. Apart from these, Abdullah Iliyasu et al. [53] proposed a quantum hybrid technique that uses quantum particle swarm optimization (QPSO) for selecting 7 out of 17 features, as well as a fuzzy KNN for the classification of cervical cells in smeared images. They used 917 images from the Herlev dataset and achieved 86% recall, 85% precision, and F1 score of 85%. On the other hand, Wen Wu et al. [54] employed three SVM-based combinations for the diagnosis of cervical cancer. All four target variables were identified, and the performance of SVM was superior to SVM-RFE and SVM-PCA. SVM achieved high precision using all 30 features, but the computation cost was high. The authors showed that the SVM-RFE and SVM-PCA gave comparable performance to the SVM using only 8 features, improving classification time considerably.

In the same year, Katrin Carow et al. [55] presented evidence that the incorporation of HPV-DNA into the host genome is an initial step in the formation of cervical cancer. They recommend using viral-cellular junction sites as biomarkers when examining circulating tumors. Meanwhile, Vidya Kudva et al. [56] proposed an image-processing approach that can be used as an image treatment step in any cervix cancer detection system. They presented a cervix region segmentation method and detected specular reflections with high precision, irrespective of lighting conditions and color variations. Apart from these, Guanglu Sun et al. [57] suggested an ML framework relying on relief feature selection and a Random Forest (RF) classifier to diagnose cervical cancer. They used 917 Pap smear images obtained from the Herlev dataset together with 10-fold cross-validation to perform binary classification. RF outperformed LR, C4.5, and Naïve Bayes classifiers with 94.44% accuracy and 0.9804 AUC using 13 features. In addition, Rubina Shaikh et al. [58] compared two optical modalities, particularly Raman (RS) and Diffuse Reflectance Spectroscopy (DRS), in differentiating between normal and abnormal cells. One hundred forty-six recorded spectra (67 tumors and 79 normal) were analyzed using a combination of Principal Component and Linear Discriminant Analysis ML techniques. They used Leave One Out Cross Validation (LOOCV) and concluded that DRS is more suited for rural areas, whereas RS is suited for developing countries. Furthermore, Muljo et al. developed an online learning management

prototype to educate health workers and the public in Indonesia about early cervical cancer detection as well as treatment [59].

In 2018, Mithlesh Arya et al. [60] used SVM as well as ANN to classify single-cell images captured from Pap smear slides into benign and malignant tumors. The accuracy obtained using the suggested texture-based features exceeds that obtained using shape-based features. Additionally, the performance obtained using a combination of features was better than that obtained using a single feature. Using quadratic SVM, they achieved 99.5% accuracy, 99% sensitivity, and 99% specificity. Meanwhile, Ashutosh Sharma et al. [61] successfully employed fluoranthene-based yellow fluorescent lipid probes with respect to the detection of lipid droplets in cervical cancer tissues. FLUN-550 and FLUN-552 quantitatively detected the excess lipid accumulation and were really useful in the early diagnosis of human cervical cancer. Additionally, Kelwin Fernandes et al. [62] developed a supervised deep learning (DL) method to diagnose cervical cancer with high accuracy using the medical records of 858 patients. To study the impact of their architecture, they applied their methodology to different datasets and demonstrated that their efficiency is not limited to cervical cancer. They used a loss function for dimensionality reduction, achieving an AUC of 0.6875. Furthermore, Yueyue Jing et al. [63] established quick, highly sensitive, and highly specific label-free imaging and spectroscopy for the detection of cervical tumors compared to the traditional clinical staining method. They studied unstained tissues extracted from 38 patients and achieved 100% sensitivity and 91% specificity.

In the same year, Rocky Dillak et al. [64] suggested an early alarm system to diagnose cervical cancer based on a combination of chaos optimization and ridge polynomial neural networks. They achieved an accuracy of 96%, a sensitivity of 95.56%, and a specificity of 96.67%. Apart from these, Vidya Kudva et al. [65] manually extracted 102 images obtained during visual inspection with acetic acid; 42 images were pathologic, and the remaining 60 were negative. They used a shallow-layer CNN to discriminate between cancer and non-cancer lesions by automatically extracting features from 684 representative patches with 100% accuracy. Following that, Sherif Abdoh et al. [66] identified 32 risk factors to build a cervical cancer diagnosis framework. They employed two feature reduction techniques, namely Recursive Feature Elimination (RFE) and PCA. Furthermore, they used an RF classifier combined with the Synthetic Minority Oversampling Technique (SMOTE) to correctly classify cervical cancers. The obtained results were validated using 10-fold cross-validation, and SMOTE-RF outperformed SMOTE-RF-RFE and SMOTE-RF-PCA in detecting all 4 cancer groups.

3.3. 2019–2020

As artificial intelligence (AI) and image processing technology advance, we have reviewed progressively intelligent diagnosis tools that are being applied in cervical cancer screening. In this section, we offer a brief review of some methods available in the literature, starting with the year 2019 and progressing to the current cervical screening. Lavanya Devi et al. (2019), for instance, investigate the various automated methods for detecting abnormal cells in Pap images. Cancer screening commonly includes a Pap smear test and an acetic acid test. Cells from the vagina and cervix are extracted and analyzed under a microscope for the occurrence of an abnormal cell in a pap test. An acetic acid test is employed to identify the existence of abnormal cells by comparing the differences in characteristics between samples before and after the application of acetic acid. According to the report, automated screening has become more common than manual screening, given that the latter is inaccurate [67]. This method of screening has been endorsed in a study conducted by Abdullah et al. (2019), where computer-based algorithms are broadly employed in cervical cancer screening. In this research, a better cellular neural network (CNN) algorithm has been set up as a potential means of detecting cancerous cells in Pap smear images in real-time. For automated detection of cancerous cervix cells, a CNN built-in in MATLAB using templates that segment cell nuclei has been established. The

simulation findings demonstrate that our suggested CNN algorithm can automatically identify cervix cancer cells with over 88% accuracy [68].

Jaya and Latha (2019) introduced a technique for enhancing Pap smear images by comparing Power Law Transformation for Gamma Correction, Histogram Equalization in the Contrast Stretching algorithm, Contrast Limited Adaptive Histogram Equalization (CLAHE), and Shading Correction. To determine the performance of upgraded Pap smear images, the quality measurement NAC, SC, PSNR, and MSE values were determined. As a programming tool, MATLAB R2016a and ANN classification were used to assess the accuracy level of each feature extraction of the algorithm. The study concluded that CLAHE produced a decent result for enhancement, and the SGLDM feature extraction algorithm achieved 93% accuracy while utilizing ANN [69]. A review of the literature undertaken found that accurate recognition of cervical cancer cells is crucial for clinical diagnosis. A better approach built around the residual neural network is presented to increase the accuracy of diagnosis. However, these current algorithms are only enhanced by the use of low-level manual features. The findings of the experiments demonstrate that the lightweight deep model performs better than the current comparative models and may obtain a model accuracy of 94.1% when applied to the cervical cell data set [70]. Hence, as recommended by William et al. (2019), it is advantageous to construct a computer-assisted diagnostic tool to increase the accuracy and reliability of the Pap smear test. In this research, Pap smear image analysis was utilized to construct a tool for the automated detection and classification of cervical cancer. Scene segmentation was accomplished using a trainable Weka segmentation classifier, while a sequential elimination strategy was employed for debris rejection. While classification was accomplished utilizing a fuzzy C-means technique, feature selection was accomplished employing simulated annealing combined with a wrapper filter [71]. The research found that three distinct datasets—single-cell images, multiple-cell images, and Pap smear slide images from a pathology lab—were utilized to evaluate the classifier. For each dataset, overall classification accuracy, sensitivity, and specificity results of "98.88%, 99.28% and 97.47%", "97.64%, 98.08% and 97.16%", and "95.00%, 100% and 90.00%", accordingly, were attained. In comparison to the manual analysis, which takes between 5 and 10 min per slide, the suggested system can analyze a whole Pap smear slide in about 3 min.

Ref. [72] identified the relevant features in the cancer classification as well as optimized the model. The vital properties in the attribute list were explored using the binary cuckoo search optimization technique. The experimental findings demonstrate the greater performance of the Decision Tree (DT) classifier over all other classifiers, with accuracy increasing from 94.7% to 97% following cuckoo optimization. Another study conducted by Adem et al. (2019) discovered that softmax classification with a stacked autoencoder model, which was implemented for the first time in the cervical cancer dataset, performed better compared to other ML methods with an appropriate 97.8% classification rate. New techniques of diagnosis are described in this article in terms of patient diagnostic support systems, taking into account the interest in ML approaches in cancer research [73].

In the year 2020, a number of studies offered new screening methods, such as the Shot multiBox detector, which can accurately detect many items of multiple scales at the same time to solve the classic saliency cervical cancer diagnosis approach in ultrasound images. The study provides a new multi-saliency object detection model with an appended deconvolution module embedded within the residual attention module. Experiments demonstrate that the suggested diagnosis method beats comparable algorithms in terms of detection accuracy. It also improves the accuracy of cervical diagnosis by increasing detection performance for multi-saliency cervical cancer objects with small scales [74]. The Enhanced Johnson's Algorithm (EJA) was proposed by Ali et al. (2019) as the new shortest path for detecting cervical cancer-associated genes in the Protein-to-Protein Interaction (PPI) network for early cervical cancer diagnosis in their study. EJA was also adopted to find the shortest path between invasive and pre-invasive genes. The Bellman-Ford approach was used in EJA to reconstruct the path with a new iterative matrix, which successfully

reduced the elapsed time by omitting the negative cycles in the gene connection [75]. Huang et al. (2019) discovered that endogenous fluorophores in cells and tissues, such as diminished nicotinamide adenine dinucleotide (phosphate) (NAD(P)H) as well as flavin adenine dinucleotide (FAD), may be imaged by FLIM to illustrate the tissue morphology features, including the biomolecular variations in the microenvironment. It was shown that by monitoring the fluorescence lifetime of NAD(P)H as well as FAD in nearby healthy cervical tissues, benign uterine tumors with abnormal cell development, which include leiomyomas and adenomyosis, may be identified [76]. According to [18], cervical cancer is caused by morphological alterations in cells or dead nuclei in the cervix. The detection of abnormalities in cells necessitated a high-level digital image processing technique that included an automated, complete ML skill set. To split the cytoplasm as well as the nucleus from the cell, an innovative fuzzy-based approach has been proposed. KNN is instructed with the color and form attributes of the segmented cell units, and then it is used to classify unknown cervix cell samples. The cytoplasm, as well as the nucleus of the cervix cell, are given shape and color using the proposed technique.

Several other methods have been introduced in detecting this disease, such as automatic feature extraction and classification for acetic acid and Lugol's iodine cervigrams, as well as (2) methods for merging diagnosis/features of distinct contrasts in cervigrams for enhanced performance, which attained a sensitivity, specificity, and accuracy of 81.3%, 78.6%, and 80.0%, respectively [77]. A study reported that a novel immunosensor had been formed for quantitative detection with respect to the squamous cell carcinoma antigen (SCCA) in cervical cancer, built on surface-enhanced Raman scattering (SERS). The SCCA monoclonal antibody was combined with polydopamine resin microspheres covered with gold nanoparticles as capture substrates. Phosphate buffer (PBS) had a detection limit of 7.16 pg mL^{-1} and human peripheral blood had a detection limit of 8.03 pg LH^{-1}. The findings showed that the SERS immunoassay approach has a possibility for use in early cervical cancer screening and diagnosis [78]. Fuzzy Swallow Swarm Based Feature Selection (FSSBFS) has been introduced for the optimal selection of cervical cancer features. The proposed ISVM-FssBFS classifier is improved when compared to SVM and Multilayer Perception Classifier (MLP) classifiers. The cervical cancer samples are characterized by 32 risk factors and four target classes: Biopsy, Cytology, Schiller, and Hinselmann [79].

Early identification of CIN dramatically improved patient survival rates in the year 2020 [80]. Most cervical cancer detection algorithms rely on natural image object detection technologies, with only minor improvements made to account for the complex application scenario with respect to cervical lesion detection. The suggested method's sensitivity at four false positives per image as well as average precision are enhanced by 2.79 and 7.2%, respectively, when compared to the baseline (Retinanet) [81]. Chen et al. (2020) first established the feasibility of using CT imaging and radiomics to create a low-cost image marker for detecting LN metastasis in cervical cancer patients. Here, the model was trained to utilize a leave-one-case-out (LOCO) cross-validation strategy with a total accuracy of 76.4%. Li et al. (2020) proposed a DL framework with regard to the accurate identification of LSIL+ (which includes CIN and cervical cancer) employing time-lapsed colposcopic images. All of the fusion methods that are compared perform better than the automated cervical cancer diagnosis systems that are currently in place and utilize a single time slot. The best fusion strategy was discovered to be a convolutional graph network with edge features (E-GCN). A novel framework built around a strong feature Convolutional Neural Networks (CNN)-Support Vector Machine (SVM) model was presented to properly categorize the cervical cells, according to research by Dongyao Jia et al. (2020). On two distinct datasets, the suggested technique was assessed using the metrics of accuracy (Acc), sensitivity (Sn), and specificity (Sp). The outcomes suggested that the CNN-SVM model with strong features might be utilized to classify cells for early cervical cancer screening [82].

A potential technique for the diagnosis of cervical cancer with parametrial infiltration is the combination of whole-tumor dynamic contrast-enhanced MRI and texture analysis [83]. Ktrans, energy, and entropy work more effectively together than separately,

particularly when it comes to increasing diagnostic sensitivity. Fuzzy logic and adaptive neuro-fuzzy inference system (ANFIS) classification method-based cancer area detection and segmentation in cervical images were suggested by Ramasamy and Chinnasamy in 2020. Fuzzy logic is employed to identify the thick and thin edges, which are then combined using an image fusion approach at the pixel level. The suggested cervical cancer detection system has a classification rate average of 98.8%. In comparison to earlier suggested approaches for cervical cancer estimation, the CCPM result demonstrated more accuracy [84]. The sensitivity, specificity, and accuracy of the suggested cervical cancer segmentation methods presented in this paper are 98.1%, 99.4%, and 99.3%, respectively. A model for early cervical cancer prediction (CCPM) has been developed by researchers, utilizing risk indicators as inputs. In comparison to earlier suggested approaches for cervical cancer estimation, the CCPM results demonstrated more accuracy. For quick and effective action at the early stages of the disease, a mobile application that may gather information on cervical cancer risk factors and offer CCPM findings has been created [85].

Apart from that, [15] adopted a voting method that takes into account the issues with earlier research on cervical cancer. To assess the suggested procedure, several measures are implemented. According to the findings, the voting approach may be used to accurately forecast the chance of having cervical cancer. In comparison to previous techniques, the one that is being presented is more scalable and practical. The key finding by Singh and Goyal (2020) is the choice of the optimal ML algorithm with the maximum accuracy. Several algorithms were able to achieve up to 100%. Although a method such as LR with L1 regularization has a 100% accuracy rate, it consumes too much CPU time [16].

To effectively recognize the nucleus as well as the cytoplasm boundary of the Pap smear cell as a way to diagnose cervical cancer, an enhanced normalized graph cut with generalized data for enhanced segmentation (INGC-GDES) method was presented. In comparison to earlier methods, the suggested INGC-G DES mechanism leads to a 28% improvement in classification accuracy [13]. To the best of our knowledge, research has demonstrated the potential of Mueller matrix image processing as a unique strategy for the detection of cancer and precancer [86]. Sections of the human uterine cervix's normal and precancerous tissue were utilized in the study. The research explained the creation of a DNA-based electrochemical biosensor that is sensitive and selective for the early detection of HPV-18. As a proprietary, accurate, sensitive, and quick diagnostic approach for HPV 18 in the polymerase chain reaction (PCR) of actual samples, the suggested biosensor can be presented. On a screen-printed carbon electrode (SPCE), a nanocomposite of reduced graphene oxide (rGO) as well as multiwalled carbon nanotubes (MWCNTs) was electrodeposited [87].

A study conducted by Rehman et al. (2020) reported that an auto-assisted cervical cancer screening system is suggested that utilizes a CNN trained on the Cervical Cells database. The system provides better performance than its previous counterparts under various testing conditions. For the 2-class problem, the classification accuracy of SR, SVM, and GEDT is determined to be 98.8%, 99.5%, and 99.6%, respectively [17]. Validation of Association Rule Mining using the Test Train Approach (VARMTTA), a data-driven methodology, was put out by Logeswaran et al. (2020). Employing the train-test validation approach lowers the number of rules that are generated from the dataset. This technique makes use of conventional measures, including sensitivity, precision, and total accuracy [14]. According to Sahoo et al. (2020), using a common path interferometric setup, low-coherence backscattered images of precancerous cervical tissue sections were recorded. These low-coherence images were subjected to a two-dimensional multifractal detrended fluctuation analysis (2D MFDFA) in order to examine the fluctuations in their fractal nature. The RI fluctuations showed long-range relationships, and multifractality was shown to be greater for cervical cancer with higher grades. It was discovered that normal and CIN-I, CIN-I and CIN-II, and normal and CIN-II had specificities and sensitivities of 94%, 88%, 93%, 96%, and 100%, respectively [88].

3.4. 2021–2022

B. Chitra and S. S. Kumar [89] reviewed the most recent soft computing techniques for detecting and classifying the most updated algorithms in current research. It is considered a literature review of the most common classification techniques for cervical cancer up to 2021. On top of that, Md. MamunAli et al. [90] employed clinical data for early cervical cancer detection. They applied a variety of data transformation techniques, such as Z-score, log, and sine functions, in addition to feature selection methods for specifying the most priority features for early detection of cervical cancer. Their results concluded that the logarithmic transformation feature is the best for biopsy data. On the other hand, sine is the best for cytology. However, the combination of sine as well as logarithmic is the best for the Hinselmann dataset, but for the Schiller dataset, the Z-score performance is the best. The classifiers utilized in this study are RF, Random Tree (RT), and instance-based nearest neighbor classifiers. For better performance, B. Chitra and S. S. Kumar [91] utilized the DL structure DesnNet 121 to classify Pap smear images. They apply various augmentation techniques to the dataset. The DL structure is optimized using the Mutation-based Atom Search Optimization (MASO) algorithm, which is employed to enhance the hyperparameters of DensNet121, for instance, the learning rate, the number of neurons in the dense layer, the number of epochs, patch size, and others. This approach obtains the best accuracy among existing techniques, which reaches 98.3%. Attempting other methods, such as recurrent neural networks, Zhang et al. [92] discussed the existing screening methods for cervical cancer that are based mainly on separated cells. Therefore, any misclassified cell causes poor accuracy. To overcome these limitations, they proposed a method that combines Long-Short Term Memory (LSTM) with a full CNN as well as fuzzy nonlinear regression. They exploited the time series method for improving cervical screening for cancer. Their procedure was accurate to 98.3%.

Sohely Jahan et al. [93] proposed an approach that is described in Figure 3. As it is clear, the raw cervical dataset is cured by outlier removal, cleaning methods, and excluding the records that have missing values. Various feature selection principles are utilized, for instance, Chi-square and RF, to find the most significant features. The selected features are scaled and split into 70:30 to train and test various types of classifiers such as Random Forest (RF), Logistic Regression (LR), Support Vector (SV), Multi-Layer Perceptron (MLP), Decision Tree (DT), Gradient Boosting (GB), K-nearest neighbour (KNN), and AdaBoost (AB) classifiers. MLP performed the best among all with a variety of features. On the other hand, all classifiers have almost the same high performance on 25 selected features.

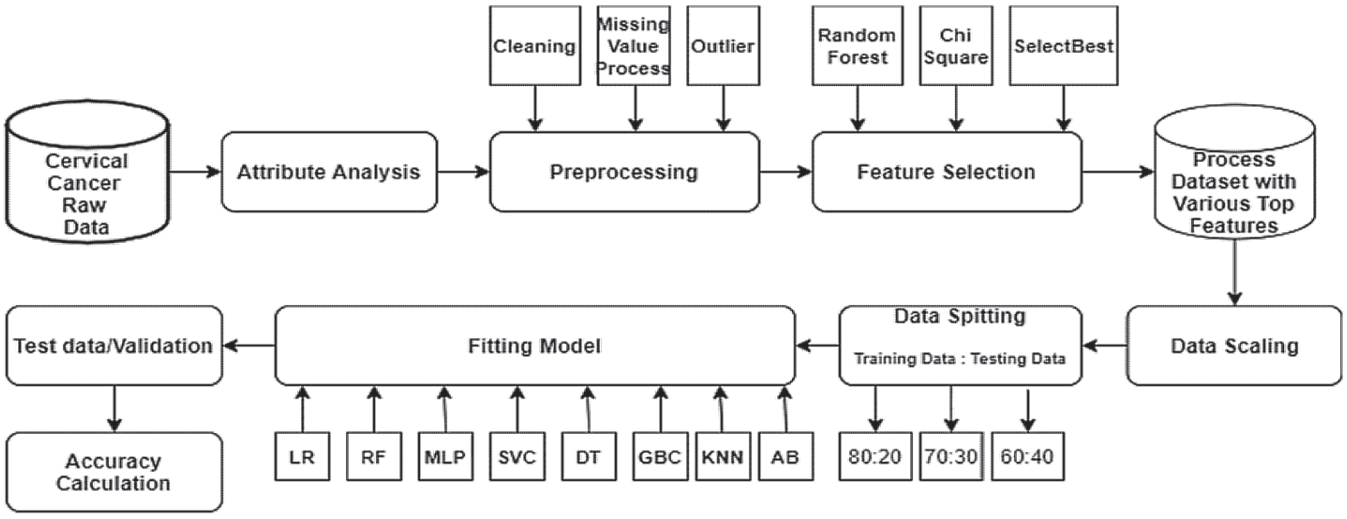


Figure 3. Automated invasive cervical cancer disease detection at an early stage via an appropriate ML model.

The research aims to improve accuracy with a reliable system. Therefore, Lei Cao et al. [94] suggested a more accurate system for detecting cervical cancer. Their method is based on a feature pyramid network to automatically classify cytological images by detecting abnormal cells. Their distinguished model has two features: the first is the reading way of the cervical cytology images, which is the same as pathologists, and the second is detecting abnormal cells at different scales using a multi-scale region-based fusion network. Their designed approach builds on clinical knowledge about abnormal cervical cells based on their shapes and sizes. The performance of their approach is better than the DL approach. Their highest accuracy was 95.8% on the independent dataset. Their process is accurate and quick, and their diagnosis time is 0.04 s per image, which is faster than pathologists' diagnoses. For dealing with big-size images such as 1000 × 1000 pixels, Antoine Pirovano et al. [95] proposed the classification under regression constraints. Their experiment enhanced the sensitivity by up to 80% for localizing malignancy in whole slide images. The proposed approach can be integrated with the pathology laboratory system to improve prediction. Figure 4 illustrates their approach.

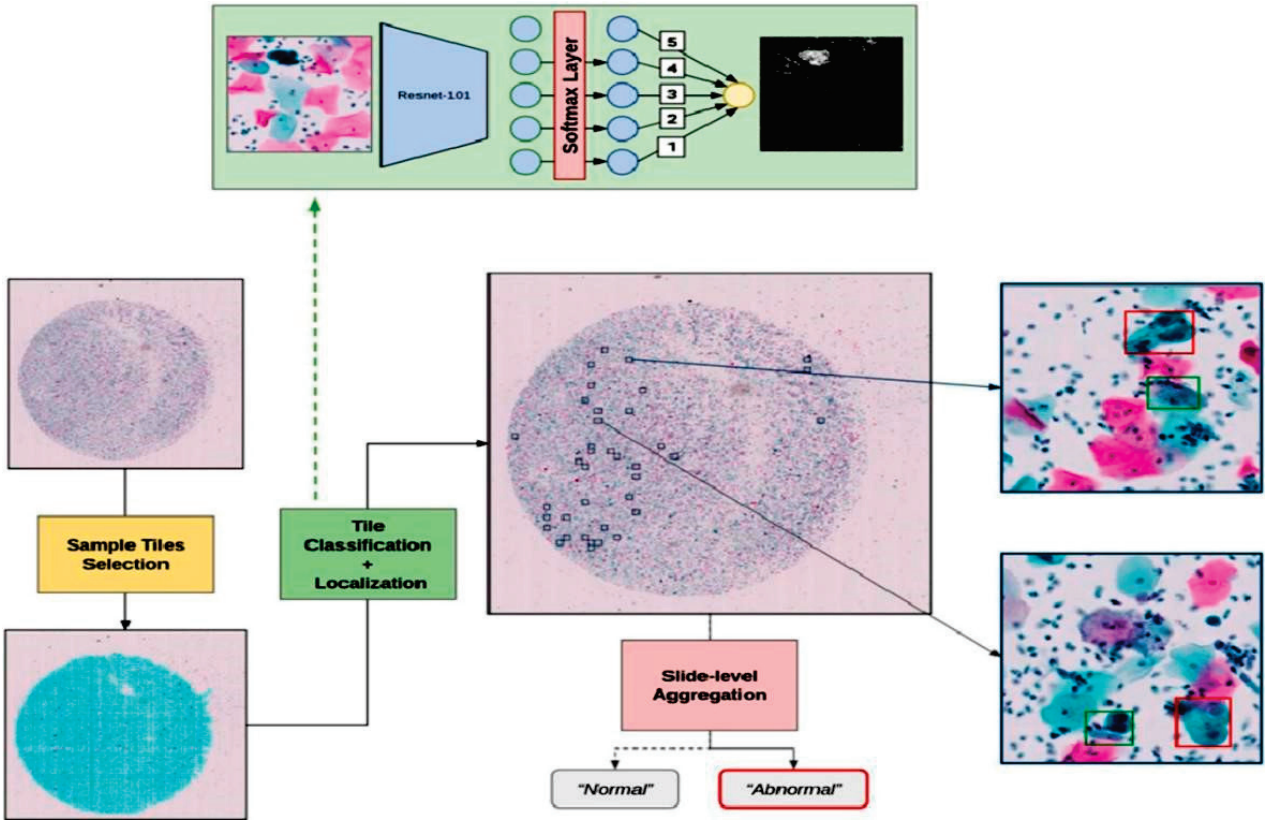


Figure 4. Graphical abstract of Antoine Pirovano et al.'s approach [95].

Some researchers used nanotechnology techniques, where Sakshi Pareek et al. [96] utilized nanotechnology to design an electrochemical biosensor that is sensitive and accurate for human papillomavirus infection (HPV-16) that causes cervical cancer. The designed biosensor is label-free for DNA. The proposed biosensor exhibits excellent sensitivity and stability. This is the core point in the HPV-16 analysis in medical diagnosis fields. On the other hand, Huiting Zhang et al. [97] employed Raman spectroscopy of pre-cancerous lesions for early cervical cancer detection. Their method depends on the Raman spectrum signal of the pre-cancerous cell, then utilizes partial least squares (PLS) with the Relife method for feature extraction from the signal. The selected features are passed to KNN and ELM classifiers. The novelty in their work is the feature fusion in the feature extraction phase. The classifier's performance was enhanced using feature fusion, where

KNN accuracy elevated from 88.17% to 93.55% using feature fusion and ELM from 90.81% to 93.51%.

AI is the challenge of many researchers, such as Sukumar Ponnusamy et al. [98], who combine the artificial neural network and fuzzy system interference (ANFIS) with a watershed algorithm to process, segment, and classify the Pap smear images. They exploited the fuzzy rules to classify abnormal images into their types. Their findings contrast with the existing approach, and it is feasible with high accuracy for classifying malignant cells into their corresponding classes. On top of that, Hongzhen Zhou et al. [99] analyzed the cervical tumor by automatic feature extraction using a deep belief network in contrast-enhanced ultrasonography images. Their goal postulated the effectiveness of intelligent cervical cancer diagnosis on chemotherapy. Their results are presented in terms of higher sensitivity and accuracy for the diagnosis system. Other researchers focused on the segmentation of affected parts of cervical cells using online machine learning (OLM), which was carried out by Asma Daly et al. [100], who segmented the cervical cells using the pelvic region in magnetic resonance imaging (MRI). They obtained high accuracy when they compared their results with existing segmentation techniques. Another type of ML is majority voting, which is based on utilizing a single classifier prediction and then an ensemble of them to vote the major, as proposed by Qazi Mudassar Ilyas et al. [101], who suggested using the ensemble classifier with majority voting of the output. Their ensemble consists of SVM, DT, RF, Naïve Bayes (NB), KNN, LR, J48 DT, and MLP. The best accuracy reached 94% when applied to different benchmark datasets. On the other hand, it utilized other types of classifiers, such as AB, XGBoost, and RF, with the Firefly algorithm as a feature reduction method in addition to SMOTE, which is utilized to deal with imbalance problems in the data. The four diagnostic data sets are exploited (Schiller, Hinselmann, Biopsy, and Cytology). The accuracy is enhanced in terms of reducing the number of selected features [102]. Due to state-of-the-art DL approaches, Khaled Mabrouk Amer Adweb et al. [103] discriminate between normal and pre-cancerous cervical cells using Leaky-RELU and PRELU in residual neural networks. The optimum accuracy reached 90.2% in Leaky-RELU and PRELU and 100% in colposcopy cervical images. On the other hand, Anant R. Bhatt et al. [104] discussed the shortcomings of all existing binary classification methods and conventional neural networks with respect to cervical cancer images. Therefore, they suggested a new approach to extracting features and classifying cervical cancer into multiclasses in a whole slide image (WSI) using ConvNet and a transfer learning strategy. They achieved 99.7% accuracy for multiclass classification in the SIPaKMed dataset. Other research focused on cervical cancer detection employing image processing methods such as Balaji, G. N., et al. [105], which utilized Boykov–Kolmogorov Graph Cuts as well as Cloud Model-based Synergy Integrated Segmentation algorithms for identifying the boundary for cytoplasm and nuclei in cervical Pap smear images. They approved that their methods enhanced the prognosis of cervical cancer by 14% over the traditional segmentation methods. Other studies employed template matching between the measured electrical impedance spectra of cervical cells and the spectra generated from a 3D model of finite elements for cancerous and non-cancerous cervical cells. The matching between spectra is expressed as a score to determine the high strength between the finite element model and the concourse and non-cancerous cells. This method can be effective for cervical cancer detection [106]. Some studies focused on the concomitant presence of miRNA-9-5p in cervical cancer, which was detected by RT-PCR. The experiment concluded that MiRNA-9-5p could be used as a biological marker for cervical cancer, which can be profitable in the inhibition track by inhibiting the CXCR4 gene and protein [107].

Some studies used the Lambert-Beer law to calculate the absorption peak. They found that the absorption is proportional to the cell concentration [108]. In contrast, other studies worked on both breast and cervical cancer together by employing DL [109]. Their work focused on utilizing the concepts of type of cancer, breast or cervical, whether it is located internally or externally, in addition to the imaging modality, whether it is mammography, ultrasound cytology, or colposcopy. Their results compared clinical diagnoses with DL.

They conclude that DL can be an efficient tool for diagnosing cervical or breast cancer that can be replaced by clinician diagnosis. One Nobel and the most effective study depend on the fluorescence signal of urine samples [110]. They collected data using urine samples from 1500 patients and compared them with the healthy subjects, which formed control samples. They achieved a high true positive rate, reaching 74%. Their experiments can be conducted with simple requirements, such as fluorescence device analyses. With an amount of 200 µL, this process for diagnosis needs almost 40 min. On the other hand, the detection of affected papillomaviruses using photothermal-triggered multi-signal readout point-of-care testing (POCT). This bioassay method is realized and sensitive in linear ranges 10^{-6} ng/mL to 1 ng/mL with detection constraints reaching 1.60×10^{-6} ng/mL. This method is effective because it is fast, precise, and optimized for POCT. Therefore, it can be used in rural areas for the early detection of malignancy. Table 3 shows the reality of this method when it is compared with the available cervical cancer biomarker detection methods [111].

Table 3. Comparison with multiple techniques with regard to cervical cancer biomarker detection.

Detection Methods	Targets	Liner Range	LOD (Limit of Detection)
Magnetic sensor	VCP	25–200 ng/mL	2.5×10^{-5} ng/mL
Colorimetric assay	HPV	20–2500 nM	1.03 nM
Electrochemical	pGEM-T/E6	40–5000 ng/mL	0.016 ng/mL
Electrochemical	GST-p16	15.6–250 ng/mL	1.3 ng/mL
Swab immunoassay	E6 protein	10^{-6}–1 ng/mL	1.60×10^{-6} ng/mL

Combining texture features of the nucleus and cytoplasm in Pap smear images is a prominent tool to diagnose cervical cells. This method comes from the reality that doctors diagnose cervical cancer based mainly on the structure as well as the size of the cervical cells. Therefore, the Pap smear images in the Herlev dataset are segmented, and then the texture features are extracted to pass through a multilayer feed-forward neural network. The optimum results show high performance compared with the existing method [112]. On the other side, some studies employed DL and endomicroscopic images to diagnose CIN grade 2. The segmented nucleus is exploited to obtain relevant information for diagnosis. The dataset consisted of 1600 patients, and 20% were used for validation and testing. This approach results in sensitivity reaching 94% and specificity reaching 58%. Therefore, HPV infection test results are considered added features. The sensitivity remains at 94%, and the specificity is enhanced to 71% [113]. Apart from that, Dongyao Jia et al. [114] employed the YOLO (You Only Look Once) algorithm to detect abnormal cervical cells to guarantee the accuracy and rapidity of the model. This novel method forms a milestone for future work in automatic cervical cancer diagnosis.

Among the most prominent studies employed dual-tree complex wavelet transform (DTCWT) with a DL approach to classify Pap smear images into four categories: carcinoma in situ, normal, dysplastic, and superficial. The database is augmented for DL requirements using shearing and flipping transformations. The pixel conductivity of the augmented images is manipulated using multimodal (DTCWT). The CNN that has been used in their experiment is ResNet18, and they obtained a high accuracy of about 99% [115]. On the contrary, Chenjie Li et al. [116] assessed the effectiveness of 3D ultrasound imaging (TUI) on the local staging diagnosis of cervical cancer. Their suggestion is compared with existing methods such as pelvic examination and MRI. Their experiment was conducted on 35 cervical cancer patients, and the back-propagation algorithm was exploited to segment the images. Their results conclude that there is a high correlation between tumor size in MRI and THI, reaching 0.842, and that the correlation between MR and clinical examination

reaches 0.654. This reveals high consistency between MR and THI and can be used for evaluating the local staging for cervical cancer.

For the combination of image processing and AI, most recent studies, such as AbuKhalil, T., et al. [117], enhanced Pap smear images using median filters and then segmented them using Outs thresholding techniques. The deep descriptors are extracted using ResNet and Inception modules. The resultant descriptors are passed to the recurrent neural network (RNN) to classify Pap smear images as cancerous or non-cancerous. In another study, Mohamed Ibrahim Waly et al. [118] used the Harvel data set to classify Pap smear images after applying preprocessing techniques such as a Gaussian filter to remove noise. Then identify the illness portion by segmenting the cell with the Tsallis entropy method with dragonfly optimization (TE-DFO). The segmented region is passed through the SqueezeNet model to extract automated graphical features. Weighted Extreme Learning Machine (ELM) is employed for cervix cell classification. On top of that, R. Elakkiya et al. [119] discussed the shortcomings of the existing methods for classifying cervical cell cancers. Mainly, they are based on accurate spotting and segmentation, in addition to handcrafted feature extraction. Therefore, they proposed Small-Object Detection-Generative Adversarial Networks (SOD-GAN) with a Fine-tuned Stacked Autoencoder (F-SAE) to detect the lesion faster and classify it into premalignant and malignant without segmentation and preprocessing. At the same time, M. Anousouya Devi et al. [120] utilized Neutrosophic Graph Cut-based for segmenting preprocessed Pap smear images into non-overlapping regions, which will lead to enhanced classification accuracy. This algorithm depends mainly on transforming preprocessed Pap smear images into the neutrophilic set. Then, the indeterminacy filter played a main role in integrating the intensity, including the spatial information of preprocessed images based on the indeterminacy value. This value specifies the weights for each pixel to define the graph. Finally, the maximum graph is determined to obtain the optimal segmentation results. This approach is better than existing detection methods by over 13%.

4. Discussion

Cervical cancer is a prominent health problem globally, with high mortality as well as incidence rates, particularly in developing countries [121,122]. Early detection is critical for the successful treatment and management of cervical cancer. The traditional method for cervical cancer screening is the Pap smear test, which involves the examination of cervical cells under a microscope for abnormalities. HPV is a very common sexually transmitted infection, with estimates estimating that up to 80% of sexually active women will become infected with HPV at some point in their lives. However, the majority of these infections will clear up on their own without causing any long-term health problems. There are many different types of HPV, and some types are more likely to cause cancer than others. However, this method is subjective and may miss precancerous lesions, leading to false negatives and a delayed diagnosis. Therefore, there has been further interest in establishing CAD methods to improve cervical cancer screening. CAD technology for cervical cancer detection has been extensively examined over the past few decades [123,124]. Between 1996 and 2022, significant advancements have been made in this field, leading to improved accuracy, sensitivity, and specificity of CAD methods. Early CAD systems utilized image processing and pattern recognition techniques to analyze digital images of cervical cells with the aim of identifying abnormal cells and lesions. However, these early systems had limited success due to low sensitivity and specificity.

In the early 2000s, ML algorithms were introduced to the field of CAD for cervical cancer detection. ML algorithms can analyze large datasets and learn from them to identify patterns and make predictions. This allowed for more accurate and automated analysis of digital images of cervical cells. ML-based CAD systems have shown promise in several studies, with improved sensitivity and specificity reported compared to traditional screening methods [125–127]. Among the most promising CAD systems for cervical cancer detection is the Hybrid Intelligent System for Cervical Cancer Diagnosis (HISCCD), which was developed in 2012. HISCCD is a combination of ML algorithms and rule-based

systems that analyze digital images of cervical cells to detect abnormal cells and lesions. Several studies have reported improved sensitivity and specificity of HISCCD compared to traditional screening methods. Another promising CAD system is the Automated Cervical Screening System (ACSS), which was introduced in 2016. ACSS uses an ML-based algorithm to analyze digital images of cervical cells and identify abnormal cells and lesions. In a study comparing ACSS to the Pap smear test, ACSS showed higher specificity and sensitivity for detecting high-grade cervical intraepithelial neoplasia. In addition to these systems, there have been several other CAD systems developed over the years, each with its own strengths and limitations. One of the major challenges with CAD systems for cervical cancer detection is the lack of standardized protocols and data sharing, which limits their widespread adoption and validation.

The previous studies describe the most updated state-of-the-art techniques that were suggested, validated, and evaluated for early cervical cancer detection. Most researchers conducted their experiments utilizing image processing in addition to ML and DL. The pre-processing techniques are employed to enhance the visualization of Pap smear images and make feature extraction an easy and more accurate task. Other researchers skipped this step by utilizing DL techniques to extract features automatically, which reduces time and gives accurate results because all of the features excreted in this step are relevant to the corresponding class. However, many researchers focused on HPV, which plays the main role in the infection of cervical cancer. They focused on the nanotechnology track by designing a biosensor that can detect the infection and is distinguished by its stability and linearity. Other researchers focus on building a finite element model for both cancerous and noncancerous cells to study the electrical impedance spectroscopy and compare it with the tested cell to find the matching score between them. They count it as an alternative method that is more accurate than using a Pap smear screening test. Chemical reactions are also considered by other researchers by studying the fluorescence signals from the urine of the infected women and comparing those signals with those of healthy women.

Various methods have been carried out in this area, either in biochemistry, image processing, DL, signals, or nanotechnology tracks, to enhance and reach a highly accurate approach to diagnosing cervical cancer in its early stages. This will reduce the mortality rate among women and increase the chance of survival. In conclusion, CAD technology for cervical cancer detection has come a long way since its introduction in the 1990s. ML-based algorithms have shown promise in improving the accuracy and sensitivity of CAD systems for cervical cancer detection. HISCCD and ACSS are two of the most promising CAD systems, but extensive research and validation are required before they can be broadly applied.

5. Conclusions

Cervical cancer is a substantial public health issue globally, with more than half a million new cases and a quarter of a million deaths each year. Early detection and treatment of cervical cancer can significantly improve outcomes and save lives. Fortunately, there are several different methods for cervical cancer detection, each with its own limitations and advantages. The Pap smear test is the most broadly employed and popular technique with respect to cervical cancer detection. It is a low-cost, simple, and efficient way to screen for precancerous or cancerous changes in the cervix. The Pap smear test has undergone several improvements over the years, including the use of liquid-based cytology, which has improved its accuracy and sensitivity. However, the Pap smear test is not foolproof and can miss some cases of cervical cancer, especially in its early stages.

The recommended screening guidelines may vary depending on age, risk factors, and previous screening results. In developed countries, the adoption of cervical cancer screening programs has led to a significant decrease in cervical cancer mortality rates. However, in low- and middle-income countries, the lack of access to screening programs and cost-effective screening methods and vaccines is a significant barrier to early detection and effective treatment. Therefore, the development of simple, low-cost, and accurate

screening methods that can be implemented in low-resource settings is essential. In recent years, machine learning (ML) and deep learning (DL) algorithms have been deployed to aid in cervical cancer diagnosis and treatment by identifying abnormal and normal cells automatically, precisely, and quickly. These algorithms have demonstrated high sensitivity and specificity in detecting abnormal cervical cells, indicating their potential use as an adjunct to traditional screening methods. However, more research is needed to evaluate the feasibility and effectiveness of these algorithms in real-world clinical settings.

In the future, the identification of important risk factors as well as the utilization of various segmentation pre-processing techniques can enhance the effectiveness of cervical cancer diagnosis and treatment. Bigger and more balanced data can also improve the performance of future classification systems. In conclusion, cervical cancer detection has come a long way over the years, with several different methods available, each with its advantages and limitations. The Pap smear test remains the most frequently employed method, but newer methods, including HPV testing, VIA, and VILI, are becoming more widely used. A colposcopy is also an important tool for follow-up and diagnostic purposes. Regular cervical cancer screening is critical for early detection and successful treatment. Women should discuss their screening options with their healthcare provider and follow the recommended guidelines for cervical cancer screening. By working together, we can continue to improve cervical cancer detection and save lives. Nevertheless, continued innovation and collaboration in this field may facilitate the enhancement of cervical cancer detection and ultimately lower the disease's burden on women worldwide.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thank you to the Fundamental Research Grant Scheme (FRGS/1/2021/SKK0/UNIMAP/02/1) of the Ministry of Higher Education of Malaysia for supporting this project.

Conflicts of Interest: The writers certify that they have no conflicting interests in relation to this research.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
2. Azizah, A.M.; Hashimah, B.; Nirmal, K.; Siti Zubaidah, A.R.; Puteri, N.A.; Nabihah, A.; Sukumaran, R.; Balqis, B.; Nadia, S.M.R.; Sharifah, S.S.S.; et al. *Malaysia National Cancer Registry Report (MNCR) 2012–2016*; National Cancer Registry Department: Putrajaya, Malaysia, 2019.
3. Ezat, S.W.P.; Aljunid, S. Comparative cost-effectiveness of HPV vaccines in the prevention of cervical cancer in Malaysia. *Asian Pacific J. Cancer Prev.* **2010**, *11*, 943–951. [CrossRef]
4. Mustafa, W.A.; Halim, A.; Ab Rahman, K.S. A Narrative Review: Classification of Pap Smear Cell Image for Cervical Cancer Diagnosis. *Oncologie* **2020**, *22*, 53–63. [CrossRef]
5. Mustafa, W.A.; Halim, A.; Jamlos, M.A.; Idrus, S.Z.S. A Review: Pap Smear Analysis Based on Image Processing Approach. *J. Physics Conf. Ser.* **2020**, *1529*, 022080. [CrossRef]
6. Mustafa, W.A.; Wei, L.Z.; Ab Rahman, K.S. Automated Cell Nuclei Segmentation on Cervical Smear Images Using Structure Analysis. *J. Biomim. Biomater. Biomed. Eng.* **2021**, *51*, 105–115. [CrossRef]
7. Halim, A.; Mustafa, W.A.; Ahmad, W.K.W.; Rahim, H.A.; Sakeran, H. Nucleus Detection on Pap Smear Images for Cervical Cancer Diagnosis: A Review Analysis. *Oncologie* **2021**, *23*, 73–88. [CrossRef]
8. Rudmann, D.G.; Foley, G.L. Female Reproductive System. In *Fundamentals of Toxicologic Pathology*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 2018; pp. 517–545. [CrossRef]

9. Shetty, A.; Shah, V. Survey of Cervical Cancer Prediction Using Machine Learning: A Comparative Approach. In Proceedings of the 2018 9th International Conference on Computing, Communication and Networking Technologies ICCCNT 2018, Bengaluru, India, 10–12 July 2018; pp. 1–6. [CrossRef]
10. Ortelli, T.A. The National Library of Medicine. *AJN Am. J. Nurs.* **2019**, *119*, 53–54. [CrossRef]
11. Muñoz, N.; Bosch, F.X.; Castellsagué, X.; Díaz, M.; de Sanjose, S.; Hammouda, D.; Shah, K.V.; Meijer, C.J. Against which human papillomavirus types shall we vaccinate and screen? the international perspective. *Int. J. Cancer* **2004**, *111*, 278–285. [CrossRef]
12. Ali, F.; Kuelker, R.; Wassie, B. Understanding cervical cancer in the context of developing countries. *Ann. Trop. Med. Public Health* **2012**, *5*, 3–15. [CrossRef]
13. Rajarao, C.; Singh, R.P. Improved normalized graph cut with generalized data for enhanced segmentation in cervical cancer detection. *Evol. Intell.* **2019**, *13*, 3–8. [CrossRef]
14. Logeswaran, K.; Suresh, P.; Savitha, S.; Prasanna Kumar, K.R.; Ponselvakumar, A.P.; Kannan, A.R. Data driven diagnosis of cervical cancer using association rule mining with trivial rule expulsion approach. *Int. J. Emerg. Technol.* **2020**, *11*, 110–115.
15. Lu, J.; Song, E.; Ghoneim, A.; Alrashoud, M. Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Futur. Gener. Comput. Syst.* **2020**, *106*, 199–205. [CrossRef]
16. Singh, S.K.; Goyal, A. Performance Analysis of Machine Learning Algorithms for Cervical Cancer Detection. *Int. J. Health Inf. Syst. Inform.* **2020**, *15*, 1–21. [CrossRef]
17. Rehman, A.-U.; Ali, N.; Taj, I.; Sajid, M.; Karimov, K.S. An Automatic Mass Screening System for Cervical Cancer Detection Based on Convolutional Neural Network. *Math. Probl. Eng.* **2020**, *2020*, 4864835. [CrossRef]
18. Bhuvaneshwari, K.V.; Poornima, B. Cervical cancer cell identification & detection using fuzzy C mean and K nearest neighbor techniques. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 1080–1084. [CrossRef]
19. Sirovich, B.E.; Welch, H.G. The frequency of Pap smear screening in the United States. *J. Gen. Intern. Med.* **2004**, *19*, 243–250. [CrossRef]
20. Chan, S.W.; Leung, K.; Wong, W.F. An expert system for the detection of cervical cancer cells using knowledge-based image analyzer. *Artif. Intell. Med.* **1996**, *8*, 67–90. [CrossRef]
21. Okimoto, G.S.; Parker, M.F.; Mooradian, G.C.; Saggese, S.J.; Grisanti, A.A.; O'Connor, D.M.; Miyazawa, K. New features for detecting cervical precancer using hyperspectral diagnostic imaging. *Clin. Diagn. Syst.* **2001**, *4255*, 67–80. [CrossRef]
22. Lonappan, A.; Thimothy, V.O.; Rajasekaran, C.; Thomas, V.; Jacob, J.; Mathew, K.T. A novel method of detecting cervical cancer using microwaves. *Microw. Opt. Technol. Lett.* **2008**, *50*, 1552–1554. [CrossRef]
23. Weber, C.R.; Schwarz, R.A.; Atkinson, E.N.; Cox, D.D.; MacAulay, C.; Follen, M.; Richards-Kortum, R.R. Model-based analysis of reflectance and fluorescence spectra for in vivo detection of cervical dysplasia and cancer. *J. Biomed. Opt.* **2008**, *13*, 064016. [CrossRef]
24. Wu, T.; Cheung, T.-H.; Yim, S.-F.; Qu, J.Y. Clinical study of quantitative diagnosis of early cervical cancer based on the classification of acetowhitening kinetics. *J. Biomed. Opt.* **2010**, *15*, 026001. [CrossRef]
25. Hai, T.H.; Phuc, L.H.; Vinh, L.K.; Long, B.D.; Kieu, T.T.; Bich, N.N.; Lan, T.N.; Hien, N.Q.; Khoa, L.H.A.; Van Tam, N.N. Immobilising of anti-HPV18 and *E. coli* O_{157}:H_7 antibodies on magnetic silica-coated Fe_3O_4 for early diagnosis of cervical cancer and diarrhoea. *Int. J. Nanotechnol.* **2011**, *8*, 383. [CrossRef]
26. DPruski, D.; Przybylski, M.; Kędzia, W.; Kędzia, H.; Jagielska-Pruska, J.; Spaczyński, M. Optoelectronic method for detection of cervical intraepithelial neoplasia and cervical cancer. *Opto-Electron. Rev.* **2011**, *19*, 478–485. [CrossRef]
27. Xie, Y.; Xu, G.; Shen, P.; Pan, X.; Yan, R.; Chen, L.; Zhang, Y.; Zhu, F. Study on The Value Assessment of hWAPL Histological Expression in Histological and Cytological Diagnosis of Cervical Cancer and Cervical Intraepithelial Neoplasia. *J. Converg. Inf. Technol.* **2011**, *6*, 330–337. [CrossRef]
28. Gertych, A.; Joseph, A.O.; Walts, A.E.; Bose, S. Automated Detection of Dual p16/Ki67 Nuclear Immunoreactivity in Liquid-Based Pap Tests for Improved Cervical Cancer Risk Stratification. *Ann. Biomed. Eng.* **2012**, *40*, 1192–1204. [CrossRef] [PubMed]
29. Lu, C.; Chelikani, S.; Jaffray, D.A.; Milosevic, M.F.; Staib, L.H.; Duncan, J.S. Simultaneous nonrigid registration, segmentation, and tumor detection in MRI guided cervical cancer radiation therapy. *IEEE Trans. Med. Imaging* **2012**, *31*, 1213–1227. [CrossRef]
30. Masilamani, V.; AlSalhi, M.S.; Vijmasi, T.; Govindarajan, K.; Rai, R.R.; Atif, M.; Prasad, S.; Aldwayyan, A.S. Fluorescence spectra of blood and urine for cervical cancer detection. *J. Biomed. Opt.* **2012**, *17*, 0980011. [CrossRef] [PubMed]
31. Kalaivani, R.; Masilamani, V.; AlSalhi, M.S.; Devanesan, S.; Ramamurthy, P.; Palled, S.R.; Ganesh, K.M. Cervical cancer detection by time-resolved spectra of blood components. *J. Biomed. Opt.* **2014**, *19*, 057011. [CrossRef] [PubMed]
32. Wu, D.; Guo, A.; Guo, Z.; Xie, L.; Wei, Q.; Du, B. Simultaneous electrochemical detection of cervical cancer markers using reduced graphene oxide-tetraethylene pentamine as electrode materials and distinguishable redox probes as labels. *Biosens. Bioelectron.* **2014**, *54*, 634–639. [CrossRef]
33. Devi, S.; Panigrahi, P.K.; Pradhan, A. Detecting cervical cancer progression through extracted intrinsic fluorescence and principal component analysis. *J. Biomed. Opt.* **2014**, *19*, 127003. [CrossRef]
34. Ramapraba, P.S.; Ranganathan, H. Feature fusion for cervical cancer detection using colposcopic images. *Int. J. Appl. Eng. Res.* **2015**, *10*, 4803–4810.
35. Sukumar, P.; Gnanamurthy, R.K. Computer Aided Detection of Cervical Cancer Using Pap Smear Images Based on Hybrid Classifier. *Int. J. Appl. Eng. Res.* **2015**, *10*, 21021–21032.

36. Das, A.; Kar, A.; Bhattacharyya, D. A novel illumination correction and intensity normalization method on cervigrams in the early detection of uterine cervical cancer. *ARPN J. Eng. Appl. Sci.* **2015**, *10*, 6376–6380.
37. Yamal, J.; Guillaud, M.; Atkinson, E.N.; Follen, M.; MacAulay, C.; Cantor, S.B.; Cox, D.D. Prediction using hierarchical data: Applications for automated detection of cervical cancer. *Stat. Anal. Data Mining ASA Data Sci. J.* **2015**, *8*, 65–74. [CrossRef]
38. Amini, N.; Supriyanto, E.; Marvibaigi, M.; Majid, F.A.A. Human papilloma virus E6/E7 messenger RNA as a biomarker for detecting the risk evaluation of cervical cancer progression: Overview of recent clinical. *J. Teknol.* **2015**, *75*, 215–224. [CrossRef]
39. Narayanan, N.; Karunakaran, V.; Paul, W.; Venugopal, K.; Sujathan, K.; Maiti, K.K. Aggregation induced Raman scattering of squaraine dye: Implementation in diagnosis of cervical cancer dysplasia by SERS imaging. *Biosens. Bioelectron.* **2015**, *70*, 145–152. [CrossRef] [PubMed]
40. Sajeena, T.A.; Jereesh, A.S. Cervical cancer detection through automatic segmentation and classification of Pap smear cells. *Int. J. Appl. Eng. Res.* **2015**, *10*, 39078–39084.
41. Li, H.; Wang, X.; Geng, J.; Zhao, X. Clinical Study of Styping Detection of Human Papillomavirus (HPV) Infection with Microarray from Paraffinembedded Specimens of Cervical Cancer and Precursor Lesions. *J. Nanosci. Nanotechnol.* **2015**, *15*, 6423–6428. [CrossRef] [PubMed]
42. Guz, N.V.; Dokukin, M.E.; Woodworth, C.D.; Cardin, A.; Sokolov, I. Towards early detection of cervical cancer: Fractal dimension of AFM images of human cervical epithelial cells at different stages of progression to cancer. *Nanomed. Nanotechnol. Biol. Med.* **2015**, *11*, 1667–1675. [CrossRef] [PubMed]
43. Vidya, R. The early detection and handling of cervical cancer with nano technology and biomarkers—A comprehensive study. *Int. J. Appl. Eng. Res.* **2015**, *10*, 387–390.
44. Duangkaew, P.; Tapaneeyakorn, S.; Apiwat, C.; Dharakul, T.; Laiwejpithaya, S.; Kanatharana, P.; Laocharoensuk, R. Ultrasensitive electrochemical immunosensor based on dual signal amplification process for p16INK4a cervical cancer detection in clinical samples. *Biosens. Bioelectron.* **2015**, *74*, 673–679. [CrossRef] [PubMed]
45. Wang, Y.; Song, C.; Wang, M.; Xie, Y.; Mi, L.; Wang, G. Rapid, Label-Free, and Highly Sensitive Detection of Cervical Cancer with Fluorescence Lifetime Imaging Microscopy. *IEEE J. Sel. Top. Quantum Electron.* **2015**, *22*, 228–234. [CrossRef]
46. Athinarayanan, S.; Srinath, M.V. Robust and efficient diagnosis of cervical cancer in pap smear images using textures features with rbf and kernel SVM classification. *ARPN J. Eng. Appl. Sci.* **2016**, *11*, 4504–4515.
47. Devi, M.A.; Ravi, S.; Vaishnavi, J.; Punitha, S. Detection of cervical cancer using the image analysis algorithms. *Int. J. Control Theory Appl.* **2016**, *9*, 3193–3203.
48. Xu, X.; Li, Z.; Qiu, X.; Wei, Z. Diagnosis performance of positron emission tomography-computed tomography among cervical cancer patients. *J. X-ray Sci. Technol.* **2016**, *24*, 531–536. [CrossRef]
49. Palacio, J.A.A.; Van Noije, W.A. High stability voltage controlled current source for cervical cancer detection using electrical impedance spectroscopy. *Analog. Integr. Circuits Signal Process.* **2016**, *89*, 541–547. [CrossRef]
50. Sobar; Machmud, R.; Wijaya, A. Behavior Determinant Based Cervical Cancer Early Detection with Machine Learning Algorithm. *Adv. Sci. Lett.* **2016**, *22*, 3120–3123. [CrossRef]
51. Sitompul, I.R.H.; A Yoga, P.; Rosa, R.T.; Shahab, S.N.; Nasution, V.A.F.; Reza, M.; Nuryanto, K.H. Overview Knowledge of Reproductive Aged Women of the Prevention and Early Detection of Cervical Cancer at Cakung Sub-District Community Health Centre in 2015. *Adv. Sci. Lett.* **2017**, *23*, 6984–6986. [CrossRef]
52. Jeftic, B.; Papic-Obradovic, M.; Muncan, J.; Matija, L.; Koruga, D. Optomagnetic Imaging Spectroscopy Application in Cervical Dysplasia and Cancer Detection: Comparation of Stained and Unstained Papanicolaou Smears. *J. Med. Biol. Eng.* **2017**, *37*, 936–943. [CrossRef]
53. Iliyasu, A.M.; Fatichah, C. A Quantum Hybrid PSO Combined with Fuzzy *k*-NN Approach to Feature Selection and Cell Classification in Cervical Cancer Detection. *Sensors* **2017**, *17*, 2935. [CrossRef]
54. Wu, W.; Zhou, H. Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches. *IEEE Access* **2017**, *5*, 25189–25195. [CrossRef]
55. Carow, K.; Gölitz, M.; Wolf, M.; Häfner, N.; Jansen, L.; Hoyer, H.; Schwarz, E.; Runnebaum, I.B.; Dürst, M. Viral-Cellular DNA Junctions as Molecular Markers for Assessing Intra-Tumor Heterogeneity in Cervical Cancer and for the Detection of Circulating Tumor DNA. *Int. J. Mol. Sci.* **2017**, *18*, 2032. [CrossRef] [PubMed]
56. Kudva, V.; Prasad, K.; Guruvare, S. Detection of Specular Reflection and Segmentation of Cervix Region in Uterine Cervix Images for Cervical Cancer Screening. *IRBM* **2017**, *38*, 281–291. [CrossRef]
57. Sun, G. Cervical Cancer Diagnosis based on Random Forest. *Int. J. Perform. Eng.* **2017**, *17*, 446. [CrossRef]
58. Shaikh, R.; Prabitha, V.G.; Dora, T.K.; Chopra, S.; Maheshwari, A.; Deodhar, K.; Rekhi, B.; Sukumar, N.; Krishna, C.M.; Subhash, N. A comparative evaluation of diffuse reflectance and Raman spectroscopy in the detection of cervical cancer. *J. Biophotonics* **2016**, *10*, 242–252. [CrossRef]
59. Muljo, H.H.; Pardamean, B.; Perbangsa, A.S. The Implementation of Online Learning for Early Detection of Cervical Cancer. *J. Comput. Sci.* **2017**, *13*, 600–607. [CrossRef]
60. Arya, M.; Mittal, N.; Singh, G. Texture-based feature extraction of smear images for the detection of cervical cancer. *IET Comput. Vis.* **2018**, *12*, 1049–1059. [CrossRef]

61. Sharma, A.; Jha, A.K.; Mishra, S.; Jain, A.; Chauhan, B.S.; Kathuria, M.; Rawat, K.S.; Gupta, N.M.; Tripathi, R.; Mitra, K.; et al. Imaging and Quantitative Detection of Lipid Droplets by Yellow Fluorescent Probes in Liver Sections of *Plasmodium* Infected Mice and Third Stage Human Cervical Cancer Tissues. *Bioconjugate Chem.* **2018**, *29*, 3606–3613. [CrossRef]
62. Fernandes, K.; Chicco, D.; Cardoso, J.S.; Fernandes, J. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Comput. Sci.* **2018**, *4*, e154. [CrossRef]
63. Jing, Y.; Wang, Y.; Wang, X.; Song, C.; Ma, J.; Xie, Y.; Fei, Y.; Zhang, Q.; Mi, L. Label-free imaging and spectroscopy for early detection of cervical cancer. *J. Biophotonics* **2018**, *11*, e201700245. [CrossRef]
64. Dillak, R.Y.; Manulangga, G.C.; Lalandos, J.L. Early warning system for cervical cancer diagnosis using ridge polynomial neural network and chaos optimization algorithm. *J. Theor. Appl. Inf. Technol.* **2018**, *96*, 1989–1998.
65. Kudva, V.; Prasad, K.; Guruvare, S. Automation of Detection of Cervical Cancer Using Convolutional Neural Networks. *Crit. Rev. Biomed. Eng.* **2018**, *46*, 135–145. [CrossRef] [PubMed]
66. Abdoh, S.F.; Rizka, M.A.; Maghraby, F.A. Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques. *IEEE Access* **2018**, *6*, 59475–59485. [CrossRef]
67. Devi, L.; Thirumurugan, D. Automated Detection of Cervical Cancer. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 1399–1401. [CrossRef]
68. Abdullah, A.A.; Giong, A.F.D.; Zahri, N.A.H. Cervical cancer detection method using an improved cellular neural network (CNN) algorithm. *Indones. J. Electr. Eng. Comput. Sci.* **2019**, *14*, 210–218. [CrossRef]
69. Jaya, S.; Latha, M. Diagnosis of cervical cancer using CLAHE and SGLDM on RGB pap smear images through ANN. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *9*, 530–534. [CrossRef]
70. Wang, H.; Jiang, C.; Bao, K.; Xu, C. Recognition and Clinical Diagnosis of Cervical Cancer Cells Based on our Improved Lightweight Deep Network for Pathological Image. *J. Med. Syst.* **2019**, *43*, 1–9. [CrossRef]
71. William, W.; Ware, A.; Basaza-Ejiri, A.H.; Obungoloch, J. A pap-smear analysis tool (PAT) for detection of cervical cancer from pap-smear images. *Biomed. Eng. Online* **2019**, *18*, 1–22. [CrossRef] [PubMed]
72. Jain, R.; Sangwan, S.R.; Bachhety, S.; Garg, S.; Upadhyay, Y. Optimized Model for Cervical Cancer Detection Using Binary Cuckoo Search. *Recent Patents Comput. Sci.* **2019**, *12*, 293–303. [CrossRef]
73. Adem, K.; Kiliçarslan, S.; Cömert, O. Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Syst. Appl.* **2018**, *115*, 557–564. [CrossRef]
74. Wei, S.; Dai, P.; Wang, Z. Cervical Cancer Detection and Diagnosis Based on Saliency Single Shot MultiBox Detector in Ultrasonic Elastography. *J. Med. Syst.* **2019**, *43*, 250. [CrossRef] [PubMed]
75. Ali, A.; Hulipalled, V.R.; Patil, S.S.; Kappaparambil, R.A. DPCCG-EJA: Detection of key pathways and cervical cancer related genes using enhanced Johnson's algorithm. *Int. J. Adv. Sci. Technol.* **2019**, *28*, 124–138.
76. Huang, M.; Zhang, Z.; Wang, X.; Xie, Y.; Fei, Y.; Ma, J.; Wang, J.; Chen, L.; Mi, L.; Wang, Y. Detecting benign uterine tumors by autofluorescence lifetime imaging microscopy through adjacent healthy cervical tissues. *J. Innov. Opt. Health Sci.* **2019**, *12*, 1940006. [CrossRef]
77. Asiedu, M.N.; Simhal, A.; Chaudhary, U.; Mueller, J.L.; Lam, C.T.; Schmitt, J.W.; Venegas, G.; Sapiro, G.; Ramanujam, N. Development of Algorithms for Automated Detection of Cervical Pre-Cancers With a Low-Cost, Point-of-Care, Pocket Colposcope. *IEEE Trans. Biomed. Eng.* **2018**, *66*, 2306–2318. [CrossRef] [PubMed]
78. Lu, D.; Xia, J.; Deng, Z.; Cao, X. Detection of squamous cell carcinoma antigen in cervical cancer by surface-enhanced Raman scattering-based immunoassay. *Anal. Methods* **2019**, *11*, 2809–2818. [CrossRef]
79. Vaijayanthimala, M.; Kumari, S.R. FSSBFS: Fuzzy Swallow Swarm based Feature Selection for Diagnosis of Cervical Cancer. *J. Adv. Res. Dyn. Control Syst.* **2019**, *11*, 77–88. [CrossRef]
80. Li, Y.; Chen, J.; Xue, P.; Tang, C.; Chang, J.; Chu, C.; Ma, K.; Li, Q.; Zheng, Y.; Qiao, Y. Computer-Aided Cervical Cancer Diagnosis Using Time-Lapsed Colposcopic Images. *IEEE Trans. Med. Imaging* **2020**, *39*, 3403–3415. [CrossRef]
81. Ma, D.; Liu, J.; Li, J.; Zhou, Y. Cervical cancer detection in cervical smear images using deep pyramid inference with refinement and spatial-aware booster. *IET Image Process.* **2020**, *14*, 4717–4725. [CrossRef]
82. Jia, A.D.; Li, B.Z.; Zhang, C.C. Detection of cervical cancer cells based on strong feature CNN-SVM network. *Neurocomputing* **2020**, *411*, 112–127. [CrossRef]
83. Li, X.-X.; Lin, T.-T.; Liu, B.; Wei, W. Diagnosis of Cervical Cancer With Parametrial Invasion on Whole-Tumor Dynamic Contrast-Enhanced Magnetic Resonance Imaging Combined With Whole-Lesion Texture Analysis Based on T2- Weighted Images. *Front. Bioeng. Biotechnol.* **2020**, *8*, 590. [CrossRef]
84. Ramasamy, R.; Chinnasamy, C. Detection and segmentation of cancer regions in cervical images using fuzzy logic and adaptive neuro fuzzy inference system classification method. *Int. J. Imaging Syst. Technol.* **2019**, *30*, 412–420. [CrossRef]
85. Ijaz, M.F.; Attique, M.; Son, Y. Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods. *Sensors* **2020**, *20*, 2809. [CrossRef] [PubMed]
86. Zaffar, M.; Pradhan, A. Assessment of anisotropy of collagen structures through spatial frequencies of Mueller matrix images for cervical pre-cancer detection. *Appl. Opt.* **2020**, *59*, 1237–1248. [CrossRef]
87. Mahmoodi, P.; Rezayi, M.; Rasouli, E.; Avan, A.; Gholami, M.; Mobarhan, M.G.; Karimi, E.; Alias, Y. Early-stage cervical cancer diagnosis based on an ultra-sensitive electrochemical DNA nanobiosensor for HPV-18 detection in real samples. *J. Nanobiotechnol.* **2020**, *18*, 11. [CrossRef]

88. Sahoo, G.R.; Dey, R.; Das, N.; Ghosh, N.; Pradhan, A. Two dimensional multifractal detrended fluctuation analysis of low coherence images for diagnosis of cervical pre-cancer. *Biomed. Phys. Eng. Express* **2020**, *6*, 025011. [CrossRef]
89. Chitra, B.; Kumar, S.S. Recent advancement in cervical cancer diagnosis for automated screening: A detailed review. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *13*, 251–269. [CrossRef]
90. Ali, M.; Ahmed, K.; Bui, F.M.; Paul, B.K.; Ibrahim, S.M.; Quinn, J.M.; Moni, M.A. Machine learning-based statistical analysis for early stage detection of cervical cancer. *Comput. Biol. Med.* **2021**, *139*, 104985. [CrossRef] [PubMed]
91. Chitra, B.; Kumar, S.S. An optimized deep learning model using Mutation-based Atom Search Optimization algorithm for cervical cancer detection. *Soft Comput.* **2021**, *25*, 15363–15376. [CrossRef]
92. Zhang, C.; Jia, D.; Wu, N.; Guo, Z.; Ge, H. Quantitative detection of cervical cancer based on time series information from smear images. *Appl. Soft Comput.* **2021**, *112*, 107791. [CrossRef]
93. Jahan, S.; Islam, M.D.S.; Islam, L.; Rashme, T.Y.; Prova, A.A.; Paul, B.K.; Mosharof, M.K. Automated invasive cervical cancer disease detection at early stage through suitable machine learning model. *SN Appl. Sci.* **2021**, *3*, 1–17. [CrossRef]
94. Cao, L.; Yang, J.; Rong, Z.; Li, L.; Xia, B.; You, C.; Lou, G.; Jiang, L.; Du, C.; Meng, H.; et al. A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening. *Med. Image Anal.* **2021**, *73*, 102197. [CrossRef]
95. Pirovano, A.; Almeida, L.G.; Ladjal, S.; Bloch, I.; Berlemont, S. Computer-aided diagnosis tool for cervical cancer screening with weakly supervised localization and detection of abnormalities using adaptable and explainable classifier. *Med. Image Anal.* **2021**, *73*, 102167. [CrossRef] [PubMed]
96. Pareek, S.; Jain, U.; Bharadwaj, M.; Chauhan, N. A label free nanosensing platform for the detection of cervical cancer through analysis of ultratrace DNA hybridization. *Sens. Bio-Sens. Res.* **2021**, *33*, 100444. [CrossRef]
97. Zhang, H.; Chen, C.; Ma, C.; Zhu, Z.; Yang, B.; Chen, F.; Jia, D.; Li, Y.; Lv, X. Feature Fusion Combined With Raman Spectroscopy for Early Diagnosis of Cervical Cancer. *IEEE Photon- J.* **2021**, *13*, 3075958. [CrossRef]
98. Ponnusamy, S.; Samikannu, R.; Venkatachary, S.K.; Sukumar, S.; Ravi, R. RETRACTED ARTICLE: Computer aided innovation method for detection and classification of cervical cancer using ANFIS classifier. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *12*, 6231–6240. [CrossRef]
99. Zhou, H.; Wang, S.; Zhang, T.; Liu, D.; Yang, K. Ultrasound image analysis technology under deep belief networks in evaluation on the effects of diagnosis and chemotherapy of cervical cancer. *J. Supercomput.* **2020**, *77*, 4151–4171. [CrossRef]
100. Daly, A.; Yazid, H.; Solaiman, B.; Ben Amara, N.E. Multiatlas-based segmentation of female pelvic organs: Application for computer-aided diagnosis of cervical cancer. *Int. J. Imaging Syst. Technol.* **2020**, *31*, 302–312. [CrossRef]
101. Ilyas, Q.M.; Ahmad, M. An Enhanced Ensemble Diagnosis of Cervical Cancer: A Pursuit of Machine Intelligence Towards Sustainable Health. *IEEE Access* **2021**, *9*, 12374–12388. [CrossRef]
102. Khan, I.U.; Aslam, N.; Alshehri, R.; Alzahrani, S.; Alghamdi, M.; Almalki, A.; Balabeed, M. Cervical Cancer Diagnosis Model Using Extreme Gradient Boosting and Bioinspired Firefly Optimization. *Sci. Program.* **2021**, *2021*, 5540024. [CrossRef]
103. Adweb, K.M.A.; Cavus, N.; Sekeroglu, B. Cervical Cancer Diagnosis Using Very Deep Networks Over Different Activation Functions. *IEEE Access* **2021**, *9*, 46612–46625. [CrossRef]
104. Bhatt, A.R.; Ganatra, A.; Kotecha, K. Cervical cancer detection in pap smear whole slide images using convNet with transfer learning and progressive resizing. *PeerJ Comput. Sci.* **2021**, *7*, e348. [CrossRef]
105. Balaji, G.; Suryanarayana, S.; Sengathir, J. Enhanced boykov's graph cuts based segmentation for cervical cancer detection. *EAI Endorsed Trans. Pervasive Health Technol.* **2021**, *21*, e3. [CrossRef]
106. Li, P.; Highfield, P.E.; Lang, Z.-Q.; Kell, D. Cervical cancer prognosis and diagnosis using electrical impedance spectroscopy. *J. Electr. Bioimpedance* **2021**, *12*, 153–162. [CrossRef]
107. Sun, W.-L.; Shen, Y.; Yuan, Y.; Zhou, X.-J.; Li, W.-P. The Value and Clinical Significance of Tumor Marker Detection in Cervical Cancer. *Sci. Program.* **2021**, *2021*, 6643782. [CrossRef]
108. Shi, W.; Wang, Y.; Hou, L.; Ma, C.; Yang, L.; Dong, C.; Wang, Z.; Wang, H.; Guo, J.; Xu, S.; et al. Detection of living cervical cancer cells by transient terahertz spectroscopy. *J. Biophotonics* **2021**, *14*, e202000237. [CrossRef]
109. Xue, P.; Wang, J.; Qin, D.; Yan, H.; Qu, Y.; Seery, S.; Jiang, Y.; Qiao, Y. Deep learning in image-based breast and cervical cancer detection: A systematic review and meta-analysis. *NPJ Digit. Med.* **2022**, *5*, 19. [CrossRef]
110. An, J.M.; Suh, J.; Kim, J.; Kim, Y.; Chung, J.Y.; Kim, H.S.; Cho, S.Y.; Ku, J.H.; Kwak, C.; Kim, H.H.; et al. First-in-Class: Cervical cancer diagnosis based on a urine test with fluorescent cysteine probe. *Sens. Actuators B Chem.* **2022**, *360*, 131646. [CrossRef]
111. Chen, Y.; Wei, J.; Zhang, S.; Dai, H.; Lv, L.; Lin, Y. Photothermal triggered clinical swab point-of-care testing diagnostics: Fluorescence-pressure multi-signal readout detection of cervical cancer biomarker. *Chem. Eng. J.* **2022**, *436*, 135205. [CrossRef]
112. Fekri-Ershad, S.; Ramakrishnan, S. Cervical cancer diagnosis based on modified uniform local ternary patterns and feed forward multilayer network optimized by genetic algorithm. *Comput. Biol. Med.* **2022**, *144*, 105392. [CrossRef]
113. Brenes, D.; Barberan, C.; Hunt, B.; Parra, S.G.; Salcedo, M.P.; Possati-Resende, J.C.; Cremer, M.L.; Castle, P.E.; Fregnani, J.H.; Maza, M.; et al. Multi-task network for automated analysis of high-resolution endomicroscopy images to detect cervical precancer and cancer. *Comput. Med. Imaging Graph.* **2022**, *97*, 102052. [CrossRef]
114. Jia, D.; He, Z.; Zhang, C.; Yin, W.; Wu, N.; Li, Z. Detection of cervical cancer cells in complex situation based on improved YOLOv3 network. *Multimedia Tools Appl.* **2022**, *81*, 8939–8961. [CrossRef]
115. Palanisamy, V.S.; Athiappan, R.K.; Nagalingam, T. Pap smear based cervical cancer detection using residual neural networks deep learning architecture. *Concurr. Comput. Pr. Exp.* **2021**, *34*, e6608. [CrossRef]

116. Li, C.; Zhang, Z.; Tian, Y.; Liu, B. The Value of Three-Dimensional Tomographic Ultrasound Imaging under Backpropagation Algorithm in the Local Staging Diagnosis of Cervical Cancer. *Sci. Program.* **2022**, *2022*, 7017580. [CrossRef]
117. AbuKhalil, T.; Alqaralleh, B.A.Y.; Al-Omari, A.H. Optimal Deep Learning Based Inception Model for Cervical Cancer Diagnosis. *Comput. Mater. Contin.* **2022**, *72*, 57–71. [CrossRef]
118. Waly, M.I.; Sikkandar, M.Y.; Aboamer, M.A.; Kadry, S.; Thinnukool, O. Optimal Deep Convolution Neural Network for Cervical Cancer Diagnosis Model. *Comput. Mater. Contin.* **2022**, *70*, 3295–3309. [CrossRef]
119. RElakkiya, R.; Teja, K.S.S.; Deborah, L.J.; Bisogni, C.; Medaglia, C. Imaging based cervical cancer diagnostics using small object detection—Generative adversarial networks. *Multimedia Tools Appl.* **2021**, *81*, 191–207. [CrossRef]
120. Devi, M.A.; Sheeba, J.; Joseph, K.S. Neutrosophic graph cut-based segmentation scheme for efficient cervical cancer detection. *J. King Saud Univ.-Comput. Inf. Sci.* **2018**, *34*, 1352–1360. [CrossRef]
121. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA A Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef]
122. American Cancer Society. Statistics for Cervical Cancer. *J. Gynecol. Womens Health* **2021**, *2*, 1–9.
123. Das, D.K. Promising Deep Learning-Based CAD System for Cervical Cancer. *Oncol. Times UK* **2022**, *44*, 30. [CrossRef]
124. Purwono, R.R.P.A.; Purwanti, E.; Rulaningtyas, R. Segmentation of cervical cancer CT-scan images using K-nearest neighbors method. *AIP Conf. Proc.* **2020**, *2314*. [CrossRef]
125. Alsmariy, R.; Healy, G.; Abdelhafez, H. Predicting Cervical Cancer using Machine Learning Methods. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 110723. [CrossRef]
126. Mehmood, M.; Rizwan, M.; Ml, M.G.; Abbas, S. Machine Learning Assisted Cervical Cancer Detection. *Front. Public Health* **2021**, *9*, 788376. [CrossRef] [PubMed]
127. Tanimu, J.J.; Hamada, M.; Hassan, M.; Kakudi, H.; Abiodun, J.O. A Machine Learning Method for Classification of Cervical Cancer. *Electronics* **2022**, *11*, 463. [CrossRef]

diagnostics

MDPI

Systematic Review

Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review

Taye Girma Debelee [1,2]

[1] Ethiopian Artificial Intelligence Institute, Addis Ababa 40782, Ethiopia; tayegirma@gmail.com
[2] Department of Electrical and Computer Engineering, Addis Ababa Science and Technology University, Addis Ababa 16417, Ethiopia

Abstract: Skin lesions are essential for the early detection and management of a number of dermatological disorders. Learning-based methods for skin lesion analysis have drawn much attention lately because of improvements in computer vision and machine learning techniques. A review of the most-recent methods for skin lesion classification, segmentation, and detection is presented in this survey paper. The significance of skin lesion analysis in healthcare and the difficulties of physical inspection are discussed in this survey paper. The review of state-of-the-art papers targeting skin lesion classification is then covered in depth with the goal of correctly identifying the type of skin lesion from dermoscopic, macroscopic, and other lesion image formats. The contribution and limitations of various techniques used in the selected study papers, including deep learning architectures and conventional machine learning methods, are examined. The survey then looks into study papers focused on skin lesion segmentation and detection techniques that aimed to identify the precise borders of skin lesions and classify them accordingly. These techniques make it easier to conduct subsequent analyses and allow for precise measurements and quantitative evaluations. The survey paper discusses well-known segmentation algorithms, including deep-learning-based, graph-based, and region-based ones. The difficulties, datasets, and evaluation metrics particular to skin lesion segmentation are also discussed. Throughout the survey, notable datasets, benchmark challenges, and evaluation metrics relevant to skin lesion analysis are highlighted, providing a comprehensive overview of the field. The paper concludes with a summary of the major trends, challenges, and potential future directions in skin lesion classification, segmentation, and detection, aiming to inspire further advancements in this critical domain of dermatological research.

Keywords: skin; cancer; skin disease; skin cancer; melanoma; machine learning; deep learning; detection; segmentation; classification

Citation: Debelee, T.G. Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review. *Diagnostics* **2023**, *13*, 3147. https://doi.org/10.3390/diagnostics13193147

Academic Editors: Wan Azani Mustafa and Hiam Alquran

Received: 30 August 2023
Revised: 22 September 2023
Accepted: 24 September 2023
Published: 7 October 2023

1. Introduction

The evolution of machine learning techniques has impacted many sectors. For instance, breast cancer detection and classification [1,2], diabetes detection and prediction [3,4], and brain tumor detection and classification [5,6] are some of the impacts that machine learning techniques have shown in the health sector in the past few years [7]. The agricultural sector [8,9] and financial sector [10] are also sectors that have benefited from machine learning techniques. In recent years, we have seen significant advancements in dermatology as researchers and clinicians try to understand the complexities of various skin conditions. Skin conditions affect millions of people worldwide and have a substantial impact on both physical health and quality of life. The skin protects our inside organs from microbes, regulates temperature, and serves as a sensation organ [11]. Human skin has three layers: epidermis, dermis, and hypodermis [12]. The epidermis is the outermost layer of skin, which provides a waterproof barrier. It creates our skin tone as it contains special cells called melanocytes, which produce the pigment melanin. The dermis is found under the epidermis and contains tough connective tissues, hair follicles, and sweat glands.

The hypodermis is made of fat and connective tissue. Any of the diseases or disorders that harm these layers of skin can be categorized as skin disease.

Apart from the disability and morbidity caused by skin diseases, skin cancer can be fatal if not treated early. Skin cancer occurs when abnormal cells grow uncontrollably in the skin [13]. According to the American Cancer Society [14], 1.9-million new cancer cases are expected to be diagnosed in 2021. The death rate due to skin cancer in America, during 2021, is predicted to be 1670 deaths per day. Kumar et al. [15] claimed that skin cancer is the most-common cancer in developing countries with the most-advanced diagnostics and prognosis. It has recorded 500,000 new cases in U.S., and that made it be classified as the 19th most-common cancer globally.

It is one of the three most-dangerous and -rapidly expanding cancer types, making it a significant public health issue [16]. One out of every three cancer diagnoses is related to skin cancer, according to the World Health Organization, and the Skin Cancer Foundation reports that the prevalence of skin cancer is rising globally [17]. Both benign and malignant skin tumors can develop from DNA damage caused by exposure to UV light, which results in unregulated cell growth, according to Hasan et al. [18]. Despite their growth, benign tumors do not spread. They include pyrogenic granulomas, cysts, cherry angiomas, seborrheic keratosis, dermatofibromas, skin tags, and dermatofibroma. Malignant tumors, on the other hand, can invade other tissues and organs and spread unpredictably throughout the patient's body. Skin cancer can broadly be categorized as melanoma and non-melanoma skin cancer. Non-melanoma skin cancer includes squamous cell carcinoma, basal cell carcinoma, and Merkel cell carcinoma, among many others. Melanoma arises from pigment-producing cells called melanocytes. If not diagnosed early and managed well, it is very lethal. When detected early, its five-year survival rate is 93%. However, the rate can decrease to 27% after spreading to distant lymph nodes and other organs [14]. That is why due emphasis is being given to screening pigmented lesions:

- Basal cell carcinoma or basalioma (BCC): It begins in the basal cells, the innermost cells of the epidermis, and accounts for around 80% of cases. Although basal cell growth is modest, BCC is typically treatable and does little harm if detected and treated in a timely manner.
- Squamous cell carcinoma or cutaneous spinocellular carcinoma (SCC): This is the primary cause of 16% of skin cancers and develops in the epidermis's outermost layer of squamous cells. Early detection makes it easy treatable, but if left untreated, it can spread to other body regions and penetrate the deeper layers of skin.
- Malignant melanoma (MM): It is a highly severe malignant skin tumor and originates in the melanocytic cells in the epidermis. It spreads quickly, has a high fatality rate because of early metastasis, and is challenging to treat. Although it only causes 4% of skin cancers, it causes death in 80% of instances. Patients with metastatic melanoma have a five-year survival rate of just 14%. It has a 95% cure rate if detected early; therefore, early diagnosis can significantly improve survival prospects.

Traditionally, dermatologists diagnose skin disease by looking at the patient's skin lesions. A dermatoscope (hand-held magnifying lens and built-in light) can be used to better see the area of interest. The revealing characteristics of skin lesions include the size, shape, color, edge, boundary, and location of the abnormality, as well as the presence or absence of other symptoms or signs. Therefore, the experience of the dermatologist can affect the examination process. Besides, skin lesions exhibit similarities in color, texture, edge contour, and other features. If visual inspection of the skin does not provide the doctor with a diagnosis, invasive tests such as a biopsy [14], scraping, etc., are used to identify skin disorders. These processes are also not efficient as they require a large amount of time and affect the patient's curing time. The diagnosis of skin diseases and cancer, which is mainly thorough inspection of the skin lesions, opens room for AI intervention. AI uses pictures of skin lesions to interpret the diagnosis. Recently, there have been several works performed on skin lesion analysis using machine learning and image-processing

techniques. They have been used in many works in the literature for skin lesion attribute identification, segmentation, and disease type detection.

1.1. Contribution

This systematic literature review provides basic technological developments and fundamentals methodically, together with recommendations for researchers. Its contribution can be summarized as follows:

- Comprehensive compilation and analysis of freely accessible and on-demand accessible skin lesion datasets for classification and detection.
- By consolidating research articles published between 2017 and 2023, this study presents vital perspectives on the detection, segmentation, and classification of skin lesions.
- This survey summarized and evaluated the contributions and limitations of the past survey papers in the domain of skin classification and detection, which were published between 2017 and 2023.
- This recapitulation outlines unaddressed research needs, offering a concise summary of the unresolved research challenges and potential avenues for further exploration in skin lesion categorization and detection across diverse skin datasets.
- The study paper indicated that, in recent years, the accuracy of skin image analysis using machine learning approaches has grown, leading it to being viewed as a complimentary approach to clinical evaluation.

1.2. Paper Organization

The rest of the paper is organized as follows, Section 2 discusses recent related works, and Section 3 presents the methodology employed in this paper to perform the systematic literature review. In Section 4, we present the main public databases containing dermoscopic images, relevant for most of the studies carried out previously for skin disease diagnosis. The machine learning techniques applied to skin disease classification are presented in Section 5.1. The commonly used machine learning techniques for skin disease detection are discussed in Section 5.2. The findings of the systematic literature review from a different perspective is presented in Section 6, and a brief discussion on open challenges and future directions is provided in Section 7. Finally, the conclusion of this systematic review paper is presented in Section 8.

2. Related Work

Skin lesion classification and detection have emerged as critical areas of research in medical imaging and computer vision, with the potential to revolutionize the early diagnosis and treatment of various skin disorders, including skin cancer. In this discussion, we delve into the existing body of related work on skin lesion classification and detection, exploring the methodologies, approaches, and advancements that researchers have employed in this domain.

Grignaffini et al. [16] conducted a systematic review of the prior literature on the use of machine learning techniques for the detection and classification of skin cancer from various datasets (MedNode, ISIC2017, HAM10000, ISIC2016, PH2, DermIS, DermQuest, ISIC archive, IDS, ISIC 2019, ISIC2020, ISIC2018, 7-point checklist, and DermNZ). After examining a total of 68 research papers, the authors provided a complete summary of the various machine learning techniques utilized for skin cancer categorization, along with their performance metrics. The methods, findings, and introduction are all divided into separate sections in this well-organized piece of work. Because the authors thoroughly outlined the inclusion and exclusion criteria used to choose the research, the review is more reliable. The authors also used a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram to show the study-selection process. The Results Section of the study, together with performance metrics, provides a complete analysis of the various machine learning techniques used for skin cancer categorization. After carefully comparing all available techniques, the authors identified the best ones. The authors also

examined the limitations of earlier studies and highlighted the need for more-dependable and -accurate methodologies.

Zafar et al. [19] provided an overview of the many techniques for analyzing skin lesions and diagnosing cancer that have been published in the literature. This review article featured studies on the diagnosis of skin lesions using several datasets (MedNode, ISIC2017, HAM10000, ISIC2016, PH2, DermIS, DermQuest, ISIC 2019, ISIC2020, ISIC2018, 7-point checklist, HPH, ISIC archive, ISBI 2016, ISBI 2017, and DermNZ) from different repositories. The report provided an outstanding summary of the various approaches that have been proposed for the examination of skin lesions and cancer diagnosis.

In a thorough review, Hauser et al. [20] investigated the use of explainable artificial intelligence (XAI) in the identification of skin cancer, and XAI refers to artificial intelligence models that can describe their decision-making processes in order to aid doctors in better understanding and interpreting the model's predictions. The authors conducted a thorough search of the literature on Google Scholar, PubMed, IEEE Explore, Science Direct, and Scopus and discovered 37 articles that used XAI techniques in skin cancer diagnosis. The authors discussed the various XAI techniques used in the trials, including decision trees, gradient-based strategies, and rule-based models. They drew attention to both the advantages of XAI, including enhanced transparency and interpretability, and its potential drawbacks, such as decreased accuracy when compared to black-box models. The authors drew the conclusion that XAI has the potential to improve skin cancer detection by providing more-transparent and -understandable models. They also emphasized the need for more research to demonstrate the viability of XAI models in clinical settings and to address the challenges of integrating these models into pre-existing healthcare systems.

In their study, Jeong et al. [21] aimed to examine the current approaches, outcomes, and restrictions of deep learning in dermatology. The investigation included studies published between 2015 and 2021, and the authors discovered 65 papers that met their criteria for inclusion. The various deep learning techniques used, the types of dermatological conditions looked into, and the performance standards used to rate the models were discussed. The authors' in-depth analysis of how deep learning techniques are applied in dermatology is a significant contribution to the field. The survey article provided links to other datasets that researchers may utilize, which were discovered. It is conceivable that additional important studies were overlooked because the assessment was restricted to works published between 2015 and 2021. While acknowledging the limitations of deep learning in dermatology, it would have been helpful to provide more-specific recommendations for future research to address these limitations.

Hasan et al. [22] conducted a thorough examination of 594 papers, 356 of which were for skin lesion segmentation and 238 for skin lesion classification. Furthermore, they evaluated and investigated potential segmentation and classification patterns for skin lesions. Important details regarding the procedures used to create CAD systems were provided by analyzing and summarizing these articles in a variety of ways. They included the method configurations (techniques, architectures, module frameworks, and losses), training methods, assessment methods, and input data, which included dataset usage, data preprocessing, augmentations, and addressing imbalanced concerns.

Relevant and essential definitions and theories were also included in this list. The aim of the researchers was to study several performance-improving strategies, such as ensemble and postprocessing. The main challenges of evaluating skin lesion segmentation and classification algorithms using small datasets were addressed, along with some potential solutions. They also discussed these dimensions to disclose their current trends based on usage frequencies.

A critical analysis of a few cutting-edge machine learning methods for skin cancer detection was presented by Bhatt et al. [23]. The importance of early melanoma skin cancer detection was also stressed by the authors because it significantly increases survival rates. The scientists also offered a comprehensive overview of the most-recent machine learning techniques for melanoma skin cancer detection and classification. The authors covered a

variety of subjects in-depth, including different algorithms (support vector machine, K-nearest neighbors, and CNN), data augmentation, and feature-extraction techniques using datasets such as PH2, MEDNODE, Dermofit, Dermquest, and others compiled from the archives of the ISIC and ISBI. There were, however, some restrictions in the work. In the study, bias in training data and other potential negative effects of using machine learning algorithms for melanoma detection were not addressed. The potential ethical repercussions of using machine learning algorithms for medical diagnosis were also not addressed by the authors.

An overview of numerous skin disease categorization methods based on machine learning approaches was presented by Mohammed and Al-Tuwaijari [24]. Support vector machines, decision trees, random forests, artificial neural networks, and deep learning were just a few of the methods covered in the study. The technique and performance indicators employed in these systems were also covered in the article. The approaches for classifying skin diseases using machine learning algorithms were well-explained in this paper. It did not, however, offer a thorough analysis of the methodologies surveyed. The survey might not include all methods for classifying skin diseases that are currently in use. Furthermore, a deeper examination and evaluation of the examined methodologies would have improved the paper's value.

A study on the possibility of deep learning and machine learning techniques for the early identification of skin cancer was reported by Mazhar et al. in the publication [25]. The authors reviewed the pertinent literature on skin cancer detection and the application of artificial intelligence (AI) in the healthcare industry using a systematic manner. The study offered a thorough assessment of the state-of-the-art in skin cancer diagnosis today, as well as the potential of AI to boost accuracy, cut down on waiting times, and increase access to healthcare services. A thorough explanation of the many methodologies employed in the study, such as convolutional neural networks (CNNs), was also provided. The authors may have discussed the problems with data quality, data imbalance, data bias, and the necessity for big datasets to train deep learning models. The study may have been made stronger by comparing the effectiveness of machine learning and deep learning approaches to conventional methods for skin cancer detection. Furthermore, a section on the ethical issues surrounding the use of AI in healthcare, particularly in relation to patient data security and privacy, may have been added to the article. Overall, the paper provided a great summary of the application of deep learning and machine learning to the identification of skin cancer.

Table 1 presents a summary of related studies, highlighting both contributions and limitations of related studies.

Table 1. Summary of related works with their contributions and limitations.

Author and Year of Publication	Contribution	Limitation
Grignaffini et al. [16], 2022	• It reported a systematic literature review of recent research on the use of machine learning to classify skin lesions. • The paper discussed datasets that are used commonly in skin lesion detection and classification.	• Explaining the basic concepts of traditional ML algorithms is not important for the readers of such papers. • Explaining the basic concepts of DL is also not important to readers that this paper targets.
Zafar et al. [19], 2023	• A thorough assessment of the literature on the methods, procedures, and approaches used to examine skin lesions was made in this paper.	• Research publications that analyzed skin lesions based on their complicated and uncommon features were excluded.
Hauser et al. [20], 2022	• A systematic review of explainable artificial intelligence (XAI) in skin cancer recognition was made in this paper.	• Limited to only XAI, it may have not covered all aspects of AI in skin cancer recognition. • It did not compare XAI with others traditional ML and DL techniques in terms of various evaluation parameters.

Table 1. *Cont.*

Author and Year of Publication	Contribution	Limitation
Jeong et al. [21], 2023	• Provided a thorough overview of dermatology in the literature review. • A summarized review report of the current state of the datasets, transfer learning strategies, difficulties, and restrictions within the body of existing AI work was presented.	• The survey was limited to papers published between 2015 and 2021. • It overlookedpertinent studies.
Hasan et al. [22], 2023	• The authors revealed that the ISIC is the most-commonly applied dataset in skin disease segmentation and classification. • The survey enables researchers to determine the best experimental setup for skin lesion diagnosis.	• Most of the publications (100+) considered in the papers were those published 6 years, between 2011 and 2016.
Bhatt et al. [23], 2022	• Detailed analysis of cutting-edge machine learning methods for the identification and categorization of melanoma skin cancer.	• Primarily focused on conventional machine learning methods for the identification and classification of melanoma skin cancer.
Mohammed and Al-Tuwaijari [24], 2022	• Provided a summary of various classification systems for skin diseases based on machine learning methods.	• Did not offer a methodical assessment of the approaches and did not include all of them.
Mazhar et al. [25], 2023	• Reviewed the pertinent literature on skin cancer detection and the application of artificial intelligence (AI) in healthcare.	• Ethical issues surrounding the use of AI in healthcare, particularly in relation to patient data security and privacy, may have been added to the article.

3. Methods

In this systematic review approach for ML-based skin disease detection and classification, we defined the research questions, search strategies with the search databases, and paper selection criteria. In order to analyze current research findings that have been suggested for skin disease detection and classification using conventional machine learning methods, deep learning methods, and hybrid methodologies, this systematic literature review work set three main objectives: (1) to identify the commonly available datasets that could be accessed freely or upon request; (2) to explore the contribution and limitations of the current state-of-the-art methods; (3) to present the summary of the open challenges in the area of skin disease and cancer detection and classification.

In this systematic review, we defined a rigorous research question that can summarize the body of literature already available on a skin lesion detection and classification, enabling a thorough and objective understanding of this topic. Several methods and processes were used to make sure the study was effective and true to its original intent. We examined the essentials of a systematic review or survey in this thorough explanation, focusing on five pre-established research questions, search strings, five inclusion and six exclusion criteria, and five search engines or databases.

Research questions:Any systematic review or survey must start with a set of clear research questions as its cornerstone. These inquiries direct the entire research procedure and aid in defining the study's scope. Five research topics were already established in this case as presented in Table 2. These inquiries were made to be precise, short, and geared

towards the review's particular goals to provide a guide for the methodical gathering and examination of pertinent facts.

Search strings: Creating search phrases or keywords is a crucial step in the systematic review process. These carefully constructed search strings are used to look up scientific papers across a variety of databases or search engines. They ought to be planned to include all pertinent material pertaining to the study's questions. Search strings that combine synonyms, Boolean operators, and truncation symbols make sure that the review is thorough and does not overlook any important studies. Algorithm 1 presents how the search strings were combined to collect the appropriate scientific papers specific to the pre-defined topic.

Inclusion and exclusion criteria: Pre-specified inclusion and exclusion criteria were developed as presented in Table 3 to preserve the caliber and applicability of the papers included in the review. Five inclusion criteria and six exclusion criteria were established in this case. The inclusion criteria specified the qualities that papers must have in order to be taken into account for the review, such as the time period between publications, the type of study, which was specific to the topic, the reputability of the journals where the scientific papers were published, and the language of the study. On the other hand, the exclusion criteria outlined the circumstances under which an article would be disregarded, such as non-English language publications or research that poses a significant risk of bias such as M.Sc. and Ph.D. theses, seminars, posters, case studies, and publications before 2020. These standards aided in ensuring that the review concentrated on the most-pertinent and -methodologically reliable studies.

Search engines or databases: In systematic reviews, the choice of the search engines or databases is also crucial. Utilizing several databases increases the chance of finding a wide variety of pertinent publications. Five search engines or databases were chosen in this systematic review process, as indicated in Figure 1. IEEE Xplore, MDPI, Google Scholar, Springer Link, and Science direct are a few popular databases for scientific literature. The evaluation reduced the chance of missing important findings by searching across different platforms.

These methods worked together to make sure that the systematic review process was orderly, impartial, and able to offer solid, evidence-based insights into the chosen study field, as presented in Algorithm 2.

Table 2. Researchquestions for the systematic review work.

No.	Research Question	Objectives
1	What are the major targets of applying machine learning techniques in skin disease diagnosis?	To investigate the major targets of applying traditional machine learning and deep learning approaches for skin disease diagnosis.
2	What machine learning techniques are used in skin disease diagnosis?	To identify the commonly and recently proposed traditional machine learning techniques and deep learning techniques for skin disease diagnosis.
3	What are the common available dataset for skin disease diagnosis?	To identify the publicly available dataset that researchers can download either by online registration or freely available with or without payment.
4	How successful are the proposed machine learning techniques in skin disease diagnosis?	To analyze and compare the proposed machine learning techniques in the diagnosis of skin disease.
5	What are the future challenges in applying machine learning techniques for skin disease diagnosis?	To investigate the open questions in the application of machine learning techniques for skin disease diagnosis.

Table 3. Inclusion and exclusion criteria for paper selection.

Inclusion Criteria (ICs)	Exclusion Criteria (ECs)
IC1: The papers should focus on skin disease or cancer detection or segmentation or classification	EC1: Publications that are not focused on skin disease and cancer detection or classification or segmentation.
IC2: The papers should include different types of diseases, mainly skin cancer or melanoma.	EC2: Publications not peer-reviewed, abstracts, editorial letters, book reviews, and scientific reports.
IC3: The papers should be published in reputable journals with an impact factor and indexed in the Web of Science or Scopus or recognized conference proceedings.	EC3: M.Sc. and Ph.D. theses, posters, and seminars.
IC4: The studies should be written in English.	EC4: Studies that are published prior to 2020 except for Sections 1 and 4.
IC5: Publication year for Sections 1 and 4 can be any year of publication.	EC5: Skin disease detection and classification based on case studies.
IC6: The publication year for study paper to be included in the systematic review must be between 2020 and 2023	EC6: Study papers that are not peer-reviewed and journals that are not indexed in the Web of Science or Scopus.

Algorithm 1 Pseudocode for defining the search string

Search_String = [("Skin" **OR** "Skin disease" **OR** "Skin cancer" **OR** "Melanoma" **OR** "Skin lesion"
AND
("Machine Learning Methods" **OR** "Machine Learning Techniques"**OR** "Deep Learning Methods"
OR "Deep Learning Techniques" **OR** "Classical Machine Learning Methods" **OR** "Traditional Machine Learning Methods"
AND
("Detection" **OR** "Classification" **OR** "Segmentation")]

Algorithm 2 Pseudocode for generating potential review papers

SearchDatabases ← *Springer_Link*, *MDPI*, *Science_Direct*, *Wiley_Online*, *IEEE_Xplore*, *Google_nScholar*
{**Initialization:**}
Area_Keyword ← [*Skin*, *Skin_disease*, *Skin_cancer*, *Melanoma*]
Method_keywords ← [*Deep_Learning_Methods*, *Classical_Machine_Learning_Methods*, *HybridMethod*]
Target_keywords ← [*Detection*, *Classification*, *Segmentation*]
Search_String ← "*Algorithm* 1"
for *keyword* ∈ *Area_keywords* **do**
  for *target* ∈ *Target_keywords* **do**
    for *method* ∈ *Method_keywords* **do**
      Search_String = *Algorithm* 1
      for *database* ∈ *Databases* **do**
        *List*1 ← *databases*.**search**(*Search_String*)
      end for
    end for
  end for
end for
Inclusion_Criteria = [IC1,IC2,IC3,IC4,IC5,IC6]
Exclusion_Criteria = [EC1,EC2,EC3,EC4, EC5,EC6]
List2 ← *Apply*.*Inclusion_Critera*(*List*1)
Final_Lists ← *Apply*.*Inclusion_Critera*(*List*2)

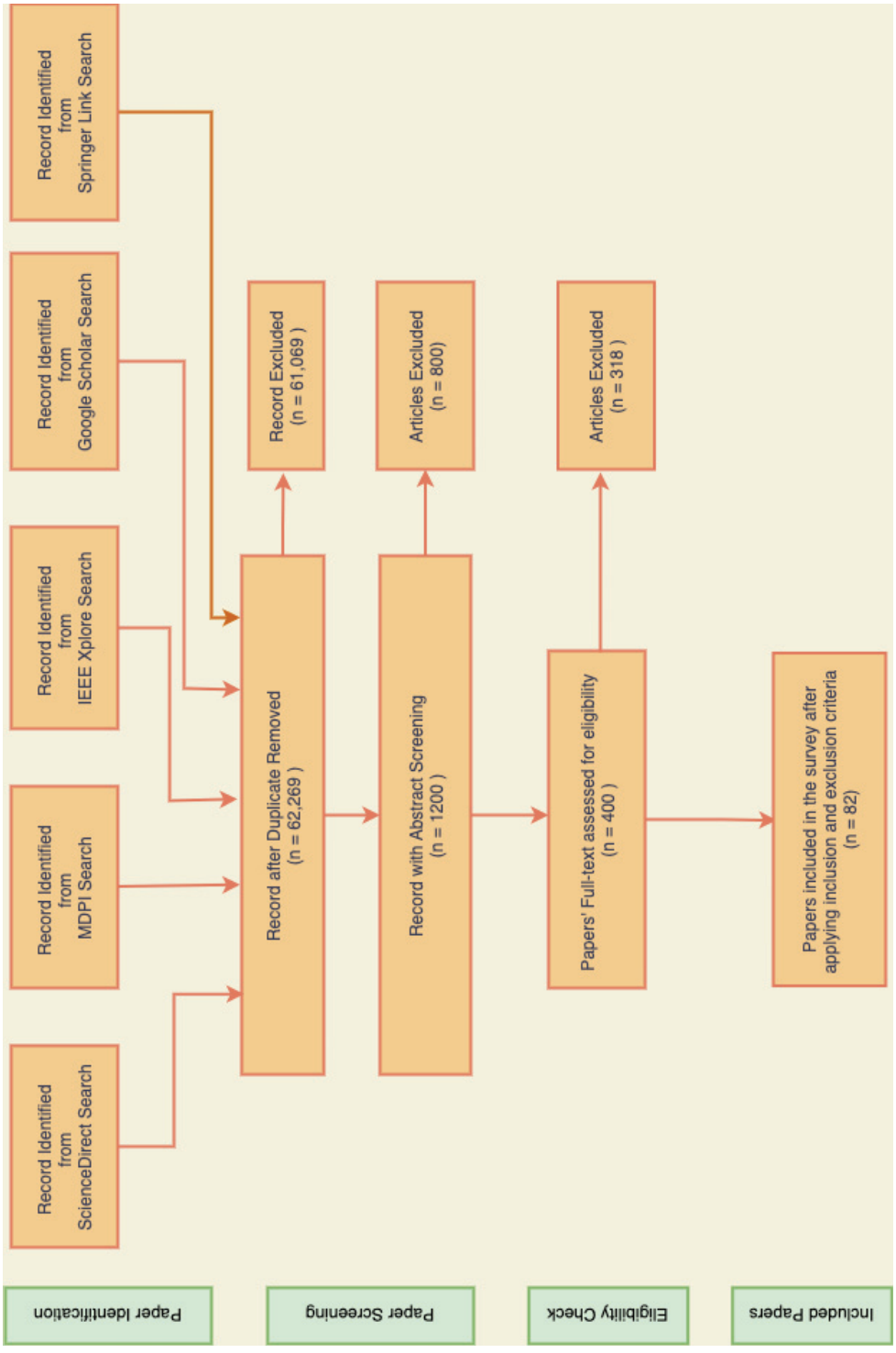


Figure 1. Flow diagram for study papers' search and selection strategy from different databases.

4. Skin Datasets

Skin lesion datasets are a useful tool for the development of algorithms to identify and categorize various forms of skin lesions in the fields of dermatology and computer vision. Various skin disorders, including benign and malignant lesions, are represented by a collection of images and labels. The "ISIC (International Skin Imaging Collaboration) Archive" is a well-known dataset of skin lesions [26,27]. The vast majority of the images in this dataset are dermoscopic images, which are polarized and enlarged views of skin lesions taken with specialized dermatology equipment. The ISIC Archive has been extensively utilized in research to create automated algorithms for melanoma diagnosis and skin lesion classification.

The "HAM10000" dataset is another important source of data [26,27]. It consists of 10,015 dermoscopic images of pigmented skin lesions divided into seven groups, including basal cell carcinoma, melanoma, and nevi. For the categorization of skin lesions, deep learning models have been developed using the HAM10000 dataset in a number of research works. A dataset that is exclusively devoted to melanocytic lesions is the "PH2 Dataset" [26,27]. It includes 200 dermoscopic photos of normal and uncommon melanocytic lesions, coupled with a ground truth that has been expertly annotated for precise diagnosis. In order to create algorithms that aid in the early identification and diagnosis of melanoma, the PH2 dataset has been extensively used. The development of computer-aided diagnosis and automated skin lesion classification has greatly benefited from these skin lesion datasets, as well as others that are available in the literature. These datasets are still being used by scientists and doctors to improve skin disease diagnosis and treatment by creating more-precise and -effective algorithms for skin lesion analysis.

In Tables 4–6, the most-popular dermoscopic public datasets are summarized, and the details of these datasets can be found in [26,27], while the website to download each dataset is also presented in Table 7.

Table 4. Summary of openly accessible datasets—ISIC Archive [26–28].

Dataset	Collection Site	Publication Year	Imaging Modality	Number of Category	Number of Images
ISIC2020	Hospital Clinic Barcelona	2020	Dermoscopic	2	7311
ISIC2020	University of Queensland	2020	Dermoscopic	-	8449
ISIC2020	Medical University Vienna	2020	Dermoscopic	2	4374
ISIC2020	Memorial Sloan Kettering Cancer Center	2020	Dermoscopic	5	11,108
ISIC2020	Sydney Melanoma Diagnosis Centre and Melanoma Institute Australia	2020	Dermoscopic	8	1884
HAM10000	Medical University of Vienna and skin cancer practice of Cliff Rosendahl in Queensland	2018	Dermoscopic	8	10,015
BCN20000	Hospital Clinic Barcelona	2019	Dermoscopic	9	19,424
JID	*Journal of Investigative Dermatology*	2018	Macroscopic	3	100
MSK 1-5	Memorial Sloan Kettering Cancer Center	2015 and 2017	Dermoscopic	15	3918
UDA	Google Research, Brain Team, and Carnegie Mellon University	2014 and 2015	Dermoscopic	7	617

Table 5. Summary of openly accessible datasets—ISIC Challenge test set and non-ISIC datasets [26–28].

Dataset	Collection Site	Publication Year	Imaging Modality	Number of Category	Number of Images
ISIC2020	Test set	2020	Dermoscopic	-	10,982
ISIC2019	Test set	2018 and 2019	Dermoscopic	-	8238
ISIC2018	Test set	2018	Dermoscopic	-	1000
PAD-UFES-20	Non-ISIC set	2020	Macroscopic	6	2298
PH2	Dermatology Service of Pedro Hispano Hospital	2013	Dermoscopic	3	200
7-point criteria evaluation database	Dr. Giuseppe Argenziano	2018	Dermoscopic and Macroscopic	15	2013
MED-NODE	University Medical Center Groningen (UMCG)	2015	Macroscopic	2	170
SKINL2	Instituto de Telecomunicações Campus Universitário de Santiago	2019	Light field photographs and dermoscopic photographs	8	814
SNU	University of Waterloo	-	Macroscopic	2	206
SDN-260	-	2019	Macroscopic	260	20,600

Table 6. Summary of non-openly accessible datasets, but downloaded upon request [26–28].

Dataset	Collection Site	Publication Year	Imaging Modality	Number of Category	Number of Images
Asan	Asan Institutional	2017	Macroscopic	12	17,125
Hallym	Asan Institutional	2017	-	1 152	
DERMOFIT image library	University of Edinburgh	-	-	10	1300
IMA205	-	2018	-	-	-
MoleMapper app patient photo	Ph.D. cancer biologist, Dan Webster	2017	Macroscopic	2	2422
SNU	University of Edinburgh	2018	Macroscopic	134	2201
Severance	-	2020	Macroscopic	43	40,331
Papadakis	-	2021	Macroscopic	1	156

Table 7. Datasets and associated download websites [26–28] accessed on 20 July 2023.

Dataset	Website
Monkeypox-dataset-2022	https://github.com/mahsan2/Monkeypox-dataset-2022
ACNE04	https://github.com/xpwu95/LDL
BCN_20000 (part of ISIC 2019)	https://challenge2019.isic-archive.com/data.html
Asan and Hallym	https://figshare.com/articles/Asan_and_Hallym_Dataset_Thumbnails_/5406136

Table 7. *Cont.*

Dataset	Website
Atlas Dermatologico	http://www.atlasdermatologico.com.br
DermQuest	http://dermquest.com
DermAtlas	http://www.dermatlas.net
Dermis	http://www.dermis.net/dermisroot/en/home/index.htm
Dermnet	http://www.dermnet.com
DermWeb	http://www.dermweb.com
Dermnet NZ	https://www.dermnetnz.org
Dermatoweb	http://www.dermatoweb.net
Danderm	http://www.danderm-pdv.is.kkh.dk/atlas/index.html
Dermatologia Praktyczna	http://derma.pl/
DermSynth3D	https://github.com/sfu-mial/DermSynth3D
DermX (525 dermatological images with diagnoses and diagnosis explanations by three dermatologists)	https://github.com/ralucaj/dermx
Dermofit	https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html
Derm7pt	http://derm.cs.sfu.ca/
Diverse Dermatology Images (DDI)	https://ddi-dataset.github.io
ENriching Health data by ANnotations of Crowd and Experts (ENHANCE): ABC criteria annotations of ISIC 2017 and PH2 datasets	https://github.com/raumannsr/ENHANCE
Fitzpatrick17k (16,577 clinical images with diagnosis and Fitzpatrick scale labels)	https://github.com/mattgroh/fitzpatrick17k
HAM10000	https://www.nature.com/articles/sdata2018161
Hellenic Derm Atlas	http://www.hellenicdermatlas.com/en
ISIC	https://isic-archive.com/
Islam et al. Monkeypox Skin Image Dataset 2022	www.Kaggle.com/datasets/arafathussain/monkeypox-skin-image-dataset-2022
MedMNIST	https://medmnist.com
Med-Node	http://www.cs.rug.nl/~imaging/databases/melanoma_naevi/
Meddean	http://www.meddean.luc.edu/lumen/MedEd/medicine/dermatology/melton/atlas.htm
MoleMap	https://molemap.co.nz
MSK	https://arxiv.org/abs/1710.05006
PH2	https://www.fc.up.pt/addi/ph2%20database.html
PAD-UFES-20 (clinical skin lesion images from smartphones)	https://data.mendeley.com/datasets/zr7vgbcyr2/1

Table 7. *Cont.*

Dataset	Website
Skin3D	https://github.com/jeremykawahara/skin3d
SD198	https://drive.google.com/file/d/1YgnKz3hnzD3umEYHAgd29n2AwedV1Jmg/view
Skin Cancer Detection	https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection
SKINCON	https://skincon-dataset.github.io/index.html
University of Iowa Clinical Skin Disease Images	http://www.medicine.uiowa.edu/dermatology/diseaseimages/
UWaterloo Skin Cancer Detection dataset (images taken from DermIS and DermQuest along with lesion segmentation)	https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection
XiangyaDerm	http://airl.csu.edu.cn/xiangyaderm/

5. Machine-Learning-Based Skin Disease Detection and Classification

In this section, the important discoveries, trends, and knowledge gaps that have been identified from prior research are highlighted thorough an examination of the body of literature on skin diseases. This study sought to provide a thorough overview of the present understanding in the topic and highlight prospective directions for further research by integrating the collective knowledge from a variety of sources.

5.1. Machine Learning and Deep Learning in Skin Disease Classification

Skin diseases and skin cancer pose significant health concerns worldwide. Early detection and accurate classification of these conditions are crucial for effective treatment and improved patient outcomes. In recent years, the field of dermatology has witnessed remarkable advancements in the development of automated systems for skin disease classification. These systems leverage the power of artificial intelligence (AI) and machine learning techniques to analyze dermatological images and provide reliable diagnoses. The classification of skin diseases and skin cancer traditionally relied on manual examination and subjective interpretation by dermatologists. However, the subjective nature of this process often led to inconsistencies and errors in diagnosis. With the advent of computer-aided diagnosis (CAD) systems, the dermatology community has gained access to powerful tools that can enhance diagnostic accuracy and assist healthcare professionals in decision-making.

This review report aimed to explore the latest advancements in the field of skin disease and skin cancer classification using AI-based approaches. We examined the methodologies employed in various studies, including deep learning algorithms, convolutional neural networks (CNNs), and image analysis techniques. By analyzing the strengths and limitations of these approaches, we can gain insights into the current state-of-the-art and identify areas for further improvement.

Balaji et al. [29] presented a method for skin disease detection and segmentation using the dynamic graph cut algorithm and classification through a Naive Bayes classifier. The authors first segmented the skin lesion from the background using the dynamic graph cut algorithm and, then, used texture and color features to classify the skin lesion into one of several categories using the Naive Bayes classifier. The use of both a dynamic graph cut algorithm and a Naive Bayes classifier provides a robust and accurate method for identifying and classifying skin lesions. The authors provided a clear description of the methodology used and the results obtained, including a comparison with existing methods. The authors evaluated their proposed approach using the ISIC 2017 dataset and reported an accuracy of 91.7%, a sensitivity of 70.1%, and a specificity of 72.7%. The dataset is available

publicly on the ISIC website for public studies. The approach also scored an accuracy of 94.3% for benign cases, 91.2% for melanoma, and 92.9% for keratosis.

Ali et al. [30] presented a study on the application of EfficientNets for multiclass skin cancer classification, with the aim of contributing to the prevention of skin cancer. The authors utilized the HAM10000 dataset consisting of 10,015 skin lesion images from seven different classes and compared the performance of different variants of EfficientNets with traditional deep learning models. The proposed approach was mainly focused on the transfer learning technique using an EfficientNet and showed promising results for the evaluation parameters that were selected during the experimental analysis. The authors presented an interesting and relevant study on the application of EfficientNet Variants B0–B7 for skin cancer classification. However, B0 was the best-performing model of the EfficientNets out of B0–B7 with an accuracy of 87.9%. The model was also evaluated in terms of other evaluation parameters and achieved a precision of 88%, a recall of 88%, an F1-score of 87%, and an AUC of 97.53%. Overall, the experimental results of this paper suggested that the proposed skin cancer classification model based on EfficientNets can accurately classify skin cancer and has the potential to be a useful tool in the prevention and early detection of skin cancer. However, the study had a few limitations. Firstly, the authors did not provide a detailed comparison of their results with other state-of-the-art methods for skin cancer classification, which makes it difficult to assess the significance of their findings. Secondly, the study only used a single dataset, which may limit the generalizability of the results. Future studies could benefit from using multiple datasets and exploring the transferability of the models to different domains. Lastly, the authors did not provide any information on the computational resources required for training and evaluating the models, which could be useful for researchers and practitioners looking to replicate or adapt their approach.

Srinivasu et al. [31] presented a classification approach for skin disease detection using deep learning neural networks with MobileNet V2 and LSTM. The proposed approach involved preprocessing of skin images followed by feature extraction using MobileNet V2 and classification using LSTM. More than 10,000 skin photos made up the dataset utilized for evaluation. These images were divided into fivedifferent skin diseases: melanocytic nevi (NV), basal cell carcinoma (BCC), actinic keratoses and intraepithelial carcinoma (AKIEC), dermatofibroma (DF), and melanoma (MEL). The contribution of the paper was the development of an accurate skin disease classification approach using deep learning techniques that were lightweight and required less computational time. The main limitation of the proposed approach was that it requires a large amount of data to train the model effectively. The proposed approach achieved an accuracy of 90.21%, which outperformed the existing approaches VGG16, AlexNet, MobileNet, ResNet50, U-Net, SegNet, DT, and RF. The sensitivity, recall, and specificity values for each skin disease category were also reported, which further validated the effectiveness of the proposed approach. The results demonstrated that the proposed approach can accurately classify skin diseases, which can aid in the early diagnosis and treatment of such diseases.

Shetty et al. [32] presented a novel approach for skin lesion classification using a convolutional neural network (CNN) and machine learning techniques. The authors aimed to develop an accurate and automated system for the classification of dermoscopic images into different categories of skin lesions. The paper's contribution lied in the development of a CNN-based skin lesion classification system that can accurately classify seven different categories of skin lesions with high accuracy. The authors utilized publicly available datasets and compared their proposed system's performance with other state-of-the-art systems such as EW-FCM + Wide-shuffleNet, shifted MobileNet V2, Shifted GoogLeNet, shifted 2-Nets, Inception V3, ResNet101, InceptionResNet V2, Xception, NASNetLarge, ResNet50, ResNet101 + KcPCA + SVMRBF, VGG16 + GoogLeNet ensemble, and Modified-MobileNet and outperformed all in terms of accuracy. The limitation of this paper was that the proposed system is limited to dermoscopic images, and it cannot classify clinical images, which are more challenging to classify due to their low contrast and other artifacts.

The experimental results showed that the proposed system achieved an overall accuracy of 95.18%, outperforming other state-of-the-art systems.

Jain et al. [33] proposed a multi-type skin disease classification algorithm using an optimal path deep-neural-network (OP-DNN)-based feature extraction approach. The proposed algorithm achieved improved accuracy compared to other state-of-the-art algorithms for the classification of various skin diseases. The contribution of this paper was the proposal of an OP-DNN-based feature extraction approach for multi-type skin disease classification. This approach improved the accuracy of classification and also reduced the number of features required for classification. The paper also provided experimental results that demonstrated the effectiveness of the proposed approach. The algorithm was evaluated on the ISIC dataset with 23,906 skin lesion images and achieved an accuracy of 95%, which outperformed other algorithms such as KNN, NB, RF, MLP, CNN, and LSTM for multi-type skin disease classification.

Wei et al. [34] proposed a novel skin disease classification model based on DenseNet and ConvNeXt fusion. The proposed model utilized the strengths of both DenseNet and ConvNeXt to achieve better performance in skin disease classification. The model was evaluated on two different datasets, where one is the publicly available HAM10000 dataset and the other was the dataset from Peking Union Medical College Hospital, and it achieved superior performance compared to the other models. The proposed model addresses the limitations of previous models by combining the strengths of the DenseNet and ConvNeXt architectures, which has not been explored before in skin disease classification. The model achieved state-of-the-art performance on the HAM10000 dataset and can potentially be used in clinical settings to assist dermatologists in diagnosing skin diseases. However, the study did not provide any explanation of how the model's decisions were made, which may limit its interpretability in a clinical setting. The proposed model achieved an accuracy of 95.29% on the HAM10000 dataset and 96.54% on the Peking Union Medical College Hospital dataset, outperforming the other state-of-the-art models. The model also achieved high sensitivity and specificity for all skin disease categories. The study also conducted ablation experiments to show the effectiveness of the proposed fusion approach, which outperformed the individual DenseNet and ConvNeXt models.

Almuayqil et al. [35] presented a computer-aided diagnosis system for detecting early signs of skin diseases using a hybrid model that combines different pretrained deep learning models (VGG19, InceptionV3, ResNet50, DenseNet201, and Xception) with traditional machine learning classifiers (LR, SVM, and RF). The proposed system consists of four main steps: preprocessing the input raw image data and metadata; feature extraction using six pretrained deep learning models (VGG19, InceptionV3, ResNet50, DenseNet201, and Xception); features concatenation; classification using machine learning techniques. The proposed hybrid system was evaluated on the HAM10000 dataset of skin images and showed promising results in detecting skin diseases accurately. However, the proposed hybrid approach DenseNet201 combined with LR achieved better performance with an accuracy of 99.94% in detecting skin diseases, which outperformed the other state-of-the-art approaches. The authors also provided a detailed comparison of the proposed model with other state-of-the-art methods, showing its superiority in terms of accuracy and other evaluation metrics.

Reddy et al. [36] proposed a novel approach for the detection of skin diseases using optimized region growing segmentation and autoencoder-based classification. The proposed approach employs an efficient segmentation algorithm that can identify the affected regions of the skin accurately. Subsequently, a convolutional autoencoder-based classification model was used to classify the skin diseases based on the extracted features. The experimental results indicated that the proposed approach achieved promising results and outperformed several state-of-the-art methods in terms of accuracy and other evaluation metrics. The proposed approach offers several contributions to the field of skin disease detection. Firstly, the proposed segmentation algorithm is optimized for skin disease detection and can accurately identify the affected regions of the skin. Secondly,

the proposed autoencoder-based classification model can classify the skin diseases with high accuracy using the extracted features. Lastly, the proposed approach outperformed several state-of-the-art methods in terms of accuracy and other evaluation metrics. One of the limitation of the approach is that it may not generalize well to new datasets with different characteristics as the the model was evaluated on the small dataset used from PH2 with 200 images. The experimental results indicated that the proposed approach achieved an accuracy of 94.2%, which outperformed several state-of-the-art methods. The proposed approach also achieved high values for other evaluation metrics such as the precision, recall, and F1-score, which demonstrated the effectiveness of the proposed approach for skin disease detection.

Malibari et al. [37] presented an optimal deep-neural-network-driven computer-aided diagnosis (ODNNsingle bondCADSCC) model for skin cancer detection and classification. The Wiener-filtering (WF)-based preprocessing step was used extensively in the described ODNNsingle bondCADSCC model, which was then segmented using U-Net. Moreover, the SqueezeNet model was used to produce a number of feature vectors. Eventually, effective skin cancer detection and classification were achieved by using the improved whale optimization algorithm (IWOA) with a DNN model. IWOA is used in this technique to effectively choose the DNN settings. The comparison study findings demonstrated the suggested ODNNsingle bondCADSCC model's promising performance against more-recent techniques with a high accuracy of 99.90%. Although the results are promising, it would be helpful to validate the proposed model on a larger dataset to assess its robustness and generalization capabilities. Another limitation is that the proposed model does not provide explanations for its decisions, which is essential for gaining the trust of clinicians and patients.

Qian et al. [38] proposed a deep convolutional neural network dermatoscopic image classification approach that groups multi-scale attention blocks (GMABs) and uses class-specific loss weighting. To increase the size of the DCNN model, the authors introduced GMABs to several scale attention branches. Hence, utilizing the GMABs to extract multi-scale fine-grained features will help the model better be able to focus on the lesion region, improving the DCNN's performance. The attention blocks, which may be used in different DCNN structures and trained end-to-end, have a straightforward structure and a limited number of parameters. The model will function successfully if the class-specific loss weighting approach is used to address the issue of category imbalance. As a result of this strategy, the accuracy of samples that are susceptible to misclassification can be greatly increased. To evaluate the model, the HAM10000 dataset was used, and the result showed that the accuracy of the proposed method reached 91.6%, the AUC 97.1%, the sensitivity 73.5%, and the specificity 96.4%. This confirmed that the method can perform well in dermatoscopic classification tasks.

An augmented-intelligence-enabled deep neural networking (AuDNN) system for classifying and predicting skin cancer utilizing multi-dimensional information on industrial IoT standards was proposed by Kumar et al. [15]. The proposed framework incorporates deep learning algorithms and IoT standards to create a robust and efficient skin cancer classification system. The approach was evaluated on a Kaggle skin cancer dataset and CIA datasets on melanoma categorization of skin lesion images, and the results showed that it outperformed other state-of-the-art methods. The proposed AuDNN framework is a significant contribution to the field of medical image analysis. The integration of IoT standards and deep learning algorithms has created a system that is both robust and efficient for skin cancer classification. The paper also provided a detailed analysis of the performance of the proposed method, which can guide the development of future approaches for skin lesion classification. One limitation of this study was that the dataset used for training and evaluation was not explicitly mentioned. It would be helpful to know more about the dataset and its properties to assess the robustness and generalization capabilities of the proposed method. Another limitation is that the implementation of IoT

standards may require significant resources and expertise, which may not be available in all settings. The proposed AuDNN framework achieved an accuracy of 93.26%.

Notwithstanding the amazing developments, the current deep-network-based approaches, which naively adopt the published network topologies in general image classification to the classification of skin lesions, still have much potential for optimization. Using self-attention to describe the global correlation of the features gathered from the conventional deep models, Nakai et al. [39] suggested an enhanced deep bottleneck transformer model to enhance the performance of skin lesions. For balanced learning, they particularly used an improved transformer module that included a dual-position encoding module to include an encoded position vector on both the key and the query vectors. By replacing the bottleneck spatial convolutions of the late-stage blocks in the baseline deep networks with the upgraded module, they created a unique deep skin lesion classification model to enhance skin lesion classification performance. To validate the effectiveness of different deep models in identifying skin lesions, they conducted comprehensive tests on two benchmark skin lesion datasets, ISIC2017 and HAM10000. With their method, the three quantitative metrics of accuracy, sensitivity, and specificity on the ISIC2017 dataset achieved 92.1%, 90.1%, and 91.9%, respectively. The findings on the accuracy and precision for the HAM10000 dataset were 95.84% and 96.1%. This demonstrated a superb harmony between sensitivity and specificity.

Hossain et al. [40] first developed an EM dataset with the assistance of knowledgeable dermatologists from the Clermont-Ferrand University Medical Center in France. Second, the authors trained 23 convolutional neural networks (CNNs) on a collection of skin lesion photos. These CNNs were modified versions of the VGG, ResNet, DenseNet, MobileNet, Xception, NASNet, and EfficientNet architectures. Lastly, the authors used transfer learning from pretrained ImageNet models to improve the CNNs' performance after pretraining them with the HAM10000 skin lesion dataset. Fourth, to examine the explainability of the model, the authors used gradient-weighted class activation mapping to pinpoint the input regions crucial to CNNs for making predictions. Lastly, the authors offered model selection suggestions based on computational complexity and predictive capability. With an accuracy of 84.42% $\pm$ 1.36, an AUC of 0.9189 $\pm$ 0.0115, a precision of 83.1% $\pm$ 2.49, a sensitivity of 87.93% $\pm$ 1.47, and a specificity of 80.65% $\pm$ 3.59, the customized ResNet50 architecture provided the best classification results. With an accuracy of 83.13% $\pm$ 1.2, AUC of 0.9094 $\pm$ 0.0129, precision of 82.83% $\pm$ 1.75, sensitivity of 85.21% $\pm$ 3.91, and specificity of 80.89% $\pm$ 2.95, a lightweight model of a modified EfficientNetB0 also performed well. The authors contributed a Lyme disease dataset with twenty-three modified CNN architecturesfor image-based diagnosis, effective customized transfer learning using the combination of ImageNet and the HAM10000 dataset, a lightweight CNN, and a criteria-based guideline for model architecture selection.

Afza et al. [41] proposed a hierarchical architecture based on two-dimensional superpixels and deep learning to increase the accuracy of skin lesion classification. The authors combined the locally and globally improved photos to improve the contrast of the original dermoscopy images. The proposed method consisted of three steps: superpixel segmentation, feature extraction, and classification using a deep learning model.The proposed method contributes to the field of skin lesion classification by introducing a hierarchical three-step superpixel and deep learning framework. This method improved the accuracy of skin lesion classification and reduced the computational complexity of the task by dividing the image into superpixels and classifying them individually. The proposed method is also generalizable and can be used on other datasets for skin lesion classification. Using an updated grasshopper optimization approach, the collected features were further optimized before being categorized using the Naive Bayes classifier. In order to evaluate the proposed hierarchical technique, three datasets (Ph2, ISBI2016, and HAM1000) consisting of three, two, and seven skin cancer classes were used. For these datasets (Ph2, ISBI2016, and HAM1000), the proposed method had corresponding accuracy levels of 95.40%, 91.1%,

and 85.80%. The findings indicated that this strategy can help in classifying skin cancer more accurately.

Alam [42] proposed S^2C-DeLeNet, a method for detecting skin cancer lesions from dermoscopic images. The proposed method integrates segmentation and classification using a parameter-transfer-based approach. The segmentation network, DeLeNet, was trained on a large-scale dataset for dermoscopic lesion segmentation, and the classification network, S^2CNet, was trained on a public dataset for skin lesion classification. The authors transferred the parameters of the segmentation network to the classification network and fine-tuned the network on the classification task. The architecture of the segmentation sub-network used an EfficientNet B4 backbone in place of the encoder. The classification sub-network contained a "Classification Feature Extraction" component that pulled learned segmentation feature maps towards lesion prediction. The "Feature Coalescing Module" block mixed and trailed each dimensional feature from the encoder and decoder, while the "3D-Layer Residuals" block developed a parallel pathway of low-dimensional features with large variance. These were the blocks created as part of the classification architecture. After tweaking on a publicly accessible dataset, the segmentation achieved a mean Dice score of 0.9494, exceeding existing segmentation algorithms, while the classification achieved a mean accuracy of 0.9103, outperforming well-known and traditional classifiers. Additionally, the network's already-tuned performance produced very pleasant outcomes when cross-inferring on various datasets for skin cancer segmentation. Thorough testing was performed to demonstrate the network's effectiveness for not only dermoscopic pictures, but also for other types of medical imaging, demonstrating its potential to be a systematic diagnostic solution for dermatology and maybe other medical specialties. For comparison, eight cutting-edge networks, AlexNet, GoogLeNet, VGG, ResNet, Inception-Net, EfficientNet, DenseNet, and MobileNet, as well as their different iterations, were taken into account, which confirmed that the proposed approach outperformed the state-of-the-art approaches.

With the aid of cutting-edge deep learning methodology, Elashiri et al. [43] intended to put into practice an efficient way for classifying skin diseases. The contrast-enhancement technique first collects and preprocesses the dataset by histogram equalization. The segmentation of the photos was carried out by the Fuzzy C Means segmentation after preprocessing (FCM). Furthermore, the segmented images were used as the input for ResNet50, VGG16, and Deeplabv3's deep feature extraction. The features were combined and obtained from the third and bottom layer of these three approaches. Hybrid squirrel butterfly search optimization performs weighted feature extraction to offer these concatenated features to the feature trans-creation phase (HSBSO). The modified long short-term memory (MLSTM) receives the changed features, and the same HSBSO optimizes the architecture there to create the final output for classification. The analysis's findings supported the notion that the proposed method is more effective than traditional methods in terms of implementing a classification of skin diseases that is accurate.

Adla et al. [44] proposed a full-resolution convolutional network with hyperparameter optimization for dermoscopy image segmentation-enhanced skin cancer classification. The hyperparameters of the network were optimized through a novel dynamic graph cut algorithm technique. By fusing the wolves' individualized hunting techniques with their collective hunting methods, the hyperparameters highlighted the need for a healthy balance between exploration and exploitation and produced a neighborhood-based searching approach. The motivation of the authors was to create a full-resolution convolutional-network-based model that is hyperparameter-optimized and is capable of accurately identifying different forms of skin cancer using dermoscopy images. The initial contribution made by the authors was FrCN-DGCA, which uses the DGCA approach to segment skin lesion images and generate image ROIs in a manner similar to how doctors define ROIs. The authors' second addition was the action bundle, which is used as a hyperparameter by the skin image-segmentation executor they provided in order to improve the segmentation process's accuracy. This segmentation process was based on the dynamic graph cut. Last, but not least, the authors carried out a quantitative statistical analysis of the skin lesion

segmentation findings to show the dependability of the segmentation methodology and to contrast the findings with those of the current state-of-the-art methods. The suggested model performed better than the other designs in tasks requiring skin lesion identification, with an accuracy of 97.986%.

Hierarchy-aware contrastive learning with late fusion (HAC-LF), a revolutionary technique presented by Hsu and Tseng [45], enhances the performance of multi-class skin classification. A new loss function called hierarchy-aware contrastive loss (HAC Loss) was developed by the developers of HAC-LF to lessen the effects of the major-type misclassification issue. The major-type and multi-class classification performance were balanced using the late fusion method. The ISIC 2019 Challenges dataset, which comprises three skin lesion datasets, was used in a series of tests by the authors to assess the performance of the suggested approach. The experimental results demonstrated that, in all assessment metrics employed in their study, the suggested method outperformed the representative deep learning algorithms for skin lesion categorization. For accuracy, sensitivity, and specificity in the major-type categorization, HAC-LF scored 87.1%, 84.2%, and 88.9%, respectively. Regarding the sensitivity of the minority classes, HAC-LF performed better than the baseline model with an imbalanced class distribution.

A convolutional neural network (CNN) model for skin image segmentation was developed by Yanagisawa et al. [46] in order to produce a collection of skin disease images suitable for the CAD of various skin disease categories. The DeepLabv3+-based CNN segmentation model was trained to identify skin and lesion areas, and the areas that met the criteria of being more than 80% skin and more than 10% lesion of the picture were segmented out. Atopic dermatitis was distinguished from malignant diseases and their consequences, such as mycosis fungoides, impetigo, and herpesvirus infection, by the created CNN-segmented image database with roughly 90% sensitivity and specificity. The accuracy of identifying skin diseases in the CNN-segmented image dataset was higher than that of the original picture dataset and nearly on par with the manually cropped image dataset.

A multi-site cross-organ calibrated deep learning (MuSClD) approach for the automated diagnosis of non-melanoma skin cancer was presented by Zhou et al. [47]. To increase the generalizability of the model, the suggested strategy makes use of deep learning models that have been trained on a variety of datasets from various sites and organs. This paper's key contribution was the creation of a reliable deep-learning-based method for the automated diagnosis of skin cancers other than melanoma. The proposed strategy was intended to go beyond the drawbacks of existing methods, which have poor generalizability because of small sample sizes and a lack of diversity. The MuSClD technique uses datasets from several sites and organs to increase the model's capacity for generalization. The main drawback of this paper was the lack of explanation for how the suggested deep learning model makes decisions. Although the model had a high degree of accuracy in detecting non-melanoma skin cancer, it is unclear how the model came to that conclusion. This lack of interpretability might prevent the suggested strategy from being used in clinical settings. Using a sizable collection of photos of skin cancers other than melanoma, the MuSClD method was assessed. As measured in terms of the AUC, the suggested method fared better than other cutting-edge approaches. Additionally, the study demonstrated that the MuSClD method is adaptable to changes in imaging modalities and patient demographics, making it appropriate for practical use.

Omeroglu et al. [48] proposed a novel soft-attention-based multi-modal deep learning framework for multi-label skin lesion classification. The proposed framework utilizes both visual and textual features of skin lesions to improve the classification accuracy. The framework consisted of two parallel branches, one for processing visual features and the other for processing textual features. A soft attention mechanism was incorporated into the framework to emphasize important visual and textual features. The 7-point criteria evaluation dataset, a well-known multi-modality multi-label dataset for skin diseases, was used to evaluate the proposed framework. For multi-label skin lesion classification, it

attained an average accuracy of 83.04%. It increased the average accuracy on the test set by more than 2.14% and was more accurate than the most-recent approaches.

Serte and Demirel [49] applied wavelet transform to extract features and deep learning to classify the features with the intention to enhance the performance of skin lesion classification. First, the wavelet transform was used as a preprocessing step to extract features from the skin lesion images. Then, skin lesions were divided into various groups using a deep learning model that was trained on the retrieved features. The authors tested their method against other cutting-edge approaches using the publicly accessible dataset ISIC 2017 of skin lesions. The use of a deep learning model for classification and the use of a wavelet transform to extract features were the key contributions of this paper. In this study, the best combination of models for melanoma and seborrheic keratosis detection were the ResNet-18-based I-A1-H-V and ResNet-50-based I-A1-A2-A3 models.

Bansal et al. [50] proposed a grayscale-based lesion segmentation, while texture characteristics were extracted in the RGB color space using global (grey-level co-occurrence matrix (GLCM) for entropy, contrast, correlation, angular second moment, inverse different moment, and sum of squares) and local (LBP and oriented FAST and rotated BRIEF (ORB)) techniques. A total of 52 color attributes for each image were extracted as the color features using histograms of the five color spaces (grayscale, RGB, YCrCb, L*a*b, and HSV), as well as information on the mean, standard deviation, skewness, and kurtosis. The BHHO-S and BHHO-V binary variations of the Harris hawk optimization (HHO) method, which used S-shaped and V-shaped transfer functions with a time-dependent behavior, respectively, for feature selection, were introduced. The classifier that determines whether the dermoscopic image contains melanoma or not was given the selected attributes. The performance of the suggested approaches was compared to that of already-developed metaheuristic algorithms by the authors. The experiment's findings demonstrated that classifiers that used features chosen using BHHO-S were superior to those that used BHHO-V and those that employed current, cutting-edge metaheuristic methods. The experimental results also showed that, in comparison to global- and other local-texture-feature-extraction strategies, texture features derived utilizing local binary patterns and color features offered higher classification accuracy.

Statistical fractal signatures ($_{STF}$) and statistical-prism-based fractal signatures were the two new fractal signatures that Gutiérrez et al. [51] used to solve the issue of amorphous pigmentary lesions and blurred edges (S_{SPF}). In order to classify multiclass skin lesions utilizing the two new fractal signatures and several classifiers, various computer-aided diagnosis techniques were compared. The combination of S_{STF} and the LDA classifier yielded the finest outcomes for reliable, impartial, and reproducible techniques.

Using a hybrid model that integrates deep transfer learning, convolutional neural networks (CNNs), and gradient boosting machines (GBMs), Thanka et al. [52] suggested a new ensemble strategy for the classification of melanoma. The proposed method was examined using 25,331 photos of skin lesions from the ISIC 2019 Challenge, a publicly accessible dataset. According to the experimental findings, the proposed hybrid strategy that merged VGG16 and XGBOOST was successful in achieving an overall accuracy of 99.1%, a sensitivity of 99.4%, and a specificity of 98.8%. The accuracy, sensitivity, and specificity of the proposed hybrid approach, which included VGG16 and LightBGM, were all higher than the figures provided by other models, at 97.2%, 97.8%, and 96.6%, respectively. The preprocessing of the dataset, the kind of CNN model, and the design of the GBM model were all covered in-depth in the authors' extensive explanation of the approach.

In a study by Brinker et al. [53], the diagnostic precision of an artificial intelligence (AI) system for melanoma detection in skin biopsy samples was examined. The performance of the AI algorithm was compared to that of 18 leading pathologists from across the world in the study. The mean sensitivity, specificity, and accuracy of the Ensemble CNNs trained on slides with or without annotation of the tumor region as a region of interest were on par with those of the experts (unannotated: 88%, 88%, and 88%, respectively; area under the curve (AUC) of 0.95; annotated: 94%, 90%, and 92%, respectively; AUC of 0.97). The

research demonstrated that the AI algorithm had a very low rate of false positives and false negatives and was very reliable in detecting melanoma. The study also discovered that the AI algorithm's performance was on par with that of skilled pathologists. The pathologists had a 90.33% diagnosis accuracy, an 88.88% sensitivity, and a 91.77% specificity. There was no statistically significant difference between the AI algorithm and the pathologists. Overall, this research showed that AI algorithms could be a useful tool for melanoma diagnosis, with performance on par with that of skilled pathologists.

In order to classify skin lesions, Alenezi et al. [54] presented a hybrid technique called the wavelet transform-deep residual neural network (WT-DRNNet). The wavelet transformation, pooling, and normalization section of the constructed model employing the suggested approach provided finer details by removing undesired detail from skin lesion images to acquire a better-performing model. The residual neural network built on transfer learning was then used to extract deep features. Finally, the global average pooling approach was combined with these deep features, and the training phase was carried out with the help of the extreme learning machine, which is based on the ReLu and other kinds of activation functions. In order to evaluate the effectiveness of the suggested model, the experimental works employed the ISIC2017 and HAM10000 datasets. The suggested algorithm's accuracy, specificity, precision, and F1-score metrics for performance were 96.91%, 97.68%, 96.43%, and 95.79% for the ISIC2017 dataset, compared to 95.73%, 98.8%, 95.84%, and 93.44% for the HAM10000 dataset. These outcomes performed better than the state-of-the-art for categorizing skin lesions. As a result, the suggested algorithm can help specialized doctors automatically classify cancer based on photographs of skin lesions.

Alhudhaif et al. [55] recommended a deep learning approach that was based on mechanisms for focusing attention and enhanced by methods for balancing data. The dataset used in the study was HAM10000, which included 10,015 annotated skin images of seven different types of skin lesions. The dataset was unbalanced and made balanced using techniques that included SMOTE, ADASYN, RandomOverSampler, and data augmentation. A soft attention module was selected as the attention mechanism in order to focus on the features of the input data and generate a feature map. The proposed model consisted of a soft attention module and convolutional layers. By integrating them with the attention mechanism, the authors were able to extract the image features from the convolutional neural networks. The key areas of the image were the focus of the soft attention module. The soft attention module and the applied data-balancing techniques significantly improved the performance of the proposed model. On open-source datasets for skin lesion classification, numerous studies were performed using convolutional neural networks and attention mechanisms. One of the contributions of the proposed approach was the attention mechanism used in the neural network. The balanced and unbalanced HAM10000 dataset's versions were used for training and the test results at different times. On the unbalanced HAM10000 dataset, training accuracy rates of 85.73%, validation accuracy rates of 70.90%, and test accuracy rates of 69.75% were attained. The SMOTE methods on the balanced dataset yielded accuracy rates of 99.86% during training, 96.41% during validation, and 95.94% during testing. Compared to other balancing methods, the SMOTE method produced better results. It can be seen that the proposed model had high accuracy rates as a result of the applied data-balancing techniques.

Huang et al. [56] proposed a computer-assisted approach for the analysis of skin cancer. In their study, they combined deep learning and metaheuristic methods. The fundamental concept was to create a deep belief network (DBN) based on an enhanced metaheuristic method called the modified electromagnetic field optimization algorithm (MEFOA) to build a reliable skin cancer diagnosis system. The proposed approach was tested on the HAM10000 benchmark dataset, and its effectiveness was verified by contrasting the findings with recent research regarding accuracy, sensitivity, specificity, precision, and F1 score.

Kalpana et al. [57] suggested a technique called ESVMKRF-HEAO, which stands for ensemble support vector kernel random-forest-based hybrid equilibrium Aquila optimization. The HAM10000 dataset, which contains different types of skin lesion images, was

used to test the suggested prediction model. First, preprocessing was applied to the dataset for noise removal and image quality improvement. Then, the malignant lesion patches were separated from the healthy backdrop using the thresholding-based segmentation technique. Finally, the dataset was given to the proposed classifier as the input, and it correctly predicted and categorized the segmented images into five (melanocytic nevus, basal cell carcinoma, melanoma, actinic keratosis, and dermatofibroma) based on their feature characteristics. The proposed model was simulated using the MATLAB 2019a program, and the performance of the suggested ESVMKRF-HEAO method was assessed in terms of parameters such as the sensitivity, F1-score, accuracy, precision, and specificity. In terms of all metrics, the suggested ESVMKRF-HEAO strategy performed better, especially when it came to the experimental data, and a 97.4% prediction accuracy was achieved.

Shi et al. [58] proposed a two-stage end-to-end deep learning framework for pathologic evaluation in skin tumor identification, with a particular focus on neurofibromas (NFs), Bowen disease (BD), and seborrheic keratosis (SK). The most-prevalent illnesses involving skin lesions are NF, BD, and SK, and they can seriously harm a person's body. In their study, the authors suggested two unique methods, the attention graph gated network (AGCN) and chain memory convolutional neural network (CMCNN), for diagnosing skin tumors. Patchwise diagnostics and slidewise diagnostics were the two steps of the framework, where they reported the result of the whole-slide image (WSI) as the input in the proposed diagnosis. Convolutional neural networks (CNNs) were used in the initial screening stage to discover probable tumor locations, and multi-label classification networks were used in the fine-grained classification stage to categorize the detected regions into certain tumor kinds. On a dataset of skin tumor images collected from Huashan Hospital, the suggested framework was tested, and the results showed promising accuracy and receiver operating characteristic curves.

Rafay and Hussain [59] proposed a technique that utilized a dataset that integrated two different datasets to establish a new dataset of 31 diseases of the skin. In their study, the authors used three different CNN models—EfficientNet, ResNet, and VGG—each with a different architecture for transfer learning on the dataset for skin diseases. EfficientNet was further tuned because it had the best testing precision, where it initially achieved a testing accuracy of 71% with a training split of 70%. However, this was considered to be low; thus, the 70% training split for the 3424 samples was increased, and the model's accuracy increased as a result to 72%. Again, the experiment was re-executed with a train–test split of 80%:20%, and the improvement in accuracy was 74%. The new dataset was augmented for a further experiment, which then increased the model's accuracy to 87.15%.

Maqsood and Damaševičius [60] proposed a methodology for localizing and classifying multiclass skin lesions. The suggested method begins by preprocessing the source dermoscopic images with a contrast-enhancement-based modified bio-inspired multiple exposure fusion method. The skin lesion locations were segmented in the second stage using a specially created 26-layer convolutional neural network (CNN) architecture. The segmented lesion images were used to modify and train four pretrained CNN models (Xception, ResNet-50, ResNet-101, and VGG16) in the third stage. In the fourth stage, all of the CNN models' deep feature vectors were recovered and combined using the convolutional sparse image decomposition method. The Poisson distribution feature selection approach and univariate measurement were also employed in the fifth stage to choose the optimal features for classification. A multi-class support vector machine (MC-SVM) was then fed the chosen features to perform the final classification. The proposed method performed better in terms of accuracy, sensitivity, specificity, and F1-score. The addition of multiclass classification increased the research's usefulness in real-world situations. However, the proposed approach lacked interpretability, making it challenging to understand the reasoning behind the classification decisions.

To identify skin diseases, Kalaiyarivu and Nalini [61] developed a CNN-based method that extracted color features and texture (local binary pattern and gray level co-occurrence

matrix) features from hand skin images. In their study, the authors reported the accuracy of the proposed CNN model as 87.5%.

Kousis et al. [62] employed 11 distinct CNN models in a different study to identify skin cancer. In this method, they used the HAM10000 dataset and DenseNet169 model, reporting an accuracy of 92.25%. Among the 11 CNN architecture configurations considered in the study, DenseNet169 reported the best results and achieved an accuracy of 92.25%, a sensitivity of 93.59%, and an F1-score of 93.27%, which outperformed the existing state-of-the-art.

A hybrid classification strategy employing a CNN and a layered BLSTM was proposed by Ahmad et al. [63]. In this study, the classification task was carried out by ensembling the BLSTM with a deep CNN network after feature extraction. The accuracy reported by the authors for their experiments on two different datasets (one customized with a size of 6454 images and the other being HAM10000) was 91.73% and 89.47%, respectively.

A deep-learning-based application that classifies many types of skin diseases was proposed by Aijaz et al. [64]. This method made use of the CNN and LSTM deep learning models. In this study, the experimental analysis was performed on 301 images of psoriasis from the Dermnet dataset and 172 images of normal skin from the BFL NTU dataset. Before extracting the color, texture, and form features, the input sample images underwent image preprocessing comprising data augmentation, enhancement, and segmentation. A convolutional neural network (CNN) and long short-term memory (LSTM) were the two deep learning methods that were used with classification models that were trained on 80% of the images. According to reports, the CNN and LSTM had accuracy rates of 84.2% and 72.3%, respectively. The accuracy results from this study showed that this deep learning technology has the potential to be used in other dermatology fields for better prediction.

Using data from the ISIC 2019 and PH2 databases, Benyahia et al. [65] examined the classification of skin lesions. The efficiency of 24 machine learning methods as classifiers and 17 widely used pretrained convolutional neural network (CNN) architectures as feature extractors were examined by the authors. The authors found accuracy rates of 92.34% and 91.71%, respectively, for a DenseNet201 combined with Fine KNN or Cubic SVM, using the ISIC 2019 dataset. The hybrid approach (DenseNet201 + Cubic SVM and DenseNet201 + Quadratic SVM) was also evaluated on the PH2 dataset, and the results showed that the suggested methodology outperformed the rivals with a 99% accuracy rate.

5.2. Machine Learning and Deep Learning in Skin Disease Detection

Inthiyaz et al. [66] presented a study on the use of deep learning techniques for the detection of skin diseases. The authors proposed a skin-disease-detection model based on convolutional neural networks (CNNs) that can classify skin diseases into ten different categories. The model was trained and evaluated using a dataset from Xiangya-Derm of skin disease images. The results showed that the proposed model achieved high accuracy and outperformed existing state-of-the-art models in skin disease detection. The main contribution of this paper was the development of a novel deep-learning-based skin-disease-detection model that can accurately classify skin diseases into different categories. One potential limitation of this study is that the proposed model was only tested on a specific dataset of skin disease images. Therefore, its generalizability to other datasets or real-world scenarios may need to be further evaluated. The paper reported that the proposed skin-disease-detection model achieved an overall accuracy of 87% on the test set, outperforming other existing models for skin disease detection. The authors also performed a comparative analysis of the proposed model with other state-of-the-art models, including ResNet-50, Inception-v3, and VGG-16. The results showed that the proposed model outperformed these models in terms of accuracy and other evaluation metrics. Overall, the experimental results of this paper suggested that the proposed skin-disease-detection model based on deep learning techniques can accurately classify skin diseases and has the potential to be a useful tool for dermatologists and healthcare professionals in diagnosing skin diseases.

Table 8. Summary of previous studies on skin disease classification with their performance.

Author/Year	Method	Dataset	Dataset Size	Acc (%)	Sn (%)	Sp (%)	R (%)	P (%)	F1 (%)
Ali et al. [30]	EfficientNets	HAM10000	10,015	87.9	88	88	88	88	87
Reddy et al. [36]	GWO	PH2	200	94.2	91.83	96.47	91.83	96.15	93.94
Inthiyaz et al. [66]	CNN	Xiangya-Derm	-	87	-	-	-	-	-
Srinivasu et al. [31]	DLNN + MobileNet V2 + LSTM	HAM10000	10,015	90.21	92.24	95.1	92.24	-	-
Shetty et al. [32]	ML + CNN	HAM10000	10,015	95.18	94	-	85	88	86
Wei et al. [34]	DenseNet + ConvNeXt	Peking-Union Medical College Hospital	2600	96.54	94.75	-	94.74	95.45	95.03
Wei et al. [34]	DenseNet + ConvNeXt	HAM10000	10,015	95.29	92.58	-	92.58	88.35	89.99
Almuayqil et al. [35]	DenseNet 201 + ML	HAM10000	10,015	99.94	91.48	98.82	91.48	97.01	-
Malibari et al. [37]	ODNNsingle bondCADSCC	-	-	99.90	-	-	-	-	-
Qian et al. [38]	DCNN-GMAB	HAM10000	10,015	91.6	73.5	96.4	73.5	-	-
Jain et al. [33]	OP-DNN	ISIC	23,906	95.6	91.2	97	91.2	92	-
Kumar et al. [15]	AUDNN	Kaggle + CIA	-	93.26	-	-	-	-	-
Nakai et al. [39]	EDBTM	HAM10000	10,015	95.84	-	-	-	96.1	-
Nakai et al. [39]	EDBTM	ISIC2017	-	92.1	90.1	91.9	-	-	-
Hossain et al. [40]	Customized ResNet50	EM + HAM10000	-	84.42 ± 1.36	87.93 ± 1.47	80.65 ± 3.9	-	83.1 ± 2.49	-
Hossain et al. [40]	Lightweight EfficientNetB0	EM + HAM10000	-	83.13 ± 1.2	85.21 ± 3.91	80.89 ± 2.95	-	82.83 ± 1.75	-
Afza et al. [41]	Hierarchical: NB	ISBI2016	1279	91.1	91	-	-	91.5	-
Afza et al. [41]	Hierarchical: NB	HAM10000	10,015	85.80	86	-	-	86.28	86.14
Afza et al. [41]	Hierarchical: NB	PH2	200	95.40	95.1	-	-	95.33	95.21

Table 8. *Cont.*

Author/Year	Method	Dataset	Dataset Size	Acc (%)	Sn (%)	Sp (%)	R (%)	P (%)	F1 (%)
Alam et al. [42]	S^2C-DeLeNet	HAM10000	10,015	91.03	90.58	90.58	90.58	90.38	90.48
Elashiri et al. [43]	HSBSO-LSTM	PH2	200	93.5	93.8	93.3	-	90.4	9.2
Elashiri et al. [43]	HSBSO-LSTM	HAM10000	10,015	93.8	93.9	93.8	-	33.9	49.8
Hsu and Tseng [45]	HAC-LF	ISIC2019	-	87.1	84.2	88.9	-	-	-
Omeroglu et al. [48]	Soft-attention-based multi-modal DL	7-point criteria evaluation (SPC)	1011	83.04	72.9	88.03	78.13	-	-
Serte and Demirel [49]	ResNet-18-based I-A1-H-V	ISIC2017	2000	81.5	-	97.5	-	-	-
Serte and Demirel [49]	ResNet-50-based I-A1-A2-A3	ISIC2017	2000	81	-	99.5	-	-	-
Bansal et al. [50]	BHHO-S algorithm + linear SVM	HAM10000	88	89	89	-	86	-	-
Gutiérrez et al. [51]	SSTF statistical fractal signatures + LDA classifier (4 classes)	ISIC2019	25,331	87	63	89	-	65	-
Gutiérrez et al. [51]	SSTF statistical fractal signatures + LDA classifier (7 classes)	ISIC2019	25,331	88	41	92	-	46	-
Thanka et al. [52]	VGG16 + XGBOOST	ISIC	1416	99.1	99.4	98.8	-	-	-
Thanka et al. [52]	VGG16 + LightBGM	ISIC	1416	97.2	97.8	96.6	-	-	-
Brinker et al. [53]	Ensembles: 3-CNNs	-	-	90.33	88.88	91.77	-	-	-
Alenezi et al. [54]	WT-DRNNet (ReLu)	ISIC2017	2750	96.91	-	97.68	-	96.43	95.79
Alenezi et al. [54]	WT-DRNNet (PReLu)	ISIC2017	2750	96.91	97.68	-	96.43	95.79	-
Alenezi et al. [54]	WT-DRNNet (Sigmoid)	ISIC2017	2750	96.91	-	97.68	-	96.43	95.79

Table 8. *Cont.*

Author/Year	Method	Dataset	Dataset Size	Acc (%)	Sn (%)	Sp (%)	R (%)	P (%)	F1 (%)
Alenezi et al. [54]	WT-DRNNet (Hardlim)	ISIC2017	2750	96.91	-	97.68	-	96.43	95.79
Alenezi et al. [54]	WT-DRNNet (ReLu)	HAM10000	10,015	95.73	-	98.80	-	95.84	93.44
Alenezi et al. [54]	WT-DRNNet (PReLu)	HAM10000	10,015	95.36	-	98.62	-	95.59	93.37
Alenezi et al. [54]	WT-DRNNet (Sigmoid)	HAM10000	10,015	93.19	-	98.00	-	93.20	89.82
Alenezi et al. [54]	WT-DRNNet (Hardlim)	HAM10000	10,015	92.14	-	97.61	-	91.82	87.45
Alhudhaif et al. [55]	Soft-attention-based CNN	HAM10000 (unbalanced)	10,015	69.75	-	-	-	-	-
Alhudhaif et al. [55]	Soft-attention-based CNN	HAM10000 (balanced-SMOTE)	46,935	-	-	96	96.14	-	95.86
Alhudhaif et al. [55]	Soft-attention-based CNN	HAM10000 (balanced-ADASYN)	46,999	-	-	94.29	94.71	-	94
Alhudhaif et al. [55]	Soft-attention-based CNN	HAM10000 (balanced-RandomOverSampler)	46,935	-	-	88.57	90.14	-	89.29
Huang et al. [56]	DBN-MEFOA	HAM10000	10,015	97.99	92.99	97.00	-	96.99	91.99
Kalpana et al. [57]	ESVMKRF-HEAO	HAM10000	10,015	97.4	95.9	96	-	96.3	97.4
Shi et al. [58]	CMCNN-whole-slide image (WSI)	-	504	82.68	-	-	-	-	-
Shi et al. [58]	AGCN-whole-slide image (WSI)	-	504	95.24	-	-	-	-	-
Rafay and Hussain [59]	EfficientSkinDis: fine-tuned EfficientNet-B2	Atlas Dermatology and ISIC	4910	74	-	-	-	-	-
Rafay and Hussain [59]	EfficientSkinDis: fine-tuned EfficientNet-B2	Atlas Dermatology and ISIC	45,912	87.15	-	-	-	-	-

Table 8. *Cont.*

Author/Year	Method	Dataset	Dataset Size	Acc (%)	Sn (%)	Sp (%)	R (%)	P (%)	F1 (%)
Kalaiyarivu and Nalini [61]	CNN	Customized hand images	-	87.5	-	-	-	-	-
Kousis et al. [62]	DenseNet169	HAM10000	10,015	92.25	93.59	-	-	-	93.27
Ahmad et al. [63]	CNN-layered BLSTM	Customized	6454	91.73	91.83	98.77	-	-	-
Ahmad et al. [63]	CNN-layered BLSTM	HAM10000	10,015	89.47	88.33	97.17	-	-	-
Aijaz et al. [64]	CNN	Dermnet (301) + BFL NTU (172)	473	84.2	84.33	86	-	-	-
Aijaz et al. [64]	LSTM	Dermnet (301) + BFL NTU (172)	473	72.3	72.33	75.16	-	-	-
Benyahia et al. [65]	DenseNet201 + Cubic SVM	ISIC2019	-	91.71	-	96.4	92.04	84.82	86.82
Benyahia et al. [65]	DenseNet201 + Fine KNN	ISIC2019	-	92.34	-	96.38	92.75	85.22	86.96
Benyahia et al. [65]	DenseNet201 + Cubic SVM	PH2	-	99	-	-	-	-	-
Benyahia et al. [65]	DenseNet201 + Quadratic SVM	PH2	-	99	-	-	-	-	-
Yanagisawa et al. [46]	DeepLabv3+- CNN	NSDD	16,313	90	90	90	-	-	-
Maqsood and Damaševičius [60]	MC-SVM	HAM10000	10,015	98.57	93.89	96.37	-	-	94.98
Maqsood and Damaševičius [60]	MC-SVM	ISIC2018	98.62	93.24	97.98	-	-	95.98	-
Maqsood and Damaševičius [60]	MC-SVM	ISIC2019	93.47	84.34	87.53	-	-	88.67	-
Maqsood and Damaševičius [60]	MC-SVM	PH2	98.98	98.03	98.70	-	-	98.87	-

Dwivedi et al. [67] proposed a deep-learning-based approach for automated skin disease detection using the Fast R-CNN algorithm. The proposed approach aimed to address the limitations of traditional approaches that are heavily dependent on domain knowledge and feature extraction. The experimental findings demonstrated that the suggested method achieved an overall accuracy of 90%, which outperformed traditional machine-learning-based approaches. The approach was evaluated on the HAM10000 dataset, which is a widely used benchmark dataset for skin disease detection. The contribution of the paper was the proposed approach for automated skin disease detection using the Fast R-CNN algorithm, which can handle large datasets and achieve high accuracy without the need for domain knowledge or feature extraction. One of the limitations of the proposed approach is that it requires a large amount of labeled data for training, which can be a challenge for some applications. Additionally, the approach is limited to detecting skin diseases included in the HAM10000 dataset, and further evaluation is required for detecting other skin diseases. Overall, the paper presented a promising approach for automated skin disease detection using deep learning, with the potential to improve clinical diagnosis and reduce human error.

Alam and Jihan [68] presented an efficient approach for detecting skin diseases using deep learning techniques. The proposed approach involves preprocessing of skin images followed by feature extraction using convolutional neural networks (CNNs) and classification using support vector machine (SVM). The dataset used for the evaluation consisted of 10,000 skin images, which were categorized into seven different skin diseases. The approach achieved an accuracy of 95.6%, which is a significant improvement compared to existing approaches. The contribution of the paper is the development of an efficient and accurate skin disease detection approach using deep learning techniques. The main limitation of the proposed approach is that it requires a large amount of data to train the model effectively. In addition, the proposed approach may not be suitable for detecting rare skin diseases that are not present in the training dataset. The proposed approach achieved an accuracy of 95.6%, which outperformed the existing approaches. The precision and recall values for each skin disease category were also reported, which further validated the effectiveness of the proposed approach. The results demonstrated that the proposed approach can accurately detect skin diseases, which can aid in the early diagnosis and treatment of such diseases.

Wan et al. [69] proposed a detection algorithm for pigmented skin diseases, based on classifier-level and feature-level fusion. The proposed algorithm combines the strengths of multiple classifiers and features to improve the detection accuracy of pigmented skin diseases. The experiments showed that the proposed algorithm outperformed the other state-of-the-art algorithms in terms of accuracy and other parameters. The novelty of the algorithm proposed in this paper for the diagnosis of pigmented skin diseases was its main contribution. The efficiency of the suggested fusion network was visualized using gradient-weighted class activation mapping (Grad_CAM) and Grad_CAM++. The results demonstrated that the accuracy and area under the curve (AUC) of the approach in this study reached 92.1% and 95.3%, respectively, when compared to those of the conventional detection algorithm for pigmented skin conditions. The contribution of this study as claimed by the authors included techniques used to perform the data augmentation, the method used for image augmentation noise, the two-feature-level fusion optimization scheme, and the visualization algorithms (Grad_CAM and Grad_CAM++) to verify the validity of the fusion network.

An optimization-based algorithm to identify skin cancer from a collection of photos was presented by Kumar and Vanmathi [70]. The input image was created from a database in the primary stage, where it was preprocessed with a Gaussian filter and region of interest (ROI) extraction to weed out noise and mine interesting sections. Using the proposed U-RP-Net, the segmentation was carried out. By combining U-Net and RP-Net in this instance, the proposed U-RP-Net model was created. Meanwhile, the output from the RP-Net and U-Net models was combined using the Jaccard-similarity-based fusion model.

To enhance the performance of detection, data augmentation was performed. SqueezeNet was used to locate skin cancer at the end. The Aquila whale optimization (AWO) method was also used to train SqueezeNet. The Aquila optimizer (AO) and whale optimization algorithm were combined to create the new AWO method (WOA). The highest testing accuracy of 92.5%, sensitivity of 92.1%, and specificity of 91.7% were achieved by the developed AWO-based SqueezeNet.

Suicmez et al. [71] proposed a hybrid learning approach for the detection of melanoma by removing hair from dermoscopic images. The approach combines image-processing techniques and the wavelet transform with machine learning algorithms, including a support vector machine (SVM) and artificial neural network (ANN). In order to speed up the algorithm's detection time, the system first uses image-processing techniques (masking for saturation and wavelet transform) to eliminate impediments such as hair, air bubbles, and noise from dermoscopic images. Making the lesion more noticeable for detection is another crucial step in this procedure. Melanoma detection was used for the first time using a unique hybrid model that combines deep learning and machine learning as an AI building block. The HAM10000 (ISIC 2018) and ISIC 2020 datasets were utilized to gauge the developed system's performance ratio after stabilization. The paper demonstrated the effectiveness of the proposed approach in removing hair from dermoscopic images, which is a crucial preprocessing step in melanoma detection. However, the approach is dependent on the quality of the input images, and low-quality images may negatively impact the performance.

Choudhary et al. [72] proposed a neural-network-based method to separate dermoscopic images including two different kinds of skin lesions. The initiative's proposed solution was divided into four steps that included initial image processing, skin lesion segmentation, feature extraction, and DNN-based classification. With a median filter, image processing was the initial stage in removing any extra noise. The specific locations of the skin lesions were then segmented using Otsu's image-segmentation method. The third stage involved further extraction of the skin lesion characteristics, which were retrieved utilizing the RGB color model, 2D DWT, and GLCM. The classification of the various types of skin diseases using a backpropagation deep neural network and the Levenberg–Marquardt (LM) generalization approach to reduce the mean-squared error was the fourth stage. The ISIC 2017 dataset was used to train and test the suggested deep learning model. With DNN, they were able to outperform other state-of-the-art machine learning classifiers with an accuracy of 84.45%.

Lembhe et al. [73] proposed a synthetic skin-cancer-screening method using a solution or sequence from visual LR images. To improve the image-processing and machine learning methods, a deep learning strategy on super-resolution images was applied. Convolutional neural network models such as VGG 16, ResNet, and Inception V3 can be accurately recreated using image super-resolution (ISR) techniques. This model was created with the help of the Keras backend, and it was evaluated using a sequence or solution from visual LR photos. To improve the altering layers of the neural networks utilized for training, a deep learning strategy on the picture super-resolution was applied. The convolutional neural network model's ISIC accuracy dataset, which is publicly available, was used to build the model.

A novel hybrid extreme learning machine (ELM) and teaching–learning-based optimization (TLBO) algorithm was developed by Priyadharshini et al. [74] as a flexible method for melanoma detection. While TLBO is an optimization technique used to fine-tune the network's parameters for enhanced performance, the ELM is a single-hidden-layer feed-forward neural network that can be trained rapidly and accurately. In contrast to earlier studies, the authors used the two methodologies to identify skin lesions as benign or malignant images, potentially increasing the accuracy of melanoma identification. However, the performance of the proposed method was only tested on a single dataset for skin cancer detection, which is a drawback of the paper. Evaluating the performance of the algorithm

on additional skin cancer datasets should have been assessed by the authors to establish its practicality and robustness.

For the purpose of detecting melanoma skin cancer, Dandu et al. [75] introduced a unique method that combines transfer learning with hybrid classification. To increase the accuracy of melanoma detection, the authors developed a hybrid framework that uses pretrained deep learning models for segmentation and incorporates a hybrid classification technique. The development of a hybrid strategy that successfully combines transfer learning and classification approaches was one of the paper's contributions. The authors increased melanoma detection accuracy by modifying a pretrained convolutional neural network for skin lesion segmentation and mixing hand-crafted features with segmented lesion features in the classification process. The proposed approach was evaluated in terms of accuracy, precision, and recall on a benchmark dataset. However, the paper did have certain limitations, where clinical validation is needed to evaluate the generalizability and dependability of the suggested strategy across a range of demographics and skin types. The paper might also used more-thorough arguments and justifications for the features used for the hybrid classification technique. Furthermore, the reproducibility and comprehension might be improved by a more-detailed explanation of the specific features used and their significance to melanoma diagnosis. Last, but not least, despite the paper's promise of increased performance in comparison to current procedures, there was a lack of a thorough comparative analysis using cutting-edge techniques. Such an analysis would offer a more-thorough evaluation of the advantages and disadvantages of the suggested strategy in comparison to other pertinent methods.

In this section skin lesion detection using machine learning and deep learning were examined, and in Table 8 presented summary of all the prior studies discussed in this study and their performance also presented in Table 9.

Table 9. Summary of previous studies on skin disease detection with their performance.

Author and Year	Method	Dataset	Dataset Size	Acc (%)	Sn (%)	Sp (%)	R (%)	P (%)	F1 (%)
Dwived et al. [67]	Fast R-CNN			-	-	90	-	-	-
Alam and Jihan et al. [68]	DL+ Image Processing	-	-	85.14	-	-	-	-	-
Wan et al. [69]	Fusion Network	HAM10000	10,015	92.01	-	89.53	-	-	88.94
Kumar and Vanmathi et al. [70]	U-Net + RP-Net	-	-	92.5	92.1	91.7	-	-	-
Suicmez et al. [71]	Hybrid CNN-Gradient Boost Classifier	HAM10000	10,015	99.4	99.4	-	99.4	99.4	99.4
Suicmez et al. [71]	Hybrid CNN-Machine Learning	ISIC 2020	-	100	100	-	100	100	100
Lembhe et al. [73]	VGG16: ISR	ISIC	70.17	69	-	-	68	73	-
Lembhe et al. [73]	ResNet: ISR	ISIC	86.57	87	-	-	87	87	-
Lembhe et al. [73]	Inception V3: ISR	ISIC	91.26	92	-	-	89	92	-
Priyadharshini et al. [74]	ELM- TLBO	Kaggle and DermIS	300	-	-	-	92.45	89.72	91.64
Dandu et al. [75]	Ensemble Classifier	SIIM ISIC	-	86.38	-	-	86.50	86.16	-

6. Discussion

In this section, we delve into a detailed analysis of the key aspects explored in our survey paper related to skin lesion classification and detection. We focused on papers exclusively dedicated to classification tasks and those solely addressing detection challenges. Additionally, we investigated the relationship between skin lesion dataset modalities and the number of papers utilizing them. Furthermore, we examined how the distribution of papers varied concerning their publication years. Lastly, we explored the relationship between the types of datasets used and the number of papers employing them. By examining these critical factors, we aimed to gain a comprehensive understanding of the trends and developments in skin lesion research, shedding light on the prevailing research priorities and areas for potential future exploration.

The findings from our survey, as illustrated in Tables 10–12, revealed the primary research emphases observed in the papers under consideration. A significant portion of the papers focused on classification tasks, indicating the prevalence of studies aimed at categorizing and labeling various entities within the dataset. However, we also noted that a smaller subset of papers placed their emphasis on detection tasks, highlighting the interest in identifying specific objects or occurrences of interest within the data. Moreover, a notable number of papers took a more-comprehensive approach, addressing both classification and detection aspects in their research, reflecting the need for a holistic understanding and analysis of the data. Furthermore, a few papers delved even deeper, incorporating segmentation alongside classification and detection in their investigations. This integration allowed for the precise delineation and localization of specific regions or structures within the dataset, providing more-detailed insights and facilitating advanced analyses.

The variation in research foci across the surveyed papers emphasized the multidimensional nature of the field, where researchers employed various methodologies and techniques to address distinct aspects of a dataset. The diversity of approaches contributes to a richer understanding of the datasets' complexities and enables the development of robust algorithms and models to tackle real-world challenges effectively. As the field continues to advance, these findings offer valuable guidance for researchers seeking to identify potential research gaps and align their studies with the evolving trends and needs of the domain.

Table 10. Summary of previous studies those focused on skin disease classification.

Author	Method	Objective
Ali et al. [30]	EfficientNets	Classification
Reddy et al. [36]	GWO	Classification
Inthiyaz et al. [66]	CNN	Classification
Srinivasu et al. [31]	DLNN + MobileNet V2 + LSTM	Classification
Shetty et al. [32]	ML + CNN	Classification
Wei et al. [34]	DenseNet + ConvNeXt	Classification
Wei et al. [34]	DenseNet + ConvNeXt	Classification
Almuayqil et al. [35]	DenseNet 201 + ML	Classification
Malibari et al. [37]	ODNNsingle bondCADSCC	Classification
Qian et al. [38]	DCNN-GMAB	Classification
Jain et al. [33]	OP-DNN	Classification
Kumar et al. [15]	AUDNN	Classification
Nakai et al. [39]	EDBTM (Dataset: HAM10000)	Classification
Nakai et al. [39]	EDBTM (Dataset: ISIC2017)	Classification
Hossain et al. [40]	Customized ResNet50	Classification
Hossain et al. [40]	Lightweight EfficientNetB0	Classification

Table 10. *Cont.*

Author	Method	Objective
Afza et al. [41]	Hierarchical: NB	Classification
Afza et al. [41]	Hierarchical: NB (Dataset: PH2)	Classification
Afza et al. [41]	Hierarchical: NB (Dataset: HAM10000)	Classification
Alam et al. [42]	S^2C-DeLeNet	Classification
Elashiri et al. [43]	HSBSO-LSTM (Dataset: PH2)	Classification
Elashiri et al. [43]	HSBSO-LSTM (Dataset: HAM10000)	Classification
Benyahia et al. [65]	DenseNet201 + Cubic SVM	Classification
Benyahia et al. [65]	DenseNet201 + Quadratic SVM	Classification

Table 11. Summary of previous studies those focused on skin disease detection.

Author	Method	Objective
Dwived et al.[67]	Fast R-CNN	Detection
Alam and Jihan et al. [68]	DL+ Image Processing	Detection
Wan et al.[69]	Fusion Network	Detection
Kumar and Vanmathi et al. [70]	U-Net + RP-Net	Detection
Suicmez et al. [71]	Hybrid CNN-Gradient Boost Classifier	Detection
Suicmez et al. [71]	Hybrid CNN-Machine Learning	Detection
Lembhe et al. [73]	VGG16: ISR	Detection
Lembhe et al.[73]	ResNet: ISR	Detection
Lembhe et al. [73]	Inception V3: ISR	Detection
Priyadharshini et al. [74]	ELM- TLBO	Detection
Dandu et al. [75]	Ensemble Classifier	Detection

Table 12. Summary of previous studies those focused on skin disease for multiple objectives.

Author	Method	Objective
Reddy et al. [36]	GAWO	Segmentation and Classification
Malibari et al. [37]	ODNNsingle bondCADSCC	Classification and Detection
Afza et al. [41]	Hierarchical: NB	Segmentation, Classification, and Detection
Afza et al. [41]	Hierarchical: NB	Segmentation, Classification, and Detection
Afza et al. [41]	Hierarchical: NB	Segmentation, Classification, and Detection
Alam et al. [42]	S^2C-DeLeNet	Segmentation, Classification, and Detection
Maqsood and Damaševičius [60]	MC-SVM	Classification and Detection
Maqsood and Damaševičius [60]	MC-SVM	Classification and Detection
Maqsood and Damaševičius [60]	MC-SVM	Classification and Detection
Maqsood and Damaševičius [60]	MC-SVM	Classification and Detection
Dandu et al. [75]	Ensemble Classifier	Segmentation, Classification, and Detection
Yanagisawa et al. [46]	DeepLabv3+- CNN	Segmentation and Classification

In this comprehensive survey paper, we performed a thorough collection of research papers published between the years 2017 and 2019. The content extracted from these papers primarily focused on their Section 1, making up approximately 11.11% of the total papers included in our analysis. By delving into these introductory sections, we aimed to gain insights into the prevalent themes, background knowledge, and contextual information used by researchers in their respective studies.

Notably, the majority of the papers we examined were relatively recent, with a substantial portion published in the year 2020, constituting around 6.18% of the papers in our survey. This suggests a growing interest and significant advancements in the research field during that particular year. The influx of publications in 2020 indicates an active and dynamic research landscape, with scholars contributing new perspectives and findings to the body of knowledge.

Moreover, we observed a substantial increase in publications in the subsequent years, with 2021 contributing to 11.11% of the papers. This steady growth indicates a sustained momentum in research activities, as researchers continued to investigate and explore various topics and areas of interest.

The year 2022 saw a remarkable surge in scholarly output, covering an impressive 38.27% of the papers in our survey. This surge may reflect emerging trends, breakthroughs, or significant developments in the field, garnering substantial attention from researchers and leading to a spike in academic contributions.

Even though the year 2023 was still ongoing at the time of our survey, it already showcased a notable presence, accounting for 33.33% of the papers. This suggests that research endeavors were thriving, and the year holds promise for numerous new discoveries and advancements.

By carefully analyzing the distribution of publications across these years, our survey paper provides a snapshot of the research landscape's temporal evolution (Figure 2). The higher concentration of recent papers highlights the dynamic nature of the field and the continuous drive to explore new avenues and challenges. Moreover, it points to the significance of staying up-to-date with the latest research findings and integrating the most-current knowledge into ongoing studies.

Furthermore, our survey contributes to understanding the trends and areas of focus within the research community over time. The increasing trend in publications from 2020 to 2023 indicates that the topics being studied were of great interest to researchers, likely due to their relevance and potential impact on the broader scientific and practical domains.

In our systematic review paper, we conducted an in-depth analysis of a diverse range of skin lesion datasets, specifically focusing on the imaging modalities employed to capture the characteristics of these lesions. Our investigation yielded valuable insights into the distribution and prevalence of different imaging modalities within these datasets, as presented in Figure 3.

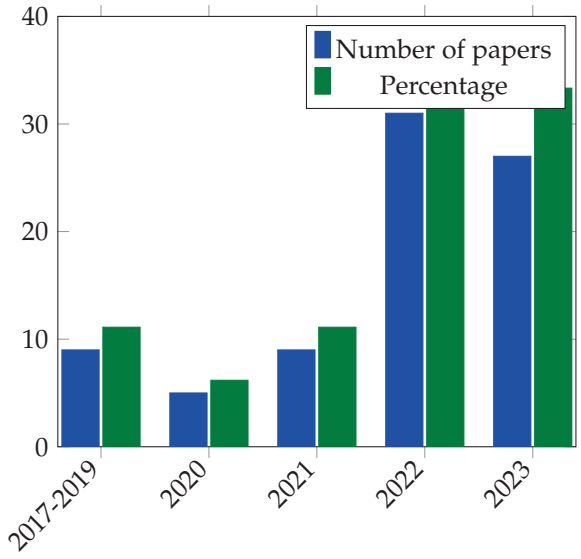


Figure 2. Reviewed papers' distribution by year of publication.

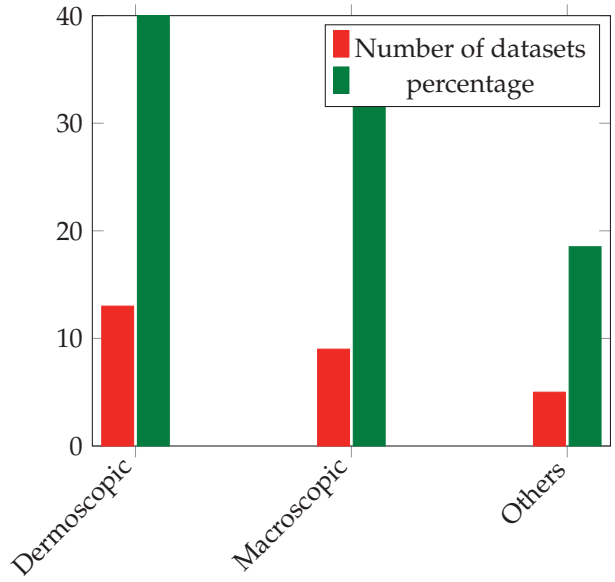


Figure 3. Imaging modalities in skin datasets.

A significant portion, accounting for 48.12% of the datasets, utilized the dermoscopic image modality. Dermoscopy, a non-invasive imaging technique, plays a crucial role in dermatology and skin lesion research. It involves the use of a specialized dermatoscope, which is a handheld magnifying device with a light source, to examine the skin lesions at a higher level of magnification. Dermoscopic images provide clinicians and researchers with enhanced visualization of the morphological structures and patterns within the skin lesions, aiding in more-accurate diagnosis, classification, and monitoring of various skin conditions. The prominence of dermoscopic imaging in nearly half of the datasets underscores its importance as a preferred and highly informative modality in the field.

Another significant imaging modality, observed in approximately 33.33% of the datasets, was the macroscopic imaging modality. Macroscopic images are captured using conventional visible light photography, which allows for a comprehensive view of the skin lesions as perceived by the naked eye. While macroscopic images lack the fine details provided by dermoscopy, they offer a practical and easily accessible means of documenting skin lesions. These images are particularly useful in a clinical setting where dermoscopes might not be readily available, and they provide essential information about the external appearance and overall presentation of the skin lesions. Moreover, macroscopic images often serve as valuable complements to dermoscopic images, providing a broader context for the lesion's evaluation.

In addition to dermoscopic and macroscopic imaging modalities, the remaining 18.52% of the datasets encompassed various other types of imaging modalities. These may include confocal microscopy, ultrasound imaging, multispectral imaging, or combinations of multiple imaging techniques. Each of these alternative modalities offers unique benefits and insights into specific aspects of skin lesions, enabling a comprehensive understanding of their underlying structures and pathological features. The inclusion of these diverse imaging modalities in a portion of the datasets indicates the continuous exploration and experimentation within the scientific community to advance the capabilities of skin lesion analysis and diagnosis.

In our systematic review paper, we conducted an extensive analysis of various research studies in the field of skin lesion detection, classification, and segmentation. As illustrated in Figure 4, we observed the utilization of different datasets in these studies. Notably, the HAM10000 dataset was employed in 38.02% of the papers, indicating its widespread adoption among researchers. The PH2 dataset, on the other hand, was found to be used in 8.50% of the papers. Although its usage was less prevalent compared to HAM10000, it still

played a significant role in contributing to the body of knowledge in this area. Furthermore, we observed that the ISIC dataset was utilized in 33.8% of the research papers. The high usage of the ISIC dataset can be attributed to its large and diverse collection of skin lesion images, making it a valuable resource for developing and evaluating skin-lesion-detection and -classification algorithms. In addition to the three major datasets mentioned above, we discovered that the remaining datasets collectively covered 19.72% of the studies. These datasets might be more-specialized or domain-specific, serving specific research purposes, or comparatively smaller in size. Overall, the data from our systematic review indicated that the HAM10000, ISIC, and PH2 datasets were the most-commonly used and influential resources in the domain of skin lesion research. Researchers have heavily relied on these datasets to train, test, and benchmark their algorithms due to their richness, diversity, and representativeness of real-world skin lesions. By understanding the prevalence and usage of these datasets, we gain valuable insights into the current trends and directions in skin lesion research, allowing for better benchmarking and comparison of novel approaches in the field. It also highlights the need for continued efforts in curating and sharing high-quality datasets to further advance the state-of-the-art in skin lesion detection, classification, and segmentation.

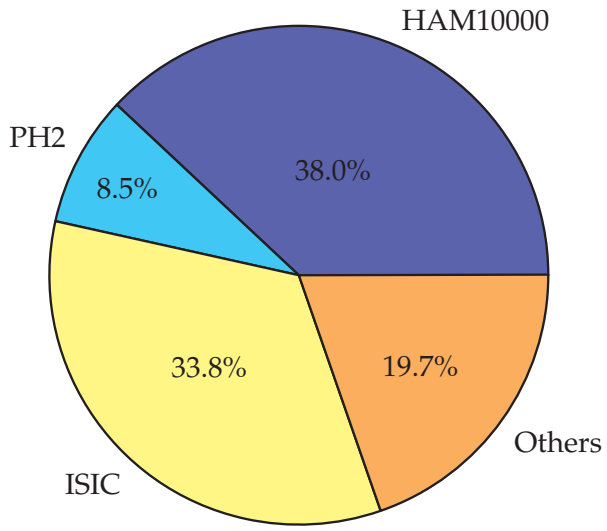


Figure 4. Datasets' distribution in the systematically reviewed papers.

7. Open Challenges for Skin Lesion Classification and Detection

Skin disease diagnosis is an area of ongoing research and development, with several open challenges that researchers and clinicians are actively addressing. Here are some of the key challenges in skin disease diagnosis, along with possible citations for further reading:

- image analysis:It is still difficult to develop reliable automated image analysis methods for the detection of skin diseases. This entails the recognition, categorization, and segmentation of skin lesions from dermatoscopic or image-based data [76,77].
- Data standardization and annotation: The lack of standardized and annotated datasets for skin diseases hinders the development and evaluation of algorithms. Creating comprehensive datasets with accurate annotations is crucial for training and validating machine learning models [78–80].
- Interpretable decision support systems: Skin disease diagnosis often requires interpretability to gain trust from clinicians. Developing decision support systems that provide transparent explanations for their predictions is a challenge that needs to be addressed [81].
- Incorporating clinical data: Integrating patient history, symptoms, and other clinical data along with visual information can improve the accuracy of skin disease diag-

nosis. However, effectively utilizing heterogeneous clinical data remains an open challenge [76].

- Real-time diagnosis: Enabling real-time skin disease diagnosis in clinical settings is another challenge. Developing fast and efficient algorithms that can provide quick and accurate assessments is crucial for improving patient outcomes [82].
- Addressing bias in dermatological datasets: Many dermatological datasets suffer from biases, including under-representation of certain skin types and diseases. Overcoming these biases is essential to ensure fairness and accuracy in skin-disease-diagnosis algorithms [83].
- Augmenting small and imbalanced datasets: Obtaining large and balanced datasets for training skin-disease-diagnosis models can be challenging. Developing effective data-augmentation techniques and strategies to handle imbalanced classes is crucial for improving model performance [84].
- Explainability and interpretability: Interpreting the decisions made by skin-disease-diagnosis models is important for gaining trust and acceptance from healthcare professionals. Developing explainable and interpretable models that can provide insights into the decision-making process is an ongoing challenge [85,86].
- Generalization to external data: Ensuring the generalizability of skin-disease-diagnosis models to external datasets and real-world clinical settings is crucial. Models need to be robust enough to handle variations in imaging conditions, patient demographics, and disease presentations [87].
- Integration with clinical workflows: Seamlessly integrating skin disease diagnosis algorithms into clinical workflows poses a challenge. The development of user-friendly interfaces and systems that can assist healthcare professionals in real-time diagnosis is essential for practical implementation [88,89].
- Ethical issues associated with AI: Currently, all doctors and users of AI products face the ethical challenges brought on by this technology. As most of us know, artificial intelligence may greatly aid in diagnosing and classifying diseases such as dermatological and other conditions. However, it also contributed to the current methods of skin-related disease detection and treatment, which indeed raise severe ethical and dermatological questions. As a result, the AI research community has been inspired to concentrate on trustworthy and responsible AI research [77].
- Skin condition similarities: One of the most-common challenges in skin disease/cancer classification and detection is that many skin conditions have similarities between them that are not distinguishable visually [11].

These challenges highlight the ongoing research and development efforts in skin disease diagnosis, focusing on data biases, interpretability, generalization, practical integration into clinical settings, etc. Researchers continue to work towards addressing these challenges to improve the accuracy and usability of diagnostic tools for dermatological conditions.

8. Conclusions

This survey on classification, segmentation, and detection of skin diseases and skin cancer has brought to light the impressive developments in dermatology made possible by the application of artificial intelligence (AI) techniques. This survey paper demonstrated the potential of AI-based systems to increase diagnostic precision, boost patient outcomes, and completely transform the identification and management of skin disorders including cancer.

Traditional machine learning algorithms, deep learning algorithms, and image analysis methods have all been used by researchers to create complex models that can analyze dermatological images captured using different imaging modalities with high levels of accuracy, sensitivity, specificity, and F1-scores. These simulations have demonstrated their capacity to categorize different skin conditions and locate malignant tumors, matching and occasionally even outperforming the performance of professional dermatologists.

The application of AI to dermatology has enhanced patient care by creating new opportunities for more-precise diagnosis. In situations where access to dermatologists may be limited, the use of computer-aided diagnostic (CAD) systems has the potential to help healthcare practitioners make decisions. These solutions can help with triage, offer second views, and increase the effectiveness of clinical workflows, all of which will ultimately enhance patient care and results.

However, despite the enormous progress made, difficulties still exist in the creation and application of AI-based systems for the diagnosis of skin conditions and skin cancer. To ensure trustworthy and moral applications in clinical settings, concerns including the quality and diversity of training datasets, class imbalance, and the interpretability of AI models must be addressed. Additionally, careful consideration of data protection, regulatory compliance, and physician acceptability is necessary for the integration of these technologies into the current healthcare infrastructure. Future studies should concentrate on overcoming these difficulties and enhancing the precision and durability of AI-based skin disease classification, segmentation, and detection systems. The creation of explainable AI models should also be prioritized since they can promote transparent decision-making and foster a relationship of trust between healthcare professionals and AI systems.

In conclusion, the systematic review report has shown how the field of dermatology could be profoundly affected by AI technologies. We can anticipate additional developments in skin illness and skin cancer analysis with continuing study, development, and collaboration between AI experts and dermatologists. These developments promise to increase diagnostic precision, create tailored treatment regimens, and improve patient care, all of which will improve the management of dermatological disorders.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Debelee, T.G.; Schwenker, F.; Ibenthal, A.; Yohannes, D. Survey of deep learning in breast cancer image analysis. *Evol. Syst.* **2020**, *11*, 143–163. [CrossRef]
2. Debelee, T.G.; Schwenker, F.; Rahimeto, S.; Yohannes, D. Evaluation of modified adaptive k-means segmentation algorithm. *Comput. Vis. Media* **2019**, *5*, 347–361. [CrossRef]
3. Rufo, D.D.; Debelee, T.G.; Ibenthal, A.; Negera, W.G. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics* **2021**, *11*, 1714. [CrossRef] [PubMed]
4. Rufo, D.D.; Debelee, T.G.; Negera, W.G. A Hybrid Machine Learning Model Based on Global and Local Learner Algorithms for Diabetes Mellitus Prediction. *J. Biomimetics Biomater. Biomed. Eng.* **2022**, *54*, 65–88. [CrossRef]
5. Biratu, E.S.; Schwenker, F.; Ayano, Y.M.; Debelee, T.G. A Survey of Brain Tumor Segmentation and Classification Algorithms. *J. Imaging* **2021**, *7*, 179. [CrossRef]
6. Biratu, E.S.; Schwenker, F.; Debelee, T.G.; Kebede, S.R.; Negera, W.G.; Molla, H.T. Enhanced Region Growing for Brain Tumor MR Image Segmentation. *J. Imaging* **2021**, *7*, 22. [CrossRef]
7. Debelee, T.G.; Kebede, S.R.; Schwenker, F.; Shewarega, Z.M. Deep Learning in Selected Cancers' Image Analysis—A Survey. *J. Imaging* **2020**, *6*, 121. [CrossRef]
8. Afework, Y.K.; Debelee, T.G. Detection of Bacterial Wilt on Enset Crop Using Deep Learning Approach. *Int. J. Eng. Res. Afr.* **2020**, *51*, 131–146. [CrossRef]
9. Waldamichael, F.G.; Debelee, T.G.; Schwenker, F.; Ayano, Y.M.; Kebede, S.R. Machine Learning in Cereal Crops Disease Detection: A Review. *Algorithms* **2022**, *15*, 75. [CrossRef]
10. Wube, H.D.; Esubalew, S.Z.; Weldesellasie, F.F.; Debelee, T.G. Text-Based Chatbot in Financial Sector: A Systematic Literature Review. *Data Sci. Financ. Econ.* **2022**, *2*, 232–259. [CrossRef]
11. Sadik, R.; Majumder, A.; Biswas, A.A.; Ahammad, B.; Rahman, M.M. An in-depth analysis of Convolutional Neural Network architectures with transfer learning for skin disease diagnosis. *Healthc. Anal.* **2023**, *3*, 100143. [CrossRef]
12. Lawton, S. Skin 1: The structure and functions of the skin. *Clin. Pract. Syst. Life Skin* **2019**, *115*, 1–2.

13. Ukharov, A.; Shlivko, I.; Klemenova, I.; Garanina, O.; Uskova, K.; Mironycheva, A.; Stepanova, Y. Skin cancer risk self-assessment using AI as a mass screening tool. *Inform. Med. Unlocked* **2023**, *38*, 101223. [CrossRef]
14. Harvey, N.T.; Chan, J.; Wood, B.A. Skin biopsy in the diagnosis of neoplastic skin disease. *Aust. Fam. Physician* **2017**, *46*, 289–294. [PubMed]
15. Kumar K, A.; Satheesha, T.Y.; Salvador, B.B.L.; Mithileysh, S.; Ahmed, S.T. Augmented Intelligence enabled Deep Neural Networking (AuDNN) framework for skin cancer classification and prediction using multi-dimensional datasets on industrial IoT standards. *Microprocess. Microsyst.* **2023**, *97*, 104755. [CrossRef]
16. Grignaffini, F.; Barbuto, F.; Piazzo, L.; Troiano, M.; Simeoni, P.; Mangini, F.; Pellacani, G.; Cantisani, C.; Frezza, F. Machine Learning Approaches for Skin Cancer Classification from Dermoscopic Images: A Systematic Review. *Algorithms* **2022**, *15*, 438. [CrossRef]
17. Goyal, M.; Knackstedt, T.; Yan, S.; Hassanpour, S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Comput. Biol. Med.* **2020**, *127*, 104065. [CrossRef]
18. Rajakani, K.; Hasan, M.R.; Fatemi, M.I.; Monirujjaman Khan, M.; Kaur, M.; Zaguia, A. Comparative Analysis of Skin Cancer (Benign vs. Malignant) Detection Using Convolutional Neural Networks. *J. Healthc. Eng.* **2021**, *2021*, 5895156.
19. Zafar, M.; Sharif, M.I.; Sharif, M.I.; Kadry, S.; Bukhari, S.A.C.; Rauf, H.T. Skin Lesion Analysis and Cancer Detection Based on Machine/Deep Learning Techniques: A Comprehensive Survey. *Life* **2023**, *13*, 146. [CrossRef]
20. Hauser, K.; Kurz, A.; Haggenmüller, S.; Maron, R.C.; von Kalle, C.; Utikal, J.S.; Meier, F.; Hobelsberger, S.; Gellrich, F.F.; Sergon, M.; et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. *Eur. J. Cancer* **2022**, *167*, 54–69. [CrossRef]
21. Jeong, H.K.; Park, C.; Henao, R.; Kheterpal, M. Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations. *JID Innov.* **2023**, *3*, 100150. [CrossRef] [PubMed]
22. Hasan, M.K.; Ahamad, M.A.; Yap, C.H.; Yang, G. A survey, review, and future trends of skin lesion segmentation and classification. *Comput. Biol. Med.* **2023**, *155*, 106624. [CrossRef] [PubMed]
23. Bhatt, H.; Shah, V.; Shah, K.; Shah, R.; Shah, M. State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: A comprehensive review. *Intell. Med.* **2022**, *3*, 180–190. [CrossRef]
24. Mohammed, S.S.; Al-Tuwaijari, J.M. Skin Disease Classification System Based on Machine Learning Technique: A Survey. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1076*, 012045. [CrossRef]
25. Mazhar, T.; Haq, I.; Ditta, A.; Mohsan, S.A.H.; Rehman, F.; Zafar, I.; Gansau, J.A.; Goh, L.P.W. The Role of Machine Learning and Deep Learning Approaches for the Detection of Skin Cancer. *Healthcare* **2023**, *11*, 415. [CrossRef]
26. Kawahara, J.; Hamarneh, G. Visual Diagnosis of Dermatological Disorders: Human and Machine Performance. *arXiv* **2019**, arXiv:1906.01256.
27. Hamarneh, G. Dataset for Skin Image Analysis. 2023. Available online: https://www.medicalimageanalysis.com/data/skinia (accessed on 5 June 2023).
28. Wen, D.; Khan, S.M.; Ji Xu, A.; Ibrahim, H.; Smith, L.; Caballero, J.; Zepeda, L.; de Blas Perez, C.; Denniston, A.K.; Liu, X.; et al. Characteristics of publicly available skin cancer image datasets: A systematic review. *Lancet Digit. Health* **2022**, *4*, e64–e74. [CrossRef]
29. Balaji, V.; Suganthi, S.; Rajadevi, R.; Kumar, V.K.; Balaji, B.S.; Pandiyan, S. Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier. *Measurement* **2020**, *163*, 107922. [CrossRef]
30. Ali, K.; Shaikh, Z.A.; Khan, A.A.; Laghari, A.A. Multiclass skin cancer classification using EfficientNets—A first step towards preventing skin cancer. *Neurosci. Inform.* **2022**, *2*, 100034. [CrossRef]
31. Srinivasu, P.N.; SivaSai, J.G.; Ijaz, M.F.; Bhoi, A.K.; Kim, W.; Kang, J.J. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. *Sensors* **2021**, *21*, 2852. [CrossRef]
32. Shetty, B.; Fernandes, R.; Rodrigues, A.P.; Chengoden, R.; Bhattacharya, S.; Lakshmanna, K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Sci. Rep.* **2022**, *12*, 18134. [CrossRef] [PubMed]
33. Jain, A.; Rao, A.C.S.; Jain, P.; Abraham, A. Multi-type skin disease classification using OP-DNN based feature extraction approach. *Multimed. Tools Appl.* **2022**, *81*, 6451–6476. [CrossRef] [PubMed]
34. Wei, M.; Wu, Q.; Ji, H.; Wang, J.; Lyu, T.; Liu, J.; Zhao, L. A Skin Disease Classification Model Based on DenseNet and ConvNeXt Fusion. *Electronics* **2023**, *12*, 438. [CrossRef]
35. Almuayqil, S.N.; Abd El-Ghany, S.; Elmogy, M. Computer-Aided Diagnosis for Early Signs of Skin Diseases Using Multi Types Feature Fusion Based on a Hybrid Deep Learning Model. *Electronics* **2022**, *11*, 4009. [CrossRef]
36. Reddy, D.A.; Roy, S.; Kumar, S.; Tripathi, R. A Scheme for Effective Skin Disease Detection using Optimized Region Growing Segmentation and Autoencoder based Classification. *Procedia Comput. Sci.* **2023**, *218*, 274–282. [CrossRef]
37. Malibari, A.A.; Alzahrani, J.S.; Eltahir, M.M.; Malik, V.; Obayya, M.; Duhayyim, M.A.; Lira Neto, A.V.; de Albuquerque, V.H.C. Optimal deep neural network-driven computer aided diagnosis model for skin cancer. *Comput. Electr. Eng.* **2022**, *103*, 108318. [CrossRef]
38. Qian, S.; Ren, K.; Zhang, W.; Ning, H. Skin lesion classification using CNNs with grouping of multi-scale attention and class-specific loss weighting. *Comput. Methods Programs Biomed.* **2022**, *226*, 107166. [CrossRef]
39. Nakai, K.; Chen, Y.W.; Han, X.H. Enhanced deep bottleneck transformer model for skin lesion classification. *Biomed. Signal Process. Control* **2022**, *78*, 103997. [CrossRef]

40. Hossain, S.I.; de Goër de Herve, J.; Hassan, M.S.; Martineau, D.; Petrosyan, E.; Corbin, V.; Beytout, J.; Lebert, I.; Durand, J.; Carravieri, I.; et al. Exploring convolutional neural networks with transfer learning for diagnosing Lyme disease from skin lesion images. *Comput. Methods Programs Biomed.* **2022**, *215*, 106624. [CrossRef]
41. Afza, F.; Sharif, M.; Mittal, M.; Khan, M.A.; Jude Hemanth, D. A hierarchical three-step superpixels and deep learning framework for skin lesion classification. *Methods* **2022**, *202*, 88–102. [CrossRef]
42. Alam.; Mohammad, M.S.; Hossain, M.A.F.; Showmik, I.A.; Raihan, M.S.; Ahmed, S.; Mahmud, T.I. S2C-DeLeNet: A parameter transfer based segmentation-classification integration for detecting skin cancer lesions from dermoscopic images. *Comput. Biol. Med.* **2022**, *150*, 106148. [CrossRef] [PubMed]
43. Elashiri, M.A.; Rajesh, A.; Nath Pandey, S.; Kumar Shukla, S.; Urooj, S.; Lay-Ekuakille, A. Ensemble of weighted deep concatenated features for the skin disease classification model using modified long short term memory. *Biomed. Signal Process. Control* **2022**, *76*, 103729. [CrossRef]
44. Adla, D.; Reddy, G.V.R.; Nayak, P.; Karuna, G. A full-resolution convolutional network with a dynamic graph cut algorithm for skin cancer classification and detection. *Healthc. Anal.* **2023**, *3*, 100154. [CrossRef]
45. Hsu, B.W.Y.; Tseng, V.S. Hierarchy-aware contrastive learning with late fusion for skin lesion classification. *Comput. Methods Programs Biomed.* **2022**, *216*, 106666. [CrossRef] [PubMed]
46. Yanagisawa, Y.; Shido, K.; Kojima, K.; Yamasaki, K. Convolutional neural network-based skin image segmentation model to improve classification of skin diseases in conventional and non-standardized picture images. *J. Dermatol. Sci.* **2023**, *109*, 30–36. [CrossRef]
47. Zhou, Y.; Koyuncu, C.; Lu, C.; Grobholz, R.; Katz, I.; Madabhushi, A.; Janowczyk, A. Multi-site cross-organ calibrated deep learning (MuSClD): Automated diagnosis of non-melanoma skin cancer. *Med. Image Anal.* **2023**, *84*, 102702. [CrossRef]
48. Omeroglu, A.N.; Mohammed, H.M.; Oral, E.A.; Aydin, S. A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105897. [CrossRef]
49. Sertan Serte, H.D. Wavelet-based deep learning for skin lesion classification. *IET Image Process.* **2020**, *14*, 720–726. [CrossRef]
50. Bansal, P.; Vanjani, A.; Mehta, A.; Kavitha, J.C.; Kumar, S. Improving the classification accuracy of melanoma detection by performing feature selection using binary Harris hawks optimization algorithm. *Soft Comput.* **2022**, *26*, 8163–8181. [CrossRef]
51. Camacho-Gutiérrez, J.A.; Solorza-Calderón, S.; Álvarez Borrego, J. Multi-class skin lesion classification using prism- and segmentation-based fractal signatures. *Expert Syst. Appl.* **2022**, *197*, 116671. [CrossRef]
52. Roshni Thanka, M.; Bijolin Edwin, E.; Ebenezer, V.; Martin Sagayam, K.; Jayakeshav Reddy, B.; Günerhan, H.; Emadifar, H. A hybrid approach for melanoma classification using ensemble machine learning techniques with deep transfer learning. *Comput. Methods Programs Biomed. Update* **2023**, *3*, 100103. [CrossRef]
53. Brinker, T.J.; Schmitt, M.; Krieghoff-Henning, E.I.; Barnhill, R.; Beltraminelli, H.; Braun, S.A.; Carr, R.; Fernandez-Figueras, M.T.; Ferrara, G.; Fraitag, S. Diagnostic performance of artificial intelligence for histologic melanoma recognition compared to 18 international expert pathologists. *J. Am. Acad. Dermatol.* **2022**, *86*, 640–642. [CrossRef]
54. Alenezi, F.; Armghan, A.; Polat, K. Wavelet transform based deep residual neural network and ReLU based Extreme Learning Machine for skin lesion classification. *Expert Syst. Appl.* **2023**, *213*, 119064. [CrossRef]
55. Alhudhaif, A.; Almaslukh, B.; Aseeri, A.O.; Guler, O.; Polat, K. A novel nonlinear automated multi-class skin lesion detection system using soft-attention based convolutional neural networks. *Chaos Solitons Fractals* **2023**, *170*, 113409. [CrossRef]
56. Huang, Q.; Ding, H.; Rashid Sheykhahmad, F. A skin cancer diagnosis system for dermoscopy images according to deep training and metaheuristics. *Biomed. Signal Process. Control* **2023**, *83*, 104705. [CrossRef]
57. Kalpana, B.; Reshmy, A.; Senthil Pandi, S.; Dhanasekaran, S. OESV-KRF: Optimal ensemble support vector kernel random forest based early detection and classification of skin diseases. *Biomed. Signal Process. Control* **2023**, *85*, 104779. [CrossRef]
58. Shi, Z.; Zhu, J.; Yu, L.; Li, X.; Li, J.; Chen, H.; Chen, L. A Two-Stage End-to-End Deep Learning Framework for Pathologic Examination in Skin Tumor Diagnosis. *Am. J. Pathol.* **2023**, *193*, 769–777. [CrossRef]
59. Rafay, A.; Hussain, W. EfficientSkinDis: An EfficientNet-based classification model for a large manually curated dataset of 31 skin diseases. *Biomed. Signal Process. Control* **2023**, *85*, 104869. [CrossRef]
60. Maqsood, S.; Damaševičius, R. Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. *Neural Netw.* **2023**, *160*, 238–258. [CrossRef]
61. M, K.; Nalini, N.J. Hand Image Based Skin Disease Identification Using Machine Learning and Deep Learning Algorithms. *ECS Trans.* **2022**, *107*, 17381. [CrossRef]
62. Kousis, I.; Perikos, I.; Hatzilygeroudis, I.; Virvou, M. Deep Learning Methods for Accurate Skin Cancer Recognition and Mobile Application. *Electronics* **2022**, *11*, 1294. [CrossRef]
63. Ahmad, B.; Usama, M.; Ahmad, T.; Khatoon, S.; Alam, C.M. An ensemble model of convolution and recurrent neural network for skin disease classification. *Int. J. Imaging Syst. Technol.* **2022**, *32*, 218–229. [CrossRef]
64. Hasikin, K.; Aijaz, S.F.; Khan, S.J.; Azim, F.; Shakeel, C.S.; Hassan, U. Deep Learning Application for Effective Classification of Different Types of Psoriasis. *J. Healthc. Eng.* **2022**, *2022*, 7541583. [CrossRef]
65. Benyahia, S.; Meftah, B.; Lézoray, O. Multi-features extraction based on deep learning for skin lesion classification. *Tissue Cell* **2022**, *74*, 101701. [CrossRef] [PubMed]
66. Inthiyaz, S.; Altahan, B.R.; Ahammad, S.H.; Rajesh, V.; Kalngi, R.R.; Smirani, L.K.; Hossain, M.A.; Rashed, A.N.Z. Skin disease detection using deep learning. *Adv. Eng. Softw.* **2023**, *175*, 103361. [CrossRef]

67. Dwivedi, P.; Khan, A.A.; Gawade, A.; Deolekar, S. A deep learning based approach for automated skin disease detection using Fast R-CNN. In Proceedings of the 2021 Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 26–28 November 2021; Volume 6, pp. 116–120. [CrossRef]
68. Alam, J. An Efficient Approach for Skin Disease Detection using Deep Learning. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 8–10 December 2021; pp. 1–8. [CrossRef]
69. Wan, L.; Ai, Z.; Chen, J.; Jiang, Q.; Chen, H.; Li, Q.; Lu, Y.; Chen, L. Detection algorithm for pigmented skin disease based on classifier-level and feature-level fusion. *Front. Public Health* **2022**, *10*, 1034772. [CrossRef] [PubMed]
70. Anup Kumar, K.; Vanmathi, C. Optimization driven model and segmentation network for skin cancer detection. *Comput. Electr. Eng.* **2022**, *103*, 108359. [CrossRef]
71. Çağrı, S.; Tolga Kahraman, H.; Suiçmez, A.; Yılmaz, C.; Balcı, F. Detection of melanoma with hybrid learning method by removing hair from dermoscopic images using image processing techniques and wavelet transform. *Biomed. Signal Process. Control* **2023**, *84*, 104729. [CrossRef]
72. Choudhary, P.; Singhai, J.; Yadav, J. Skin lesion detection based on deep neural networks. *Chemom. Intell. Lab. Syst.* **2022**, *230*, 104659. [CrossRef]
73. Lembhe, A.; Motarwar, P.; Patil, R.; Elias, S. Enhancement in Skin Cancer Detection using Image Super Resolution and Convolutional Neural Network. *Procedia Comput. Sci.* **2023**, *218*, 164–173. [CrossRef]
74. Priyadharshini, N.; Selvanathan, N.; Hemalatha, B.; Sureshkumar, C. A novel hybrid Extreme Learning Machine and Teaching–Learning-Based Optimization algorithm for skin cancer detection. *Healthc. Anal.* **2023**, *3*, 100161. [CrossRef]
75. Dandu, R.; Vinayaka Murthy, M.; Ravi Kumar, Y. Transfer learning for segmentation with hybrid classification to Detect Melanoma Skin Cancer. *Heliyon* **2023**, *9*, e15416. [CrossRef] [PubMed]
76. 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]
77. Li, Z.; Koban, K.C.; Schenck, T.L.; Giunta, R.E.; Li, Q.; Sun, Y. Artificial Intelligence in Dermatology Image Analysis: Current Developments and Future Trends. *J. Clin. Med.* **2022**, *11*, 6826. [CrossRef]
78. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [CrossRef]
79. Rasmussen, C.B.; Kirk, K.; Moeslund, T.B. The Challenge of Data Annotation in Deep Learning: A Case Study on Whole Plant Corn Silage. *Sensors* **2022**, *22*, 1596. [CrossRef]
80. Nanni, L.; Loreggia, A.; Lumini, A.; Dorizza, A. A Standardized Approach for Skin Detection: Analysis of the Literature and Case Studies. *J. Imaging* **2023**, *9*, 35. [CrossRef]
81. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.Y.; Bagul, A.; Langlotz, C.P.; Shpanskaya, K.S.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
82. Liopyris, K.; Gregoriou, S.; Dias, J.; Stratigos, A.J. Artificial Intelligence in Dermatology: Challenges and Perspectives. *Dermatol. Ther.* **2022**, *12*, 2637–2651. [CrossRef]
83. Rezk, E.; Eltorki, M.; El-Dakhakhni, W. Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach. *JMIR Dermatol.* **2022**, *5*, e39143. [CrossRef]
84. Shen, S.; Xu, M.; Zhang, F.; Shao, P.; Liu, H.; Xu, L.; Zhang, C.; Liu, P.; Zhang, Z.; Yao, P.; et al. Low-cost and high-performance data augmentation for deep-learning-based skin lesion classification. *arXiv* **2021**, arXiv:2101.02353.
85. Das, A.; Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv* **2020**, arXiv:2006.11371.
86. Vollert, S.; Atzmueller, M.; Theissler, A. Interpretable Machine Learning: A brief survey from the predictive maintenance perspective. In Proceedings of the 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Vasteras, Sweden, 7–10 September 2021; pp. 1–8. [CrossRef]
87. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Comput. Surv.* **2020**, *53*, 63. [CrossRef]
88. Campagna, M.; Naka, F.; Lu, J. Teledermatology: An updated overview of clinical applications and reimbursement policies. *Int. J. Women's Dermatol.* **2017**, *3*, 176–179. [CrossRef]
89. Blezek, D.J.; Olson-Williams, L.; Missert, A.; Korfiatis, P. AI Integration in the Clinical Workflow. *J. Digit. Imaging* **2021**, *34*, 1435–1446. [CrossRef]

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Diagnostics Editorial Office
E-mail: diagnostics@mdpi.com
www.mdpi.com/journal/diagnostics

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-4124-0