diagnostics

Artificial Intelligence in Image-Based Screening, Diagnostics, and Clinical Care of Cardiopulmonary Diseases

Edited by
Sameer Antani and Sivaramakrishnan Rajaraman

Printed Edition of the Special Issue Published in *Diagnostics*

www.mdpi.com/journal/diagnostics

MDPI

Artificial Intelligence in Image-Based Screening, Diagnostics, and Clinical Care of Cardiopulmonary Diseases

Artificial Intelligence in Image-Based Screening, Diagnostics, and Clinical Care of Cardiopulmonary Diseases

Editors

Sameer Antani
Sivaramakrishnan Rajaraman

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

Sameer Antani
U.S. National Library of Medicine, National Institutes of Health
USA

Sivaramakrishnan Rajaraman
U.S. National Library of Medicine, National Institutes of Health
USA

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Diagnostics* (ISSN 2075-4418) (available at: https://www.mdpi.com/journal/diagnostics/special_issues/AI_Cardiopulmonary).

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

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

ISBN 978-3-0365-6434-0 (Hbk)
ISBN 978-3-0365-6435-7 (PDF)

Cover image courtesy of U.S. National Library of Medicine, National Institutes of Health

Contents

About the Editors

Sameer Antani

Dr. Antani is a Principal Investigator with the Computational Health Research Branch, within the Lister Hill National Center for Biomedical Communications of the National Library of Medicine (NLM), which is a part of the National Institutes of Health (NIH), Bethesda, Maryland, USA. He has over 20 years of research experience in medical image processing, machine learning/artificial intelligence (ML/AI), computer vision, information retrieval, and data science aimed at advancing their role in clinical applications and improving the field of computational life sciences. He has authored nearly 350 publications, of which nearly 130 are in high-impact journals. He is a Fellow of the American Institute of Medical and Biological Engineering (AIMBE), Senior Member of the SPIE and the Institute of Electrical and Electronics Engineers (IEEE).

Sivaramakrishnan Rajaraman

Dr. Rajaraman is working as a research scientist at the National Library of Medicine (NLM), National Institutes of Health (NIH). He is involved in projects aimed at applying computational sciences and engineering techniques toward advancing life science applications. These projects involve the use of medical images to aid healthcare professionals in low-cost decision-making at POC screenings/diagnostics. He is a versatile researcher with expertise in machine learning, data science, biomedical image analysis, and computer vision. Dr. Rajaraman is an Editorial Board member of the journals *PLoS ONE* and *MDPI Electronics*. He is a Life Member of the Society of Photo-optical Instrumentation Engineers (SPIE), a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE), IEEE Engineering in Medicine & Biology Society (EMBS), and the Biomedical Engineering Society (BMES).

diagnostics

MDPI

Editorial

Editorial on Special Issue "Artificial Intelligence in Image-Based Screening, Diagnostics, and Clinical Care of Cardiopulmonary Diseases"

Sivaramakrishnan Rajaraman and Sameer Antani *

Computational Health Research Branch, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
* Correspondence: sameer.antani@nih.gov

Citation: Rajaraman, S.; Antani, S. Editorial on Special Issue "Artificial Intelligence in Image-Based Screening, Diagnostics, and Clinical Care of Cardiopulmonary Diseases". *Diagnostics* **2022**, *12*, 2615. https://doi.org/10.3390/diagnostics12112615

Received: 19 October 2022
Accepted: 25 October 2022
Published: 27 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cardiopulmonary diseases are a significant cause of mortality and morbidity worldwide. The COVID-19 pandemic placed significant demands on clinicians and care providers, particularly in low-resource or high-burden regions. Simultaneously, advances in artificial intelligence (AI), machine learning (ML), and the increased availability of relevant images enhanced the focus on cardiopulmonary diseases. According to the recent American Lung Association report, more than 228,000 people will be diagnosed with lung cancer in the United States alone this year, with the rate of new cases varying by state [1]. Further, heart disease is indiscriminate in ethnic and racial origin, causing mortality. Additionally, infectious diseases, such as tuberculosis (TB) often coupled with the human immunodeficiency virus (HIV) comorbidity, are found with drug-resistant strains that greatly impact treatment pathways and survival rates [2]. The screening, diagnosis, and management of such cardiopulmonary diseases have become difficult owing to the limited availability of diagnostic tools and experts, particularly in low and middle-income regions. Early screening and the accurate diagnosis and staging of cardiopulmonary diseases could play a crucial role in treatment and care and potentially aid in reducing mortality. Radiographic imaging methods such as computed tomography (CT), chest-X-rays (CXRs), and echo-ultrasound are widely used in screening and diagnosis [3–6]. Research on using image-based AI, ML, particularly convolutional neural network (CNN)-based deep learning (DL) methods, can help increase access to care, reduce variability in human performance, and improve care efficiency while serving as surrogates for expert assessment [7]. We find that significant progress has been made [5,8–10] in DL-based medical image modality classification, segmentation, detection, and retrieval techniques which have resulted in a positive impact on clinical and biomedical research. We wanted to capture a snapshot of these advances through a Special Issue collection of peer-reviewed high-quality primary research studies and literature reviews focusing on novel AI/ML/DL methods and their application in image-based screening, diagnosis, and clinical management of cardiopulmonary diseases. These published studies present state-of-the-art AI in cardiopulmonary medicine with an aim toward addressing this global health challenge.

Studying the articles in this collection, the reader will observe that the choice of the DL model depends largely on the characteristics of the data under study [11]. A study of the literature reveals that no individual DL model is optimal for a wide range of medical imaging modalities [12]. Despite delivering superior performance, the performance of DL models is shown to improve with the availability of meaningful data and computational resources [13]. The quality of medical images and their annotations also plays an important role in the success of DL models. The visual characteristics of medical images, viz., shape, size, color, texture, and orientation are unique compared to the natural stock photographic images [14]. The regions of interest (ROIs) concerning the disease manifestations or the organs in medical images are relatively small compared to natural images. Hence, it is crucial to select the optimal DL model for the medical image modality and problem under

study. Unlike natural images, medical images and their associated labels are often scarcely available. Strategies including transfer learning [13,15] and multicenter collaboration [11] have been proposed to handle data scarcity issues. The transfer learning-based approaches are prominently used as they leverage the knowledge learned from a large collection of stock photographic images such as ImageNet [16] to improve performance and generalization in medical visual recognition tasks with a sparse collection of medical data and their associated labels. In this regard, Gozzi et al. [17] proposed the identification of the optimal transfer learning strategy for a CXR classification task. They followed a systematic procedure which is as follows: (i) Several ImageNet-pretrained CNN models were retrained on the publicly available CheXpert [18] CXR dataset. This approach facilitated learning CXR modality-specific feature representations. A study of the literature [19–21] reveals that the medical image modality-specific retraining of ImageNet-pretrained models demonstrates significant gains in related classification, segmentation, and detection tasks. The authors evaluated the classification performance achieved through multiple transfer learning methods such as image feature (embedding) extraction, fine-tuning, stacking, and tree-based classification using a private CXR dataset. They qualitatively evaluated performance using gradient-weighted class activation maps (Grad-CAM) [22]. In this regard, the authors demonstrated superior performance with a 0.856 area under the curve (AUC) using the image embeddings extracted from the penultimate layer of the CNN models and an averaging ensemble of the RF predictions, showcasing it as the optimal transfer learning strategy for the task under study. The Grad-CAM maps showed that the CNN models learned task-specific features to improve prediction performance.

In another study, Huang et al. [23] evaluated the gains achieved through transfer learning in a multi-label CXR classification task. They used a private CXR collection containing multiple abnormalities including aortic sclerosis/calcification, arterial curvature, consolidations, pulmonary fibrosis, enlarged hilar shadows, scoliosis, cardiomegaly, and intercostal pleural thickening, etc. The ImageNet-pretrained CNN models were retrained on the CheXpert and NIH CXR-14 [24] datasets to learn CXR modality-specific representations. The learned knowledge was transferred and finetuned for a related CXR classification task. They further evaluated the gains achieved through multiple transfer learning strategies such as the reuse of pretrained weights, layer transfer where some of the model weight layers were frozen, and model retraining, using the models trained on differently sized CheXpert and NIH CXR-14 datasets. It was observed that CXR modality-specific finetuning of the ImageNet-pretrained models, using the NIH CXR-14 dataset, demonstrated superior prediction performance with an accuracy of 0.935, compared to other models/methods. The authors recommend retraining the CNN models using multiple cross-institutional datasets for promising performance and generalization under conditions of sparse medical data and label availability.

DL models have demonstrated poor performance and generalization in cases where the distribution of the data used to train the models (source distribution) is different compared to the unseen real-world data (target distribution). This lack of generalization could be attributed to several factors including changes in image acquisition protocols, data formatting and labeling, patient heterogeneity based on age, gender, race, and ethnicity, and varying characteristics of the underlying disease manifestations, etc., between the source and target distribution [25]. The discrepancy in the characteristics of the source and target data may eventually lead to domain shift issues resulting in performance degradation and sub-optimal generalization. Under these circumstances, training and evaluating the models using the source data may not accurately reflect real-world settings. Karki et al. [26] discussed the generalization issues with the DL models that were trained to classify Drug-Resistant TB (DR-TB) manifestations from drug-sensitive TB (DS-TB) using CXRs. They observed sub-optimal classification performance with an AUC = 0.65 using an unseen test set in a CNN model that was trained on internal data. The authors observed poor localization using Grad-CAM activation maps as compared to the radiologist-annotated ROIs. Training a multi-task attention model using lesion location information from prior

TB infection helped to improve classification performance (AUC = 0.68) on the blinded test set. The authors highlight differences in acquisition protocols and the variation in non-pathological and non-anatomical image attributes across the datasets that contributed to sub-optimal performance and generalization.

Mueller et al. [27] assessed the diagnostic performance of dual-energy subtraction radiography (DE) [28] in detecting pulmonary emphysema and compared it to the performance achieved using conventional radiography (CR). Pulmonary emphysema, a chronic obstructive pulmonary disease (COPD), blocks airflow in the lungs and causes breathing disorders. CT imaging is reported to be the most sensitive radiological imaging method for detecting and quantifying pulmonary emphysema [29]. The authors used the posteroanterior and lateral radiographic projections acquired from patients using CR, DE, and CT radiography imaging. Expert radiologists were involved in identifying the presence and degree of manifestations consistent with pulmonary emphysema in the DR and CR images while keeping CT as the reference standard. The specificity and recall in detecting and localizing the disease and the inter-reader consensus were measured. The authors observed a high consensus between the readers in identifying pulmonary emphysema manifestations using CR images (Kappa = 0.611) and a moderate consensus (Kappa = 0.433) using the DR images. The authors conclude that the diagnostic performance in terms of detecting, quantifying, and localizing pulmonary emphysema manifestations using CR and DE imaging was comparable.

Li et al. [30] performed a systematic review of the literature to analyze the additional effect of AI-based methods on the performance of physicians to detect cardiopulmonary pathologies using CXR and CT images. They followed the Place of Relevant Intermediary Approach (PRIMA) [31] to record different stages during their literature review process. The authors retrieved relevant literature on AI-based cardiopulmonary screening/diagnosis, published in the last 20 years, using Web of Science, SCOPUS, PubMed, and other literature archives. The authors analyzed human performance in terms of evaluation time, recall, specificity, accuracy, and AUC, in the presence or absence of AI-based assistive tools. It was observed that the average recall increased from 67.8% to 74.6% when human decisions were supplemented by AI assistive tools. A similar improvement was observed in terms of specificity (82.2% to 85.4%), accuracy (75.4% to 81.7%), and AUC (0.75 to 0.80). A significant reduction in the evaluation time was also observed with AI assistance.

In our work [32], we evaluated the gains achieved using modality-specific CNN backbones in a RetinaNet model toward detecting pneumonia-consistent manifestations with CXRs. We retrained ImageNet-pretrained DL models, viz., VGG-16, VGG-19, DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNet on CheXpert and TBX11K datasets to learn CXR modality-specific features. The best-performing model architectures, viz., VGG-16 and ResNet-50, were used as the modality-specific classifier backbones in a RetinaNet-based object detection model. We used focal loss and focal Tversky loss functions to train the classifier backbones. The RetinaNet model was finetuned on the RSNA CXR [33] collection to detect pneumonia-consistent manifestations. We compared detection performance using various weight-initialization methods, viz., random, ImageNet-pretrained, and CXR modality-specific weights, for the classifier backbones. We observed that the VGG-16 and ResNet-50 classifier backbones, initialized with the CXR modality-specific weights, delivered superior performance compared to random and ImageNet-pretrained weight initializations. We further constructed a weighted averaging ensemble of the predictions of the top three performing models, viz., ResNet-50 with CXR image modality-specific weights trained with focal loss, ResNet-50 with CXR image modality-specific weights trained with focal Tversky loss, and ResNet-50 with random weights trained with focal loss, to arrive at the final predictions. We observed that weighted averaging delivered superior values for the mean average precision (mAP) metric (mAP: 0.3272), which was observed to be markedly superior to the state-of-the-art (mAP: 0.2547). We attribute this performance improvement to the key modifications in terms of CXR modality-specific

weight initializations and ensemble learning that reduced prediction variance compared to the constituent models.

A study of the literature reveals that COVID-19 viral infection could cause acute respiratory distress syndrome and may lead to rapidly progressive and lethal pneumonia in infected patients [34]. The laboratory-based real-time reverse transcription polymerase chain reaction (rRT-PCR) test has been reported to be the most sensitive test for identifying COVID-19 infection [35]. However, there are several challenges reported in performing this test, some of which include high false negative rates, delayed processing, variability in test protocols, and reduced recall, among others. CT imaging has been reported to be an effective alternative in identifying COVID-19 disease-consistent evolution, manifestation, and progression [36]. AI-based methods applied to CT imaging could supplement clinical decision-making in identifying COVID-19, particularly in resource-constrained settings to facilitate swift referrals and improve patient care. Suri et al. [37] performed an inter-variability analysis by segmenting the lungs for assessing COVID-19 severity using CT images. The authors used two ground-truth (GT) annotations from different experts and trained U-Net [38] models to segment the lung regions of interest. The authors hypothesized that an AI model could be considered unbiased if the test performance reported with the models when trained on two different GT annotations lay within the 5% range. They further validated their hypothesis through empirical observations. It was observed that the difference in the correlation coefficient obtained using the models trained on two different GT annotations was below the 5% range, thereby showcasing a robust lung segmentation performance.

In another study, Wang et al. [39] measured the three-dimensional (3D) vascular diameter of the aorta and the pulmonary artery in Non-Contrast-Enhanced Chest CT Images to detect pulmonary hypertension. The authors proposed a novel two-stage, 3D-CNN segmentation pipeline to segment the aorta and pulmonary artery and measure the diameter in the 3D space. The authors reported superior segmentation performance in terms of the Dice similarity coefficient (DSC) metric in this segmentation task (0.97 DSC for the aorta and 0.93 DSC for the pulmonary artery). The authors discussed the benefits of such a segmentation approach in terms of providing a non-invasive, pre-operative evaluation of pulmonary hypertension for the optimal planning of surgery and reducing surgical risks.

Khan et al. [40] proposed a joint segmentation and classification network to detect pulmonary lung nodules in publicly available lung CT datasets. Performing unified segmentation and classification would not only help to learn and delineate the semantic regions of interest but also classify them into their respective categories. The authors used the VGG-SegNet [41] for nodule segmentation. The classification model was constructed by appending the classification layers to the VGG-SegNet encoder backbone. The extracted features from the penultimate layer of the trained model were concatenated with hand-crafted features extracted using a gray-level cooccurrence matrix (GLCM), local binary patterns (LBP), and pyramidal histogram of gradient (PHOG) algorithms. A radial basis function kernel-initialized support vector machine (RBF-SVM) classifier learned these concatenated features to improve classification performance with a 97.83% accuracy.

AlOthman et al. [42] proposed a novel feature extraction technique with minimal computational overload to detect and assess the severity of coronary artery disease (CAD) using CT images. The authors used the enhanced features from the accelerated segment test (FAST) to reduce the dimensions of the features extracted from a CNN model. The authors observed improved performance with this feature extraction method, demonstrating accuracies of 99.2% and 98.73% with two benchmark datasets. These findings highlighted the importance of optimal feature selection methods to improve model performance.

Germain et al. [43] analyzed whether CNN models could supersede the performance of experienced clinicians in diagnosing Cardiac Amyloidosis (CA) using Cine-Cardiovascular cine magnetic resonance (Cine-CMR) images. This disease results in the accumulation of amyloid fibrils in cardiac tissues that might lead to progressive cardiomyopathy. Cine

imaging is a type of magnetic resonance imaging (MRI) sequence that captures motion. Cine-CMR is a sensitive diagnostic modality that is used to assess cardiac tissue characterizations and dysfunctions such as CA [44]. The preprocessed systolic and diastolic cine-CMR images were used to train a VGG-based CNN model to classify them as manifesting CA or left ventricular hypertrophy (LVH). The model performance was compared to the outputs of three experienced radiologists. The VGG-based CNN model significantly superseded ($p < 0.05$) human performance on frame-based evaluations, demonstrating an accuracy of 0.746 and AUC of 0.824 as compared to human experts (accuracy = 0.605 and AUC = 0.630). A similar performance improvement was observed in patient-based evaluations. The authors concluded that CNN models have a unique capability to identify CA manifestations in Cine-CMR images compared to trained human experts.

The electrical conductivity is observed to vary considerably among the biological tissues and the movement of gases and fluids within these tissues. Electrical impedance tomography (EIT) is a non-invasive medical imaging modality that uses surface electrodes to measure the electrical permittivity, impedance, and conductivity of biological tissues. However, there exists an inverse problem in EIT imaging in which the non-linear and noisy nature of the EIT imaging acquisition results in sub-optimal reconstruction. Recently, artificial neural networks (ANN) have gained prominence in tackling the inverse problem in EIT imaging. Rixen et al. [45] proposed an ANN model to resolve the EIT inverse problem. The authors reused the dense layers in the ANN model multiple times while considering the rotational symmetries exhibited by the EIT in the circular domain. The authors used an α-blending method to generate synthetic data and augment the training samples. Superior reconstruction performance and robustness to noise were reported with augmented training in which the ANN model demonstrated high values for the amplitude response (AR: 0.14) and low values for the position error (PE: 7.1) compared to conventional methods (AR: 0.1 and PE: 11.0).

In conclusion, the manuscripts published in this Special Issue discuss the novel, state-of-the-art methods for binary, multiclass, and multi-label classification, 2D and 3D image segmentation, object detection and localization, image reconstruction, generalization, recommendation, and inter-reader consensus analysis for identifying, segmenting, classifying, quantifying, reconstructing, and interpreting cardiopulmonary diseases using several medical imaging modalities including CT, MRI, CXRs, and EIT, among others. Nevertheless, deploying these proposed approaches in real-time settings remains an open avenue for research. We would like to express our sincere thanks to the authors for their significant contributions. We hope readers benefit from these research findings, and that the work included in this Special Issue inspires novel methods for diagnosis, treatment, and processes that could eventually promote healthcare.

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

Funding: This work was supported by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health (NIH), USA.

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

References

1. Patel, B.; Priefer, R. Impact of Chronic Obstructive Pulmonary Disease, Lung Infection, and/or Inhaled Corticosteroids Use on Potential Risk of Lung Cancer. *Life Sci.* **2022**, *294*, 120374. [CrossRef]
2. Pande, T.; Cohen, C.; Pai, M.; Ahmad Khan, F. Computer-Aided Detection of Pulmonary Tuberculosis on Digital Chest Radiographs: A Systematic Review. *Int. J. Tuberc. Lung Dis.* **2016**, *20*, 1226–1230. [CrossRef]
3. Rajaraman, S.; Folio, L.R.; Dimperio, J.; Alderson, P.O.; Antani, S.K. Improved Semantic Segmentation of Tuberculosis—Consistent Findings in Chest x-Rays Using Augmented Training of Modality-Specific u-Net Models with Weak Localizations. *Diagnostics* **2021**, *11*, 616. [CrossRef]

4. Zamzmi, G.; Rajaraman, S.; Hsu, L.-Y.; Sachdev, V.; Antani, S. Real-Time Echocardiography Image Analysis and Quantification of Cardiac Indices. *Med. Image Anal.* **2022**, *80*, 102438. [CrossRef]
5. Freedman, M.T.; Lo, S.C.B.; Seibel, J.C.; Bromley, C.M. Lung Nodules: Improved Detection with Software That Suppresses the Rib and Clavicle on Chest Radiographs. *Radiology* **2011**, *260*, 265–273. [CrossRef]
6. Hua, K.L.; Hsu, C.H.; Hidayati, S.C.; Cheng, W.H.; Chen, Y.J. Computer-Aided Classification of Lung Nodules on Computed Tomography Images via Deep Learning Technique. *Onco Targets Ther.* **2015**, *8*, 2015–2022. [CrossRef]
7. Rajaraman, S.; Sornapudi, S.; Alderson, P.O.; Folio, L.R.; Antani, S.K. Analyzing Inter-Reader Variability Affecting Deep Ensemble Learning for COVID-19 Detection in Chest Radiographs. *PLoS ONE* **2020**, *15*, e0242301. [CrossRef]
8. Rajaraman, S.; Jaeger, S.; Thoma, G.R.; Antani, S.K.; Silamut, K.; Maude, R.J.; Hossain, M.A. Understanding the Learned Behavior of Customized Convolutional Neural Networks toward Malaria Parasite Detection in Thin Blood Smear Images. *J. Med. Imaging* **2018**, *5*, 034501. [CrossRef]
9. Rajaraman, S.; Antani, S. *Visualizing Salient Network Activations in Convolutional Neural Networks for Medical Image Modality Classification*; Springer: Singapore, 2019; Volume 1036.
10. Shome, D.; Kar, T.; Mohanty, S.N.; Tiwari, P.; Muhammad, K.; Altameem, A.; Zhang, Y.; Saudagar, A.K.J. Covid-Transformer: Interpretable Covid-19 Detection Using Vision Transformer for Healthcare. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11086. [CrossRef]
11. Rajaraman, S.; Kim, I.; Antani, S.K. Detection and Visualization of Abnormality in Chest Radiographs Using Modality-Specific Convolutional Neural Network Ensembles. *PeerJ* **2020**, *8*, e8693. [CrossRef]
12. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. *Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions*; Springer International Publishing: New York, NY, USA, 2021; Volume 8.
13. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]
14. Suzuki, K. Overview of Deep Learning in Medical Imaging. *Radiol. Phys. Technol.* **2017**, *10*, 257–273. [CrossRef] [PubMed]
15. Zamzmi, G.; Rajaraman, S.; Antani, S. UMS-Rep: Unified Modality-Specific Representation for Efficient Medical Image Analysis. *Informatics Med. Unlocked* **2021**, *24*, 100571. [CrossRef]
16. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8689 LNCS, pp. 818–833.
17. Gozzi, N.; Giacomello, E.; Sollini, M.; Kirienko, M.; Ammirabile, A.; Lanzi, P.; Loiacono, D.; Chiti, A. Image Embeddings Extracted from CNNs Outperform Other Transfer Learning Approaches in Classification of Chest Radiographs. *Diagnostics* **2022**, *12*, 2084. [CrossRef]
18. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Annual Conference on Innovative Applications of Artificial Intelligence, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI, Honolulu, HI, USA, 27 January–1 February 2019; pp. 590–597. [CrossRef]
19. Kim, I.; Rajaraman, S.; Antani, S. Visual Interpretation of Convolutional Neural Network Predictions in Classifying Medical Image Modalities. *Diagnostics* **2019**, *9*, 38. [CrossRef]
20. Rajaraman, S.; Sornapudi, S.; Kohli, M.; Antani, S. Assessment of an Ensemble of Machine Learning Models toward Abnormality Detection in Chest Radiographs. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Berlin, Germany, 23–27 July 2019.
21. Rajaraman, S.; Antani, S.K. Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs. *IEEE Access* **2020**, *8*, 27318–27326. [CrossRef]
22. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-CAM: Why Did You Say That? *arXiv* **2016**, arXiv:1611.07450.
23. Huang, G.-H.; Fu, Q.-J.; Gu, M.-Z.; Lu, N.-H.; Liu, K.-Y.; Chen, T.-B. Deep Transfer Learning for the Multilabel Classification of Chest X-Ray Images. *Diagnostics* **2022**, *12*, 1457. [CrossRef]
24. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1–19.
25. Therrien, R.; Doyle, S. Role of Training Data Variability on Classifier Performance and Generalizability. *Digit. Pathol.* **2018**, *10581*, 58–70. [CrossRef]
26. Karki, M.; Kantipudi, K.; Yang, F.; Yu, H.; Wang, Y.X.J.; Yaniv, Z.; Jaeger, S. Generalization Challenges in Drug-Resistant Tuberculosis Detection from Chest X-Rays. *Diagnostics* **2022**, *12*, 188. [CrossRef]
27. Mueller, J.A.; Martini, K.; Eberhard, M.; Mueller, M.A.; De Silvestro, A.A.; Breiding, P.; Frauenfelder, T. Diagnostic Performance of Dual-Energy Subtraction Radiography for the Detection of Pulmonary Emphysema: An Intra-Individual Comparison. *Diagnostics* **2021**, *11*, 1849. [CrossRef] [PubMed]

28. Rajaraman, S.; Cohen, G.; Spear, L.; Folio, L.; Antani, S. DeBoNet: A Deep Bone Suppression Model Ensemble to Improve Disease Detection in Chest Radiographs. *PLoS ONE* **2022**, *17*, e0265691. [CrossRef] [PubMed]
29. Arthur, R. Interpretation of the Paediatric Chest X-Ray. *Paediatr. Respir. Rev.* **2000**, *1*, 41–50. [CrossRef] [PubMed]
30. Li, D.; Pehrson, L.M.; Lauridsen, C.A.; Tøttrup, L.; Fraccaro, M.; Elliott, D.; Zając, H.D.; Darkner, S.; Carlsen, J.F.; Nielsen, M.B. The Added Effect of Artificial Intelligence on Physicians' Performance in Detecting Thoracic Pathologies on CT and Chest X-Ray: A Systematic Review. *Diagnostics* **2021**, *11*, 2206. [CrossRef] [PubMed]
31. Santosh, K.C.; Allu, S.; Rajaraman, S.; Antani, S. Advances in Deep Learning for Tuberculosis Screening Using Chest X-Rays: The Last 5 Years Review. *J. Med. Syst.* **2022**, *46*, 82. [CrossRef]
32. Rajaraman, S.; Guo, P.; Xue, Z.; Antani, S.K. A Deep Modality-Specific Ensemble for Improving Pneumonia Detection in Chest X-Rays. *Diagnostics* **2022**, *12*, 1442. [CrossRef]
33. Shih, G.; Wu, C.C.; Halabi, S.S.; Kohli, M.D.; Prevedello, L.M.; Cook, T.S.; Sharma, A.; Amorosa, J.K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiol. Artif. Intell.* **2019**, *1*, e180041. [CrossRef]
34. Wang, B.; Jin, S.; Yan, Q.; Xu, H.; Luo, C.; Wei, L.; Zhao, W.; Hou, X.; Ma, W.; Xu, Z.; et al. AI-Assisted CT Imaging Analysis for COVID-19 Screening: Building and Deploying a Medical AI System. *Appl. Soft Comput.* **2021**, *98*, 106897. [CrossRef]
35. Liu, C.; Yin, Q. Automatic Diagnosis of COVID-19 Using a Tailored Transformer-like Network. *J. Phys. Conf. Ser.* **2021**, *2010*, 012175. [CrossRef]
36. Vayá, M.D.L.I.; Saborit, J.M.; Montell, J.A.; Pertusa, A.; Bustos, A.; Cazorla, M.; Galant, J.; Barber, X.; Orozco-Beltrán, D.; García-García, F.; et al. BIMCV COVID-19+: A Large Annotated Dataset of RX and CT Images from COVID-19 Patients. *arXiv* **2020**, arXiv:2006.01174.
37. Suri, J.S.; Agarwal, S.; Elavarthi, P.; Pathak, R.; Ketireddy, V.; Columbu, M.; Saba, L.; Gupta, S.K.; Faa, G.; Singh, I.M.; et al. Inter-Variability Study of Covlias 1.0: Hybrid Deep Learning Models for Covid-19 Lung Segmentation in Computed Tomography. *Diagnostics* **2021**, *11*, 2025. [CrossRef] [PubMed]
38. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2015.
39. Wang, H.J.; Chen, L.W.; Lee, H.Y.; Chung, Y.J.; Lin, Y.T.; Lee, Y.C.; Chen, Y.C.; Chen, C.M.; Lin, M.W. Correction: Wang et Al. Automated 3D Segmentation of the Aorta and Pulmonary Artery on Non-Contrast-Enhanced Chest Computed Tomography Images in Lung Cancer Patients. *Diagnostics* **2022**, *12*, 967. [CrossRef] [PubMed]
40. Khan, M.A.; Rajinikanth, V.; Satapathy, S.C.; Taniar, D.; Mohanty, J.R.; Tariq, U.; Damaševičius, R. VGG19 Network Assisted Joint Segmentation and Classification of Lung Nodules in CT Images. *Diagnostics* **2021**, *11*, 2208. [CrossRef] [PubMed]
41. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
42. AlOthman, A.F.; Sait, A.R.W.; Alhussain, T.A. Detecting Coronary Artery Disease from Computed Tomography Images Using a Deep Learning Technique. *Diagnostics* **2022**, *12*, 2073. [CrossRef]
43. Germain, P.; Vardazaryan, A.; Padoy, N.; Labani, A.; Roy, C.; Schindler, T.H.; El Ghannudi, S. Deep Learning Supplants Visual Analysis by Experienced Operators for the Diagnosis of Cardiac Amyloidosis by Cine-CMR. *Diagnostics* **2022**, *12*, 69. [CrossRef]
44. Oda, S.; Kidoh, M.; Nagayama, Y.; Takashio, S.; Usuku, H.; Ueda, M.; Yamashita, T.; Ando, Y.; Tsujita, K.; Yamashita, Y. Trends in Diagnostic Imaging of Cardiac Amyloidosis: Emerging Knowledge and Concepts. *Radiographics* **2020**, *40*, 961–981. [CrossRef]
45. Rixen, J.; Eliasson, B.; Hentze, B.; Muders, T.; Putensen, C.; Leonhardt, S.; Ngo, C. A Rotational Invariant Neural Network for Electrical Impedance Tomography Imaging without Reference Voltage: RF-REIM-NET. *Diagnostics* **2022**, *12*, 777. [CrossRef] [PubMed]

diagnostics

MDPI

Article

Image Embeddings Extracted from CNNs Outperform Other Transfer Learning Approaches in Classification of Chest Radiographs

Noemi Gozzi [1,2], Edoardo Giacomello [3], Martina Sollini [1,4,*], Margarita Kirienko [5], Angela Ammirabile [1,4], Pierluca Lanzi [3], Daniele Loiacono [3] and Arturo Chiti [1,4]

[1] IRCCS Humanitas Research Hospital, Via Manzoni 56, Rozzano, 20089 Milan, Italy
[2] Laboratory for Neuroengineering, Department of Health Sciences and Technology, Institute for Robotics and Intelligent Systems, ETH Zurich, 8092 Zurich, Switzerland
[3] Dipartimento di Elettronica, Informazione e Bioingegneria, Via Giuseppe Ponzio 34, 20133 Milan, Italy
[4] Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, Pieve Emanuele, 20090 Milan, Italy
[5] Fondazione IRCCS Istituto Nazionale Tumori, Via G. Venezian 1, 20133 Milan, Italy
* Correspondence: martina.sollini@hunimed.eu; Tel.: +39-0282245614

Abstract: To identify the best transfer learning approach for the identification of the most frequent abnormalities on chest radiographs (CXRs), we used embeddings extracted from pretrained convolutional neural networks (CNNs). An explainable AI (XAI) model was applied to interpret black-box model predictions and assess its performance. Seven CNNs were trained on CheXpert. Three transfer learning approaches were thereafter applied to a local dataset. The classification results were ensembled using simple and entropy-weighted averaging. We applied Grad-CAM (an XAI model) to produce a saliency map. Grad-CAM maps were compared to manually extracted regions of interest, and the training time was recorded. The best transfer learning model was that which used image embeddings and random forest with simple averaging, with an average AUC of 0.856. Grad-CAM maps showed that the models focused on specific features of each CXR. CNNs pretrained on a large public dataset of medical images can be exploited as feature extractors for tasks of interest. The extracted image embeddings contain relevant information that can be used to train an additional classifier with satisfactory performance on an independent dataset, demonstrating it to be the optimal transfer learning strategy and overcoming the need for large private datasets, extensive computational resources, and long training times.

Keywords: medical imaging; X-rays; artificial intelligence; transfer learning; explainability

Citation: Gozzi, N.; Giacomello, E.; Sollini, M.; Kirienko, M.; Ammirabile, A.; Lanzi, P.; Loiacono, D.; Chiti, A. Image Embeddings Extracted from CNNs Outperform Other Transfer Learning Approaches in Classification of Chest Radiographs. *Diagnostics* **2022**, *12*, 2084. https://doi.org/10.3390/diagnostics12092084

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 7 July 2022
Accepted: 24 August 2022
Published: 28 August 2022

1. Introduction

The world's population increased by about threefold between 1950 and 2015 (from 2.5 to 7.3 billion), and this trend is projected to continue in the coming decades (a population of 19.3 billion people is expected in 2100), with a growing share of the aging population (≥65 years) (https://www.eea.europa.eu/data-and-maps/indicators/total-population-outlook-from-unstat-3/assessment-1 (accessed on 30 April 2021)). This projected trend is strongly linked to the increasing demand for medical doctors, including imagers. The medical community has offered some warnings about the urgent need to act (https://www.rcr.ac.uk/press-and-policy/policy-priorities/workforce/radiology-workforce-census (accessed on 30 April 2021)), suggesting that artificial intelligence (AI) might partially fill this gap [1]. The joint venture between AI and diagnostic imaging relies on the advantages offered by machine learning approaches to the medical field, which include the automation of repetitive tasks, the prioritization of unhealthy cases requiring urgent referral, and the development of computer-aided systems for lesion detection and diagnosis [2]. Nonetheless, the majority of such AI-based methods are still research prototypes, and only a few

have been introduced in clinical practice [3], despite increasing evidence the superior performance of AI relative to that of doctors [4,5]. A number of reasons may be called upon to explain this fact [6–8]. A successful AI-based tool relies on three main ingredients: an effective algorithm, high computational power, and a reliable dataset. Whereas the first two ingredients are generally available and can leverage several applications in different domains, the latter is perhaps the most critical in medical imaging. An adequate quality and amount of data necessary for machine learning approaches are still challenging or unfeasible in most clinical trials [6]. Accordingly, some strategies can be used to cross the hurdle of datasets in the medical imaging field. These comprise virtual clinical trials [9,10], privacy-preserving multicenter collaboration [11], and transfer learning approaches [12]. In particular, transfer learning, i.e., leveraging patterns learned on a large dataset to improve generalization for another task, is an effective approach for computer vision tasks on small datasets. Besides enabling training with a smaller amount of data, avoiding overfitting, transfer learning has shown remarkable performance in generalizing from one task and/or domain to another [13]. However, the optimal transfer learning strategy has not yet been defined due to the lack of dedicated comparative studies. In this work, we propose:

1. Identification of the best transfer learning approach for medical imaging classification, encompassing three steps: (1) pretraining of CNN models on a large publicly available dataset, (2) development of multiple transfer learning methods, and (3) performance evaluation and comparison;
2. Interpretation of CNN black-box predictions using explainable AI (XAI) on a population level and randomly selected set of examples.

We tested this proof-of-concept approach on chest radiographs (CXRs). CXR is the most frequently performed radiological examination. Thus, the semiautomatic interpretation of CXRs could significantly impact medical practice by potentially offering a solution to the shortage of radiologists.

2. Materials and Methods

2.1. Datasets

The experimental analysis discussed in this paper involved two datasets: (i) CheXpert, a large public dataset, which was used to pretrain several classification models; and (ii) HUM-CXR, a smaller local dataset, which was used to evaluate the investigated transfer learning approaches.

CheXpert. This dataset comprises 224316 CXRs of 65,240 patients collected from the Stanford Hospital from October 2002 to July 2017 [14]. For this study, 191027 CXRs from the original dataset that presented a full reported diagnosis were selected. Each image was annotated with a vector of 14 labels corresponding to major findings in a CXR. Mentions of diseases were extracted from radiology reports with an automatic rule-based system and mapped—for each disease—with positive, negative, and uncertain labels according to their level of confidence. Table 1 shows the data distribution among the 14 labels included in the dataset.

HUM-CXR. We retrospectively collected all chest X-rays performed between 1 May 2019 and 20 June 2019 from the IRCCS Humanitas Research Hospital institutional database. We excluded records (1) not focused on the chest, (2) without images stored in the Institutional PACS, (3) without an available medical report, and (4) without an anteroposterior view. HUM-CXR is composed of 1002 CXRs, including anteroposterior, lateral, and portable (i.e., in bed) CXRs. Labels were manually extracted from medical reports (CJ). Uncertain cases were reassessed by two independent reviewers (M.S. and M.K.), and discordant findings were solved by consensus. Each image was annotated as normal or abnormal; abnormalities were further specified as mediastinum, pleura, diaphragm, device, other, gastrointestinal (GI), pneumothorax (PNX), cardiac, lung, bone, or vascular, resulting in a vector of 12 labels. It was not possible to use available automatic labelers [14] because they are designed for English-language use, whereas our radiological reports were written in Italian. Mediastinum, diaphragm, other, GI, and vascular labels were not included in this

work due to a limited number of available X-rays (<30) and significant inconsistencies with CheXpert labels. Ultimately, 941 CXRs were included in the analysis. Table 2 shows the data distribution of the labels selected for this study.

Table 1. Absolute frequencies of positive, uncertain, and negative samples for each finding (relative frequencies are reported in parentheses) in the CheXpert dataset (n = 191,027).

Label	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16,974 (8.89)	0 (0.0)	174,053 (91.11)
Enlarged card.	30,990 (16.22)	10,017 (5.24)	150,020 (78.53)
Cardiomegaly	23,385 (12.24)	549 (0.29)	167,093 (87.47)
Lung opacity	137,558 (72.01)	2522 (1.32)	50,947 (26.67)
Lung lesion	7040 (3.69)	841 (0.44)	183,146 (95.87)
Edema	49,675 (26.0)	9450 (4.95)	131,902 (69.05)
Consolidation	16,870 (8.83)	19,584 (10.25)	154,573 (80.92)
Pneumonia	4675 (2.45)	2984 (1.56)	183,368 (95.99)
Atelectasis	29,720 (15.56)	25,967 (13.59)	135,340 (70.85)
Pneumothorax	17,693 (9.26)	2708 (1.42)	170,626 (89.32)
Pleural effusion	76,899 (40.26)	9578 (5.01)	104,550 (54.73)
Pleural other	2505 (1.31)	1812 (0.95)	186,710 (97.74)
Fracture	7436 (3.89)	499 (0.26)	183,092 (95.85)
Support devices	107,170 (56.1)	915 (0.48)	82,942 (43.42)

Table 2. Absolute frequencies of positive, uncertain, and negative samples for each finding (relative frequencies are reported in parentheses) in the HUM-CXR dataset (n = 941).

Label	Positive (%)	Negative (%)
Normal	273 (29.01)	668 (70.99)
Cardiac	93 (9.88)	848 (90.12)
Lung	427 (45.38)	514 (54.62)
Pneumothorax	38 (4.04)	903 (95.96)
Pleura	135 (14.35)	806 (85.65)
Bone	137 (14.6)	804 (85.4)
Device	147 (15.56)	794 (84.44)

This study was approved by the Ethical Committee of IRCCS Humanitas Research Hospital (approval number 3/18, amendment 37/19); due to the retrospective design, specific informed consent was waived.

Preprocessing. For both datasets, we selected only anteroposterior images. Concerning CheXpert, following the approach described in [15], we resized the images to 256 × 256, and a chest region of 224 × 224 was extracted using a template-matching algorithm. We then normalized the images by scaling their values in the range [0, 1]; because the original models were pretrained on ImageNet, we further standardized them with respect to ImageNet mean and standard deviation. Concerning HUM-CXR, we selected X-rays acquired with an anteroposterior view, screening the images according to the series description in DICOM format, which had to be anteroposterior, posteroanterior, or portable; the final sample comprised 941 image of 746 patients. First, we clipped pixel values with a maximum threshold of 0.9995 quantile to minimize the noise due to the landmark (see Figure 1).

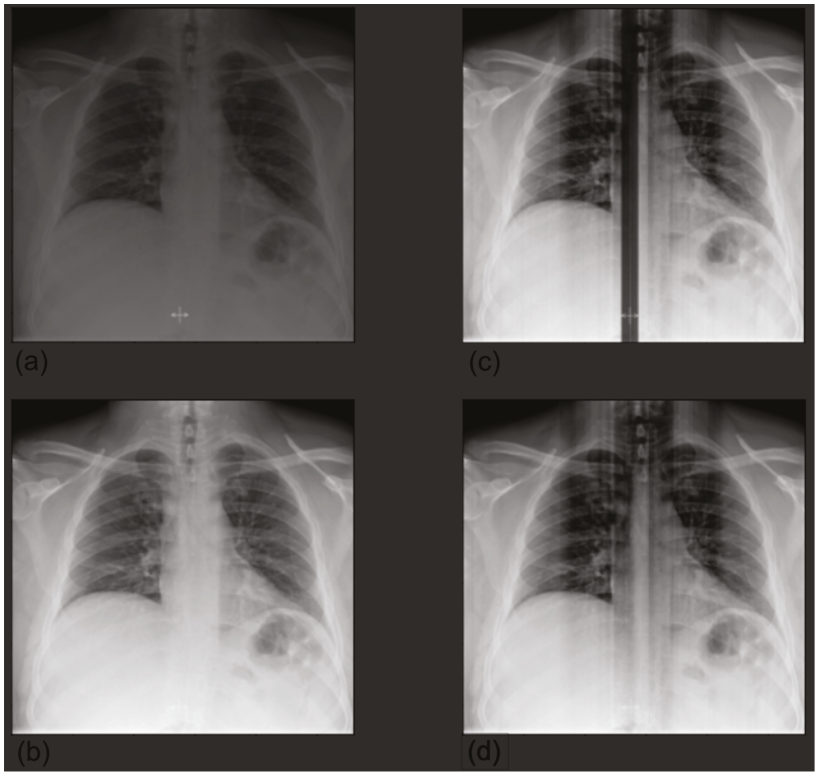


Figure 1. Preprocessing by clipping values larger than the 0.9995 quantile. The presence of a landmark, significantly whiter than the other pixels, created significant noise after normalization (**a**); original image (**b**); clipped image (**c**); normalized original image (**d**). To match the input dimension of the models, we resized the images to 224 × 224 and encoded them as RGB images by repeating the images for three channels. This was a necessary step in order to use the state-of-the-art image classification networks already pretrained on the ImageNet dataset. Then, we normalized each image by scaling the values in the range.

2.2. Pretraining on CheXpert

In this work, we trained several classifiers on the CheXpert dataset to predict CXR findings. Following the protocol described in [15], we considered seven convolutional neural networks (CNNs) with different topologies and numbers of parameters: DenseNet121 (7M parameters), DenseNet169 (12,5M parameters), DenseNet201 (18M parameters) [16], InceptionResNetV2 (54M parameters) [17], Xception (21M parameters) [18], VGG16 (15M parameters) [19], and VGG19 (20M parameters) [19]. We selected these seven network architectures because (i) they are the most common architectures used to perform classification, and (ii) the performance of each architecture differed depending on the labels. With no predominant architectures, aggregating multiple models can improve the final performances. To use these networks as classifiers, we removed the original dense layer and replaced it with a global average pooling (GAP) [20] layer, followed by a fully-connected layer with a number of outputs that matched the number of labels. These seven networks were not trained from scratch; instead, following a common practice in CNN training, we performed a first transfer learning step by initializing the convolutional layers of the networks with weights of pretrained models on the ImageNet dataset [21]. Then, we trained all the weights (both convolutional and classification layers) on the CheXpert dataset, using 90% of the

sample for training and 10% for validation (further details on the training process can be found in our previous work [15]).

Once trained to classify images, the convolutional blocks of CNN models can be employed as a mean to extract a vector of features from images, usually called image embedding. CNNs learn to classify images by learning an effective input representation directly from raw data; the sequence of convolutional layers progressively reduces the size of the input and extracts features from images from low-level features (e.g., edges, pixel intensities, etc.) in early convolutional layers to high-level semantic features in the latest convolutional layers. Accordingly, the last convolutional block, resulting from the training process, is designed to output a vector with the relevant features.

2.3. Transfer Learning

In this paper, we propose three transfer learning approaches, as depicted in Figure 2.

As a reference standard, we mapped the CheXpert labels to the HUM-CXR labels and using the pretrained CNNs. The first transfer learning approach consisted of combining the outputs of pretrained CNNs using stacking [22]. The second approach exploited the pretrained CNNs to compute the image embeddings from HUM-CXR data and used them to train tree-based classifiers. The last approach consisted of tuning the CNNs pretrained on CheXpert on HUM-CXR data. In the remainder of this section, we describe these four approaches in detail.

Pretrained CNNs. This was the most straightforward of the investigated approaches and was used mainly as a baseline. It consisted of providing the HUM-CXR images as input to the CNNs trained on CheXpert and using the output of the networks to classify them based on a mapping between the labels of the two datasets. Table 3 shows the mapping designed as a result of an analysis of the images and labels in the two datasets.

Table 3. Correspondence between CheXpert and HUM-CXR labels.

CheXpert	HUM-CXR
Pleural effusion, pleural other	Pleura
Support devices	Device
Pneumothorax	PNX
Enlarged cardiomediastinum, cardiomegaly	Cardiac
Lung opacity, lung lesion, consolidation, pneumonia, atelectasis, edema	Lung
Fracture	Bone
No findings	Normal

For multiple labels, we selected the maximum output probability of the network for CheXpert labels as the predicted value for the respective HUM-CXR outcome. As reported in previous works [15,23], none of the trained CNNs outperformed any of the other networks on the label problem. Thus, to improve the overall classification performances, we combined the outputs of the trained CNSs through two ensemble methods: simple average and entropy-weighted average. In the case of simple average, the predictions of the classifiers were combined as:

$$\widetilde{y_i} = \frac{1}{N}\sum_{k=1}^{N} p_{k,i} \tag{1}$$

where $p_{k,i}$ is the prediction of classifier k for label i, N is the number of classifiers, and yei is the resulting prediction of the ensemble for label i.

When using entropy-weighted average, the predictions were combined as:

$$\widetilde{y_i} = \sum_{k=1}^{N} (1 - H(p_{k,i}))p_{k,i} \tag{2}$$

where $p_{k,i}$ is the prediction of classifier k for label i, N is the number of classifiers, $H(p) = -p\log_2(\mathrm{p}) - (1-p)\log_2(1-p)$ is the binary entropy function, and $\widetilde{y}_i$ is the resulting prediction of the ensemble for label i.

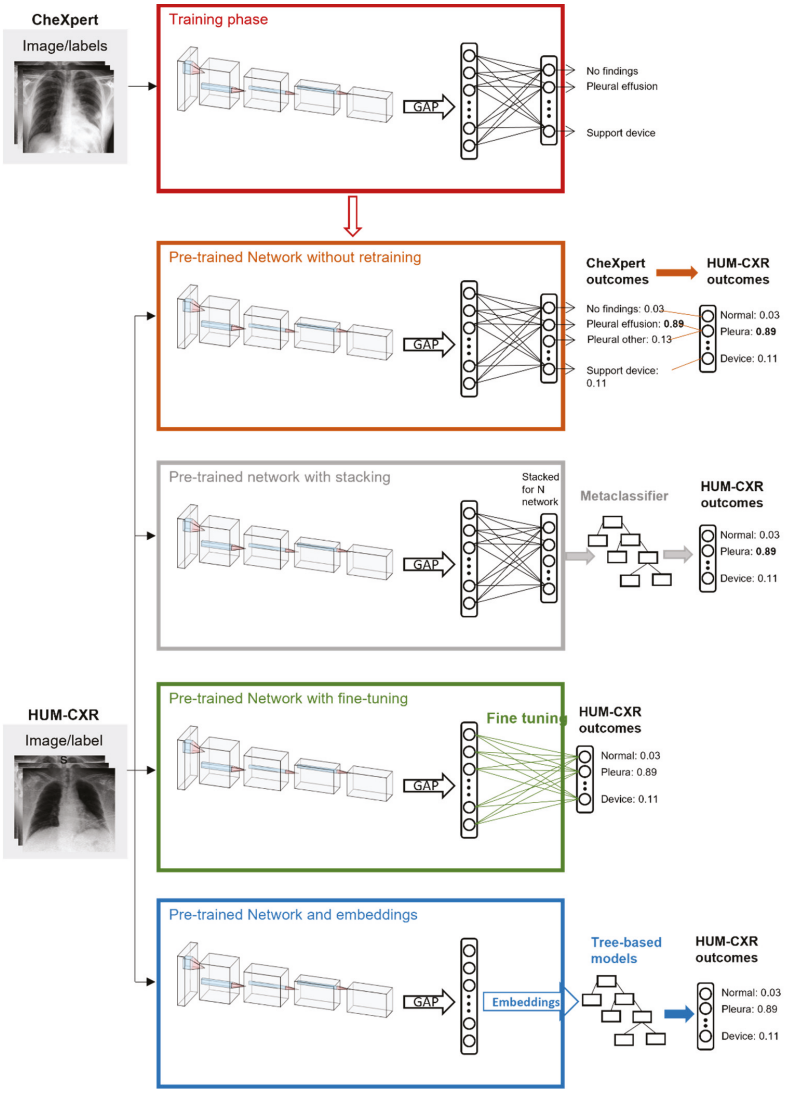


Figure 2. An overview of our experimental design. In the first phase, state-of-the-art image classification networks were tuned on a large public dataset of X-rays (CheXpert [14]). Then, we performed four different steps on the HUM-CXR dataset: (1) we tested the originally trained networks on the X-rays of the new dataset, mapping the HUM_CXR labels to CheXpert labels; (2) we used the originally pretrained networks with a metaclassifier to combine the predictions of each network on the new dataset; (3) we fine-tuned the networks by removing the fully connected classification layer from the seven CNNs trained on CheXpert and replacing it with a seven-output layer that matched HUM-CXR labels; and (4) we extracted the image embeddings from each network and trained tree-based classifiers to predict the HUM-CXR labels starting from the extracted embeddings.

Stacking. This approach extends the previous approach by using a method called stacked generalization or stacking [22]. Instead of combining the outputs of the CNNs with a simple or an entropy-weighted average as described above, we combined them using a metaclassifier trained for this purpose. Thus, we trained a random forest (RF) to predict the label for HUM-CXR samples based on the predictions of the seven CNNs trained on CheXpert and mapped to labels of HUM-CXR, as shown in Table 3. The data were divided into a training set (70%) and a test set (30%).

Tree-based classifiers. This approach exploits the CNNs trained on CheXpert to compute the image embeddings of CXRs included in the HUM-CXR dataset. Image embeddings can be used to predict the label of the corresponding images using much simpler models than CNNs, such as tree-based models. The benefit of using tree-based models with respect to CNNs is that they do not require either high computational resources or extremely large datasets for training, making them suitable for smaller single-institution datasets. In this work, we focused on three kinds of tree-based methods: decision tree (DT), random forest (RF), and extremely randomized trees (XRT). For each method, we trained seven classifiers using the seven CNNs pretrained on CheXpert to compute the image embeddings from the HUM-CXR dataset, with 70% of the samples used for training and 30% for testing. As previously mentioned, for the pretrained CNNs, we applied ensemble methods, i.e., the simple average and the entropy-weighted average, to combine these seven classifiers. We tuned the training hyperparameters of the tree-based classifiers with a grid-search optimization using stratified K-fold cross validation (Table 4 shows the parameters).

Table 4. Embedding model hyperparameters.

Model	Hyperparameters
DT	Max depth = [1, 2, 3, 4, 5, 10, 20], min samples leaf = [1, 2, 4], min samples split = [2, 5, 10], criterion = [gini, entropy] **Final values:** Max depth = 10, min samples leaf = 1, min samples split = 2, criterion = gini
RF	Max depth = [1, 2, 3, 4, 5, 10, 20], min samples leaf = [1, 2, 4], min samples split = [2, 5, 10], criterion = [gini, entropy], number estimators = [10, 20, 30, 50, 100, 200, 300] **Final values:** Max depth = 10, min samples leaf = 4, min samples split = 10, criterion = gini, number of estimators = 100
XRT	Max depth = [1, 2, 3, 4, 5, 10, 20], min samples leaf = [1, 2, 4], min samples split = [2, 5, 10], criterion = [gini, entropy], number estimators = [10,20,30, 50, 100, 200, 300] **Final values:** Max depth = 10, min samples leaf = 2, min samples split = 2, criterion = entropy, number of estimators = 200

The results were combined with simple average and entropy-weighted average.

Fine-tuning. This is a common transfer learning approach in deep learning that consists of adapting and retraining the last layers of a pretrained neural network on different data or tasks [13]. Therefore, we removed the fully connected classification layer from the seven CNNs trained on CheXpert and replaced it with a seven-output layer that matched the HUM-CXR labels. Then, the HUM-CXR (70% training, 10% validation) dataset was used to finetune the original networks. The models were fine-tuned for five epochs with early stopping on the validation AUC set to three epochs. Binary cross entropy was used as loss function, and the learning rate was initially set to 1×10^{-4}, to be reduced by a factor of 10 after each epoch. For each CNN, the best-performing model upon validation was tested on the remaining 20% of the HUM-CXR dataset. The performances were evaluated with simple average, entropy-weighted average, and stacking.

2.4. Performance Assessment

To assess the performances of our classifiers, we computed the area under the receiving operating characteristic (ROC) curve. The ROC curve was obtained by plotting the true positive rate (TPR) (or sensitivity) versus the false positive rate (FPR) (or 1-specificity). Values higher than 0.8 were considered excellent [24], and the training time was recorded.

2.5. Explainability

Despite having proven successful predictive performance, CNNs are recognized as black-box models, i.e., the reasoning behind the algorithm is unknown or known but not interpretable by humans. In order to build trust in AI systems, it is necessary to provide the user with details and reasons to make their functioning clear or easy to understand [25]. We applied gradient-weighted class activation map (Grad-CAM) [26], a state-of-the-art class-discriminative localization technique for CNN interpretation that outputs a visualization of the regions of the input (heat map) that are relevant for a specific prediction. Grad-CAM uses the gradient of an output class in the final convolutional layer to produce a saliency map that highlights areas of the image relevant to detection of the output class. Then, the map is upsampled to the dimensions of the original image, and the mask is superimposed on the CXR. Grad-CAM is considered an outcome explanation method, providing a local explanation for each instance. Therefore, we applied Grad-CAM to randomly selected HUM-CXR data. Grad-CAM heat maps were computed for each CNN model and averaged. In addition to superimposing them on the original image, we used Grad-CAM heat maps to automatically generate a bounding box surrounding the area associated with the outcome. We created a mask with the salient part of the heat map (pixel importance larger than the 0.8 quantile) and used its contours to draw a bounding box highlighting the region of the input that contributed most to the prediction. Grad-CAM saliency maps were compared to saliency masks manually extracted by a radiologist (A.A.). The agreement was evaluated as intersection area over the total area identified by the imager. DeGrave et al. [27] suggested that single local explanations are not enough to validate the correctness of a model against shortcuts and spurious correlations. Therefore, we propose a population-level explanation averaging the saliency maps of 200 randomly sampled images, with the prediction with the highest probability selected.

3. Results

In this section, we first introduce the baseline results (of the networks originally trained on CheXpert) and the performance of the networks following stacking, embedding, and fine tuning. Then, we present an in-depth analysis of the classification by applying Grad-CAM and comparing the extracted saliency maps with those generated by radiologists.

3.1. Baseline with Pretrained CNN

Table 5 shows the performance on the test set achieved by transfer learning without retraining in terms of AUC for each HUM-CXR class and on average.

The results are shown for each CNN, ensembling with averaging, and weighted entropy averaging. Generally, the networks pretrained on CheXpert showed promising performance on the new dataset (HUM-CXR). Failures occurred mainly for bone. Ensembling generally achieved better average results compared to single-model performance.

Table 5. CNN results with pretrained networks without retraining in terms of AUC. Each column represents an HUM-CXR label. We report the results for each network and for the two ensembling strategies. The best results for each class and average are highlighted in bold.

Model	Normal	Cardiac	Lung	PNX	Pleura	Bone	Device	Mean
DenseNet121	0.81	0.84	0.70	0.89	0.87	0.39	0.87	0.766
DenseNet169	0.80	0.79	0.69	0.90	0.87	0.36	0.88	0.755
DenseNet201	0.81	0.78	0.70	0.90	0.87	0.35	0.86	0.754
InceptionResNetV2	0.81	0.83	0.69	0.89	0.87	0.39	0.86	0.762
Xception	0.80	0.77	0.69	**0.91**	0.87	0.44	0.86	0.764
VGG16	0.82	**0.85**	0.70	0.89	0.89	0.41	0.86	0.775
VGG19	0.81	0.83	0.71	0.88	0.89	0.42	0.85	0.772
Averaging	0.82	0.84	0.71	**0.91**	0.89	0.38	**0.89**	0.777
Entropy	0.82	0.83	0.71	**0.91**	0.89	0.37	**0.89**	0.772

3.2. Stacking and Embeddings

Combining the predictions with a metaclassifier (stacking) significantly improved bone classification and the mean classification AUC compared to the baseline. Furthermore, the embeddings extracted from pretrained CNNs were used to train tree-based classifiers. Table 6 shows the performance achieved by stacking and embeddings with DT, RF, and XRT ensembled with simple average and entropy-weighted average.

Table 6. Results of stacking and tree-based models trained on embeddings extracted from pretrained CNNs in terms of AUC. Each column represents an HUM-CXR finding. We report the results for each tree model and for both ensembling strategies. Best results for each class and average are highlighted in bold.

Model	Normal	Cardiac	Lung	PNX	Pleura	Bone	Device	Mean
Stacking	0.85	0.81	**0.74**	0.88	**0.94**	**0.85**	0.84	0.843
DT + averaging	0.81	0.69	0.68	0.75	0.88	0.68	0.78	0.734
RF + averaging	**0.86**	**0.85**	0.72	**0.92**	**0.94**	**0.85**	**0.86**	**0.856**
XRT + averaging	0.85	0.84	0.73	**0.92**	**0.94**	**0.85**	0.85	0.853
DT + entropy	0.81	0.69	0.68	0.75	0.88	0.69	0.78	0.753
RF + entropy	0.85	**0.85**	0.72	**0.92**	**0.94**	**0.85**	0.85	0.853
XRT + entropy	0.85	0.84	0.73	**0.92**	**0.94**	**0.85**	0.84	0.852

The best model (RF + simple averaging) achieved a mean AUC of 0.856 with a maximum of 0.94 for pleura. The results show that stacking and embedding achieved better classification performance compared to the baseline. Complex machine learning models (XRT and RF) achieved better performance than simple decision tree classifiers.

3.3. Fine Tuning

The last set of experiments consisted of fine tuning the classification layers of the pretrained CNNs (Table 7). Single-model performance improved with respect to transfer learning without retraining, except for VGG16 and VGG19. Ensemble AUC increased for all strategies. Fine tuning combined with stacking achieved the best AUC for PNX (0.97), whereas on average, it was performant than the best embedding model. However, these results show that fine tuning alone is not enough to achieve competitive performance, and an additional metaclassifier is required to combine the results. All the described models are available at https://github.com/DanieleLoiacono/CXR-Embeddings.

Table 7. CNN results with fine tuning of the classification layer of pretrained networks in terms of AUC. Each column represents an HUM-CXR finding. We report the results for each single network, for the two ensembling strategies, and for stacking. The best results for each class and average are highlighted in bold.

Model	Normal	Cardiac	Lung	PNX	Pleura	Bone	Device	Mean
DenseNet121	**0.81**	0.73	0.76	0.94	0.90	0.73	0.78	0.807
DenseNet169	0.73	**0.88**	0.77	0.95	0.94	0.72	0.72	0.814
DenseNet201	0.83	0.81	0.71	0.94	0.94	0.74	0.82	0.828
InceptionResNetV2	**0.81**	0.86	**0.79**	0.90	0.93	0.69	0.76	0.818
Xception	**0.81**	0.82	0.73	0.94	0.95	0.68	0.80	0.819
VGG16	0.67	0.81	0.72	0.33	0.95	0.62	0.82	0.704
VGG19	0.64	0.79	0.44	0.83	0.93	0.52	**0.87**	0.717
Averaging	0.83	0.86	0.78	**0.96**	0.96	0.71	0.81	0.842
Entropy	**0.81**	0.86	0.78	0.95	0.96	0.73	0.83	0.845
Stacking	0.80	0.85	0.74	0.93	**0.97**	**0.83**	0.86	**0.853**

3.4. Grad-CAM

We averaged the saliency maps of two batches of 200 randomly sampled images computed with Grad-CAM. The Grad-CAM heat map emphasizes the salient area within the image in shades of red and yellow, whereas the rest of the image is colored in blues and greens. Figure 3 shows that at a population level, the model was generally focused on the lung field and did not take into account shortcuts or spurious correlations that could be present in the borders.

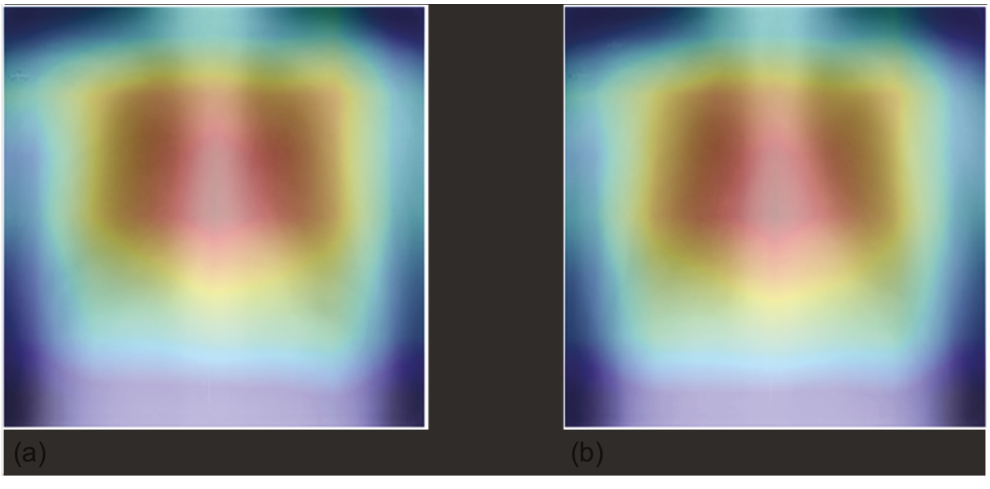


Figure 3. Average of Grad-CAM saliency maps for two batches (panels **a**,**b**) of 200 randomly sampled images. Images confirm that the model focuses on the lung field (image in shades of red and yellow) and does not take into account shortcuts or spurious correlations that could be present in the borders.

We visualized the areas of the CXRs that the model predicted to be most indicative of each prediction using gradient-weighted class activation mappings (GradCAMs) [26] and by creating a bounding box surrounding it. Randomly selected examples are shown in Figures 4–7.

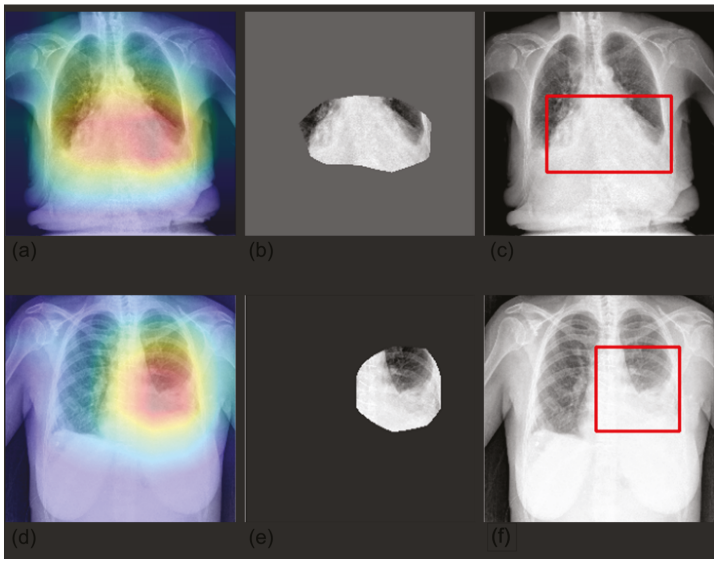


Figure 4. Visualization of pleura prediction maps for two selected CXRs. The panels represent the saliency mask obtained with Grad-CAM (panels **a**,**d**), the relevant area (mask values higher than the 0.8 quantile (panels **b**,**e**)), and the respective bounding box (panels **c**,**f**). The saliency mask focuses on plaura abnormalities, as shown by the heat map (panel **a**,**d**).

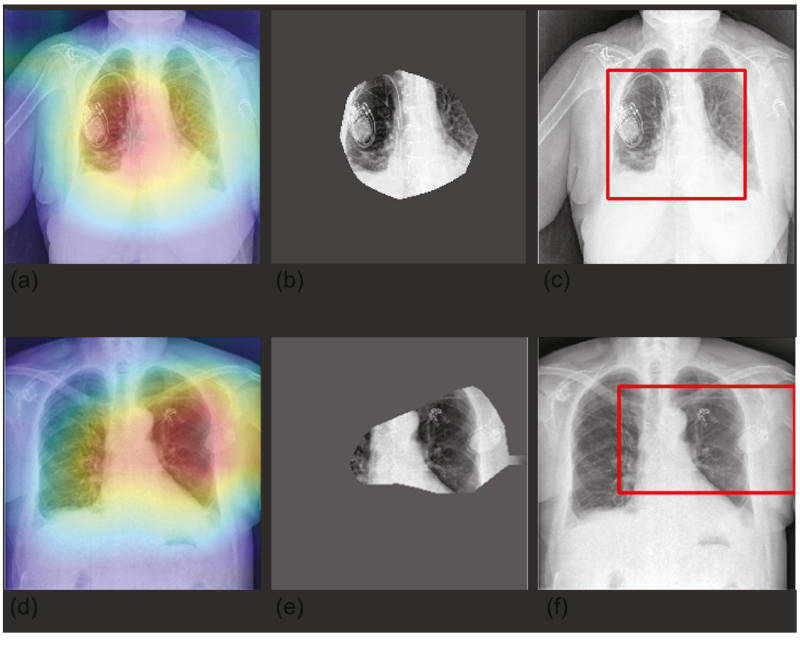


Figure 5. Visualization of device prediction maps for two selected CXRs. The panels represent the saliency mask obtained with Grad-CAM (panels **a**,**d**), the relevant area (mask values higher than the 0.8 quantile) (panels **b**,**e**), and the respective bounding box (panels **c**,**f**). The saliency mask focuses on device (hardware and/or leads), as shown by the heat map (panel **a**,**d**).

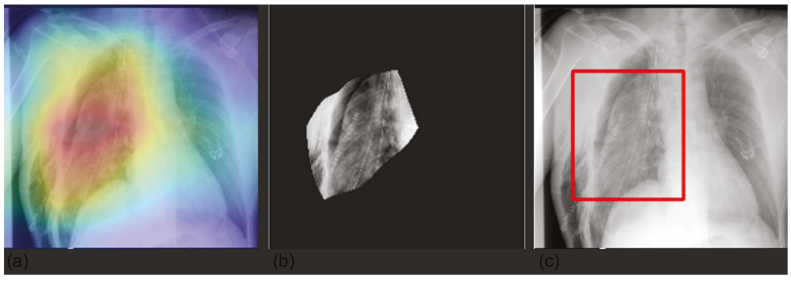


Figure 6. Visualization of pneumothorax prediction maps for a selected CXR. The panels represent the saliency mask obtained with Grad-CAM (panel **a**), the relevant area (mask values higher than the 0.8 quantile) (panel **b**), and the respective bounding box (panel **c**). The saliency mask, as emphasized by the heat map (panel **a**), focuses on the right lung field, which shows the pneumothorax.

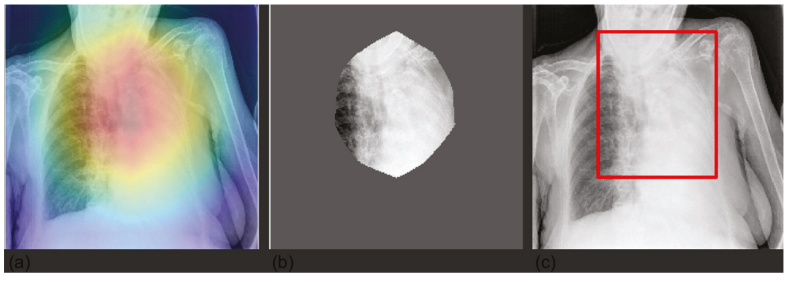


Figure 7. Visualization of lung prediction maps for a selected CXR. The panels represent the saliency mask obtained with Grad-CAM (panel **a**), the relevant area (mask values higher than the 0.8 quantile) (panel **b**), and the respective bounding box (panel **c**). The saliency mask, as emphasized by the heat map (panel **a**), focuses on the left lung, which shows lung abnormality.

In Figure 8, we superimposed the bounding boxes for two classes to show how the model looks at different input areas depending on the specific class.

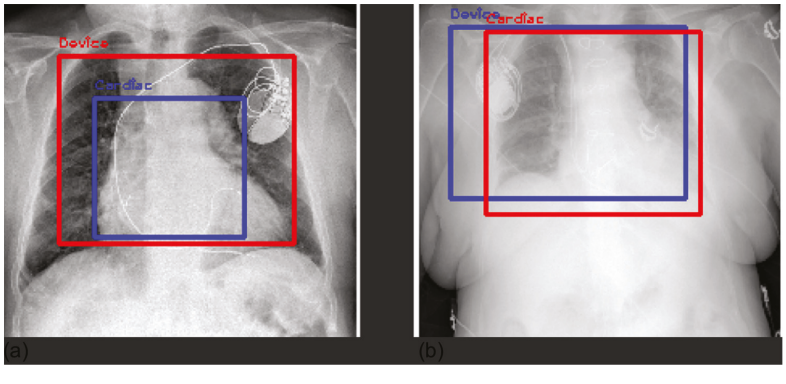


Figure 8. Superimposition of bounding boxes for cardiac (panel **a**, cardiac in blue and device in red) and device (panel **b**, device in blue and cardiac in red) outcomes for two examples.

Furthermore, we compared Grad-CAM maps with saliency masks extracted by a radiologist in terms of common area over the full area identified by the expert. Our models achieved an overall average agreement of 75% (80% lung, 65% pleura, 84% cardiac,

75% PNX, and 67% device), showing how the models automatically learned meaningful features from the images similarly to an expert radiologist. Explainable AI (XAI) algorithms for visualization are successful approaches to identify potential spurious shortcuts that the network may have learned. Overall, our CNNs focused on meaningful areas of the image for the respective prediction. We found some inconsistencies in some examples of device predictions, especially with pacemakers. Figure 9 shows an example of a correct classification but based on an area that does not match well the hardware of the CIED.

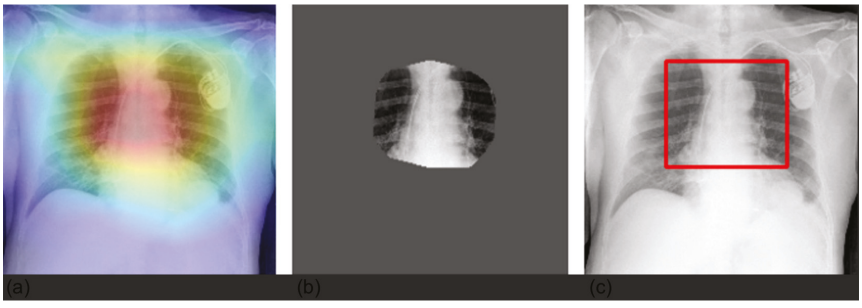


Figure 9. Shortcut for the identification of a pacemaker focusing on the leads: the saliency mask obtained with Grad-CAM (panel **a**), the relevant area (mask values higher than the 0.8 quantile) (panel **b**), and the respective bounding box (panel **c**).

The saliency map highlights the intracardiac leads as the region responsible for device prediction.

4. Discussion

In this work, we first developed and trained CNN models to extract features; thereafter, we proposed the application of different transfer learning approaches to the feature extractor stage of pretrained CNNs to a test dataset, proving the efficiency of transfer learning for domain and task adaptation in medical imaging. Finally, we used Grad-CAM saliency maps to interpret, understand, and explain CNNs and to investigate the presence of potential Clever Hans effects, spurious shortcuts, and dataset biases. Our results support the use of transfer learning to overcome the need for large datasets toward promising AI-powered medical imaging to assist imagers in automating repetitive tasks and prioritizing unhealthy cases. CNNs were first introduced in handwritten zip code recognition in [28], dramatically increasing the performance of deep learning models, especially with N-dimensional matrix input (e.g., three channels images). Since then, CNNs have proven successful capabilities for image analysis, understanding, and classification. Convolutional layers are used in sequence to progressively reduce the input size and simultaneously perform feature extraction, starting from simple patterns in early convolutional layers (edges, curves, etc.) to semantically strong high-level features in deeper layers. The feature maps, i.e., the output at each convolutional step, can be represented as a continuous vector that contains a low-dimensional representation of the image, namely the image embedding. Image embeddings meaningfully represent the original input in a transformed space, reducing the dimensionality. Image embeddings can be used as input to train classifiers based on trees, kernels, Bayesian statistics, etc. Thereby, the advantage of using embeddings lies in benefiting the feature extraction capabilities of CNNs trained on a large dataset of images while designing a specific classifier for new data and, eventually, for a slightly different task. We trained our CNNs with a large publicly available dataset [14] to create an efficient feature extractor that could learn from a large corpus of images. Next, we proposed three transfer learning approaches to apply the feature extractor stage of pretrained CNNs to a new local, independent dataset—HUM-CXRs. Transfer learning has shown remarkable capabilities in computer vision, boosting performance for applications with small datasets.

Transfer learning avoids overfitting, in addition to enabling generalization from one task to another [13], although the generalization capabilities decrease according to the dissimilarity between the base task and target task. Transfer learning has been successful in several fields, including image classification [21,29,30], natural language processing [31–34], cancer subtype discovery [35], and gaming [36]. We applied transfer learning to medical imaging understanding and classification, envisioning the possibility of developing a library of pretrained models for different medical imaging modalities and tasks. Our first TL approach consisted of stacking the predictions of the pretrained CNNs and training an additional metaclassifier to learn the correspondence between them and the HUM-CXR outcomes. The second approach involved two steps: first, the image embeddings of the last convolutional layer were extracted, and additional tree-based classifiers were trained to classify them into the output vector. Finally, we applied a more conventional fine tuning of the last classification layer of each CNN. In this way, the classification layer was customized to the label vector of the new dataset, and the final weights were updated to learn the correspondence between the features extracted by the CNN and the output. In addition to achieving a best classification performance of 0.856 average AUROC, transfer learning with image embeddings has the advantage of minimizing the computational power, dataset dimensions, and time required to adapt the pretrained models to a new dataset and task. The time required to train our tree-based model was in the order of a few minutes, overcoming the need for considerable computational resources, long training times, and GPU availability.

As a proof of concept, we applied this framework to CXRs. CXRs are commonly used for diagnosis, screening, and prognosis; thus, large labeled datasets are already available, such as CheXpert [14], MIMIC-CXR [37], and ChestX-ray [38]. Several previous studies were focused on CXR diagnosis with deep learning, along with these publicly available datasets. CheXNet [39] achieved state-of-the-art performance on fourteen disease classification tasks with ChestXray data [38], and the modified version CheXNeXt [40] achieved radiologist-like performance on ten diseases. On the same dataset, Ye et al. [41] proposed localization of thoracic diseases, in addition to CXR classification. Along with the publication of the dataset, Irvin et al. [37] proposed a solution to achieve performance comparable to that of expert radiologists for the classification of five thoracic diseases. Recently, Pham et al. [23] improved state-of-the-art results on CheXpert, proposing an ensemble of CNN architectures. We used the same dataset as Irvin et al. [37], Pham et al. [23], and Giacomello et al. [15] for pretraining; however, whereas they focused on only five representative findings, we enlarged the classification to seven classes. We can compare the performance of cardiomegaly and pleural effusion, the two findings that are most similar between HUM-CXRs and CheXpert. With respect to cardiomegaly [14,15,23], achieved a best AUROC of 0.828, 0.854, and 0.910, respectively. With respect to pleural effusion, [14,15,23] achieved a best AUROC of 0.940, 0.964, and, 0.936, respectively. Our models obtained by transferring the knowledge acquired on CheXPert to an independent local dataset achieved a best AUROC of 0.88 and 0.97 for cardiac and pleura, respectively. However, [14,15,23] trained and tested on data from the same dataset, i.e., the same distribution, demographic and geographic characteristics (USA residents, Stanford Hospital) and—potential—bias in the data. Hence, these models are potentially prone to the "Clever Hans" effect [42], which limits their actual transition to clinical application. Weber et al. [43] discussed the importance of evaluating the performance of a DL model for applications for which it was not explicitly trained to characterize its generalization capabilities and avoid the Clever Hans effect. Similarly, in a recent analysis of COVID-19 machine learning predictors, Roberts et al. [44] claimed that none of the works under review was reliable enough for the transition from scientific research to clinical routine due to dataset biases, insufficient model evaluation, limited generalizability, and lack of reproducibility, among other reasons. Furthermore, they argued that the scientific community is focusing too much on outperforming benchmarks on public datasets. Using only public datasets without generalizing to new data can lead to overfitting, strongly hindering clinical translation. For

these reasons, in this work, we did not focus on outperforming the state of the art in CXR classification, instead proposing a reproducible framework to overcome some of the main limitations of DL in medical imaging toward a more robust AI-powered clinical routine. In particular, we achieved the following insights. The original models performed poorly on the baseline task (best average AUROC: 0.777), i.e., using the CNNs directly in inference on the new external independent dataset; therefore, even if they were trained on an extremely large dataset, the CNNs were not able to generalize to a new domain and additional data. On the other hand, using transfer learning, in particular with image embeddings, it is possible to adapt the original models to a new domain, i.e., a new hospital, geographic and demographic characteristics, and new tasks, i.e., different labels, with minimum effort and competitive performance (best average AUROC: 0.856). Our approach is not limited to our dataset and the highlighted application; it could be adopted and successfully applied by any other research group or hospital that might need to classify medical images but does not have either a sufficient volume of data or the computational resources to train the model. Following this framework, the resulting models will have excellent feature extraction capability learned from large public datasets, but they will be validated, tailored, and improved with respect to the specific application to achieve optimal results.

Although adherence to the FAIR principles [45] is recommended for scientific data management, a recent systematic review proved the scarce reproducibility of deep learning research. The majority of published deep learning studies focused on medical imaging were non-randomized retrospective trials (only 7% of prospective were tested in a real-world clinical setting) affected by a high risk of bias (72%), with a low adherence to existing reporting standards and without access to data and code (available in 5% and 7% of cases, respectively). Furthermore, deep learning studies typically scantly and elusively describe the used methods, affecting external validity and implementation in clinical settings [7]. To comply with the FAIR principles, respect legal requirements, and preserve the institutional policy, we exhaustively described our methods, providing details for each step, from image analysis to model building, and we made our models available (https://github.com/DanieleLoiacono/CXR-Embeddings) Regardless of the singular value of AUC for each class and the direct comparison between HUM-CXRs and CheXpert among labels, we demonstrated the efficiency of the proposed method. We believe that by making the data available, we guarantee the reproducibility of the proposed methodology, strongly encouraging other groups to repeat our approach with CXRs and/or other images (e.g., computed tomography (CT)).

Finally, CNNs are black-box models that are difficult to interpret, significantly hindering their acceptance in critical fields, such as medicine. Degrave et al. [27] demonstrated that DL models for COVID-19 detection relied on spurious shortcuts, such as lateral markers, image annotations, and borders, to distinguish between positive and negative patients instead of identifying real markers of COVID-19 in the lung field. They suggested that explainable AI (XAI) models should be applied to every AI application in medicine and should be a prerequisite for clinical translation to routine practice. The trustworthiness of AI models for clinical diagnosis and prognosis has to be accurately assessed before they can be applied in a real setting. Several algorithms have been proposed to overcome the intrinsic black-box nature of CNNs. DeepLIFT [46] and SHAP GradientExplainer [47] are based on feature importance, with the aim of measuring the relevance and importance of each input feature in the final predicted output, usually using the coefficients of linear models as interpretability models. Another proposed approach is the use of DGN-AM [48], which evaluates which neurons are maximally activated with respect to a particular input observation, with the aim of identifying input patterns that maximize the output activation. CAM [49], Grad-CAM [26], and LRP [50] create coarse localization maps of the important regions of the input defining the discriminative regions for a specific prediction.

For these reasons, we applied Grad-CAM [26] to our problem, with the aim of (1) interpreting, understanding, and explaining CNN black-box behavior through comprehensible explanations to increase the trust and acceptance in AI for medical imaging for transla-

tion to clinical routine; and (2) investigating the presence of potential Clever Hans effects, spurious shortcuts, and dataset biases. Overall, the explanations provided by Grad-CAM showed a satisfactory ability of the model to identify specific markers and features with respect to the identified class. Grad-CAM saliency maps were found inside the lung field, with particular attention to the correct side of the chest. Double-class images correctly showed the differences between chest findings. However, De Grave et al. [27] were skeptical about presenting only a few examples of explanations, as they may not truthfully represent the real behavior of the model. They discussed the need for a population-level explanation to demonstrate the correctness and reasoning of the entire model, in addition to selecting single examples. In this work, we presented randomly selected examples and population-level explanations averaging two batches of 200 CXRs. The averaged saliency maps presented in Figure 3 demonstrate a high level of attention in the center of the image, whereas the borders are almost useless. Our findings demonstrate that the models were generally focused on the lung field without deploying shortcuts and spurious correlations that may be present outside the lung field, such as annotations, different border dimensions, and lateral markers. Overall, examples of local explanations did not indicate the use of shortcuts as the general model. The only exception we identified concerns the device class, particularly when detecting a CIED. Whereas the model generally correctly focused on the hardware components, in some examples (Figure 9), it correctly classified "device" but exploited the intracardiac leads. This finding is not incorrect, but we would expect the model to focus more on the hardware, i.e., the main box. We believe this might be caused by the original dataset on which the models were pretrained. The device class is extensive and includes lines, tubes, valves, catheters, CIEDs, hardware, coils, etc. However, the percentage proportion of each subclass is not publicly available, so it is possible that "some objects", such as tubes, leads, electrodes, and catheters, are more present than CIEDs, inducing the model to focus on them. Furthermore, we investigated false-positive predictions with respect to the device class. In most cases, we assert that the model was correctly classified devices, although the ground truth was incorrect. The main reason for such false positives is that our labels were extracted from unstructured medical reports. Whereas diseases are clearly written and discussed in the report, cardiac devices, electrodes, prosthesis, and other "objects" may be omitted in the report because are not considered "abnormal" as medical pathologies or clinically relevant. We reasonably believe that with further effort in the definition of the ground truth, the performance of "normal" and "device" labels can be improved.

In contrast to reports by Saporta et al. [51] and Arun et al. [52], who recently demonstrated the unreliability of current saliency methods to explain deep learning algorithms in chest X-ray interpretation, we proved a satisfactory match between Grad-CAM saliency maps and a human benchmark (overall average agreement of 75%), although our data confirmed the same issues (a larger gap between Grad-CAM and radiologist saliency maps in cases of diseases characterized by multiple instances, complex shapes, and small size [51]). We found more variability in some classes, such as pleura and device (65% and 67%, respectively), whereas lung, cardiac, and PNX exhibited greater confidence (80%, 84%, and 75%, respectively).

Results of our study may be of value for both the medical and the scientific communities, as well as for the general population. Overall, our results may impact AI applicability in the medical field, speeding up the grounding system of machine and deep learning algorithms toward clinical application, partially overcoming the problem of the increasing demand for medical doctors. In this work, we analyzed the adaptability and applicability of state-of-the-art imaging classification techniques to a new dataset of images collected in a different country with different scanners. Our models achieved competitive performances (AUC > 85%), correctly identifying and labeling seven classes from X-ray images. We also showed that our model correctly interpreted X-rays similarly to expert radiologists. We proved the feasibility of our approach to train large models and apply them in different countries and hospitals. Next steps of this work will include the investigation of more

recent CNN models [53–55] and the validation of this proof of concept on other datasets, as well as on different kinds of images (e.g., computed tomography). Finally, we have to acknowledge some limitations in our study. First, our results are limited by the retrospective design of our study. Secondly, we did not evaluate the optimal transfer learning approach when we trained the seven CNNs on the CheXpert dataset; however, this was outside of the scope of the present work. Thirdly, different and more recent CNN models should also be used in future research.

5. Conclusions

In this work, we proposed three transfer learning approaches for medical imaging classification. We demonstrated that CNNs pretrained on a large public dataset of medical images can be exploited as feature extractors for a different task (i.e., different classes) and domain (different country, scanner, and hospital) than the original one. In particular, the extracted image embeddings contain relevant information to train an additional classifier with satisfactory performance on an independent local dataset. This overcomes the need for large private datasets, considerable computational resources, and long training times, which are major limitations for the successful applications of AI in clinical practice. Finally, we proved that we can rely on saliency map for deep learning explainability in medical imaging, showing that the models automatically learned how to interpret X-rays in agreement with expert radiologists.

Author Contributions: P.L., D.L. and A.C. ideated the project; D.L., E.G., M.S. and M.K. planned the project; M.S., A.A. and M.K. contributed to data collection and image labelling and retrieval; N.G. performed the analyses; E.G. and D.L. supervised the analyses; N.G., M.S., D.L. and M.K. prepared the draft; A.C. and P.L. commented on the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This retrospective study was performed in accordance with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of IRCCS Humanitas Research Hospital (authorization number 3/18 of 17/04/2018; emended on 22/10/2019 with authorization number 37/19).

Informed Consent Statement: Informed consent was waived (observational retrospective study).

Data Availability Statement: This manuscript represents valid work, and neither this manuscript nor one with substantially similar content under the same authorship has been published or is being considered for publication elsewhere. Arturo Chiti had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All the described models are available at https://github.com/DanieleLoiacono/CXR-Embeddings.

Acknowledgments: We thank Elisa Maria Ragaini, Calvin Jones, and Alessandro Gaja Levra for their support in data collection.

Conflicts of Interest: Arturo Chiti reports a fellowship grant from Sanofi and personal fees from AAA, Blue Earth Diagnostics, and General Electric Healthcare outside the scope of the submitted work. The other authors do not report any conflict of interest.

References

1. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [CrossRef]
2. Sollini, M.; Bartoli, F.; Marciano, A.; Zanca, R.; Slart, R.H.J.A.; Erba, P.A. Artificial intelligence and hybrid imaging: The best match for personalized medicine in oncology. *Eur. J. Hybrid Imaging* **2020**, *4*, 24. [CrossRef]
3. Benjamens, S.; Dhunnoo, P.; Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digit. Med.* **2020**, *3*, 118. [CrossRef]
4. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A.; et al. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [CrossRef]

5. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef]
6. Porenta, G. Is there value for artificial intelligence applications in molecular imaging and nuclear medicine? *J. Nucl. Med.* **2019**, *60*, 1347–1349. [CrossRef]
7. Sollini, M.; Antunovic, L.; Chiti, A.; Kirienko, M. Towards clinical application of image mining: A systematic review on artificial intelligence and radiomics. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 2656–2672. [CrossRef]
8. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D.S.W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ Digit. Med.* **2021**, *4*, 65. [CrossRef]
9. Gelardi, F.; Kirienko, M.; Sollini, M. Climbing the steps of the evidence-based medicine pyramid: Highlights from Annals of Nuclear Medicine 2019. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 1293–1301. [CrossRef]
10. Abadi, E.; Segars, W.P.; Tsui, B.M.W.; Kinahan, P.E.; Bottenus, N.; Frangi, A.F.; Maidment, A.; Lo, J.; Samei, E. Virtual clinical trials in medical imaging: A review. *J. Med. Imaging* **2020**, *7*, 042805. [CrossRef]
11. Kirienko, M.; Sollini, M.; Ninatti, G.; Loiacono, D.; Giacomello, E.; Gozzi, N.; Amigoni, F.; Mainardi, L.; Lanzi, P.L.; Chiti, A. Distributed learning: A reliable privacy-preserving strategy to change multicenter collaborations using AI. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 3791–3804. [CrossRef]
12. Tajbakhsh, N.; Shin, J.Y.; Gurudu, S.R.; Hurst, R.T.; Kendall, C.B.; Gotway, M.B.; Liang, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. Med. Imaging* **2016**, *35*, 1299–1312. [CrossRef]
13. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; MIT Press: Cambridge, MA, USA, 2014; Volume 2, pp. 3320–3328.
14. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 590–597. [CrossRef]
15. Giacomello, E.; Lanzi, P.L.; Loiacono, D.; Nassano, L. Image embedding and model ensembling for automated chest X-ray interpretation. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021.
16. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Manhattan, NY, USA, 2017.
17. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; AAAI Press: Palo Alto, CA, USA, 2017; pp. 4278–4284.
18. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Manhattan, NY, USA, 2017.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Conference Track Proceedings; Bengio, Y., LeCun, Y., Eds.; 2015.
20. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
21. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Manhattan, NY, USA, 2009.
22. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
23. Pham, H.H.; Le, T.T.; Tran, D.Q.; Ngo, D.T.; Nguyen, H.Q. Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels. *medRxiv* **2019**. [CrossRef]
24. Mandrekar, J.N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **2010**, *5*, 1315–1316. [CrossRef]
25. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fus.* **2020**, *58*, 82–115. [CrossRef]
26. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Manhattan, NY, USA, 2017.
27. DeGrave, A.J.; Janizek, J.D.; Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **2021**, *3*, 610–619. [CrossRef]
28. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]

29. Cui, Y.; Song, Y.; Sun, C.; Howard, A.; Belongie, S. Large scale fine-grained categorization and domain-specific transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4109–4118.
30. Ge, W.; Yu, Y. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Manhattan, NY, USA, 2017.
31. Do, C.B.; Ng, A.Y. Transfer learning for text classification. In Proceedings of the 18th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005; MIT Press: Cambridge, MA, USA, 2005; pp. 299–306.
32. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
33. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
34. Peddinti, V.M.K.; Chintalapoodi, P. Domain adaptation in sentiment analysis of twitter. In Proceedings of the 5th AAAI Conference on Analyzing Microtext, San Francisco, CA, USA, 8 August 2011; AAAI Press: Palo Alto, CA, USA, 2011; pp. 44–49.
35. Hajiramezanali, E.; Dadaneh, S.Z.; Karbalayghareh, A.; Zhou, M.; Qian, X. Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 9133–9142.
36. Sharma, M.; Holmes, M.; Santamaria, J.; Irani, A.; Isbell, C.; Ram, A. Transfer learning in real-time strategy games using hybrid CBR/RL. In Proceedings of the 20th International Joint Conference on Artifical Intelligence, Hyderabad, India, 6–12 January 2007; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2007; pp. 1041–1046.
37. Johnson, A.E.W.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
38. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Manhattan, NY, USA, 2017.
39. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.
40. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef]
41. Ye, W.; Yao, J.; Xue, H.; Li, Y. Weakly supervised lesion localization with probabilistic-CAM pooling. *arXiv* **2020**, arXiv:2005.14480.
42. Gozzi, N.; Chiti, A. Explaining a XX century horse behaviour. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 3046–3047. [CrossRef]
43. Weber, M.; Kersting, D.; Umutlu, L.; Schäfers, M.; Rischpler, C.; Fendler, W.; Buvat, I.; Herrmann, K.; Seifert, R. Just another "Clever Hans"? Neural networks and FDG PET-CT to predict the outcome of patients with breast cancer. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 3141–3150. [CrossRef]
44. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. [CrossRef]
45. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
46. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; JMLR.org: Cambridge, MA, USA, 2017; Volume 70, pp. 3145–3153.
47. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
48. Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 3395–3403.
49. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Manhattan, NY, USA, 2016.
50. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140.

51. Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S.Q.; Nguyen, C.D.; Ngo, V.-D.; Seekins, J.; Blankenberg, F.G.; Ng, A.Y.; et al. Benchmarking saliency methods for chest X-ray interpretation. *medRxiv* **2021**. [CrossRef]
52. Arun, N.; Gaw, N.; Singh, P.; Chang, K.; Aggarwal, M.; Chen, B.; Hoebel, K.; Gupta, S.; Patel, J.; Gidwani, M.; et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Artif. Intell.* **2021**, *3*, e200267. [CrossRef] [PubMed]
53. Nafisah, S.I.; Muhammad, G. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence. *Neural Comput. Appl.* **2022**, 1–21. [CrossRef]
54. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
55. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *Proc. IEEE Int. Conf. Comput. Vis.* **2021**, 9992–10002. [CrossRef]

diagnostics

MDPI

Article

Detecting Coronary Artery Disease from Computed Tomography Images Using a Deep Learning Technique

Abdulaziz Fahad AlOthman [1,*], Abdul Rahaman Wahab Sait [1] and Thamer Abdullah Alhussain [2]

[1] Department of Documents and Archive, Center of Documents and Administrative Communication, King Faisal University, P.O. Box 400, Al Hofuf 31982, Al-Ahsa, Saudi Arabia
[2] Programming and Electronic Services Department, Admission and Registration Deanship, King Faisal University, P.O. Box 400, Al Hofuf 31982, Al-Ahsa, Saudi Arabia
* Correspondence: afalothman@kfu.edu.sa

Abstract: In recent times, coronary artery disease (CAD) has become one of the leading causes of morbidity and mortality across the globe. Diagnosing the presence and severity of CAD in individuals is essential for choosing the best course of treatment. Presently, computed tomography (CT) provides high spatial resolution images of the heart and coronary arteries in a short period. On the other hand, there are many challenges in analyzing cardiac CT scans for signs of CAD. Research studies apply machine learning (ML) for high accuracy and consistent performance to overcome the limitations. It allows excellent visualization of the coronary arteries with high spatial resolution. Convolutional neural networks (CNN) are widely applied in medical image processing to identify diseases. However, there is a demand for efficient feature extraction to enhance the performance of ML techniques. The feature extraction process is one of the factors in improving ML techniques' efficiency. Thus, the study intends to develop a method to detect CAD from CT angiography images. It proposes a feature extraction method and a CNN model for detecting the CAD in minimum time with optimal accuracy. Two datasets are utilized to evaluate the performance of the proposed model. The present work is unique in applying a feature extraction model with CNN for CAD detection. The experimental analysis shows that the proposed method achieves 99.2% and 98.73% prediction accuracy, with F1 scores of 98.95 and 98.82 for benchmark datasets. In addition, the outcome suggests that the proposed CNN model achieves the area under the receiver operating characteristic and precision-recall curve of 0.92 and 0.96, 0.91 and 0.90 for datasets 1 and 2, respectively. The findings highlight that the performance of the proposed feature extraction and CNN model is superior to the existing models.

Keywords: coronary artery disease; deep learning; machine learning; cardiopulmonary disease; faster CNN

Citation: AlOthman, A.F.; Sait, A.R.W.; Alhussain, T.A. Detecting Coronary Artery Disease from Computed Tomography Images Using a Deep Learning Technique. *Diagnostics* **2022**, *12*, 2073. https://doi.org/10.3390/diagnostics12092073

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 23 June 2022
Accepted: 23 August 2022
Published: 26 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Coronary artery disease (CAD) has recently become regarded as one of the most dangerous and life-threatening chronic diseases [1]. Blockage and narrowing of the coronary arteries is the primary cause of heart failure. The coronary arteries must be open to provide the heart with adequate blood [2–4]. According to a recent survey, the United States has the highest heart disease prevalence and the highest ratio of heart disease patients [5]. Shortness of breath, swelling feet, fatigue, and other symptoms of heart disease are among the most frequent. CAD is the most common type of heart disease, which can cause chest discomfort, stroke, and heart attack. Besides heart disease, there are heart rhythm issues, congestive heart failure, congenital heart disease, and cardiovascular disease [6].

Traditional methods of investigating cardiac disease are complex [7–10]. The lack of medical diagnostic instruments and automated systems makes pulmonary heart disease detection and treatment challenging in developing nations. However, to reduce the impact

of CAD, an accurate and appropriate diagnosis of cardiac disease is necessary. Developing countries experience an alarming rise in the number of people dying from heart disease [11–16]. According to WHO, CAD is the most frequent type of heart disease, claiming the lives of 360,900 individuals globally in 2019 [17]. The sum accounts for nearly 30% of all deaths worldwide. The number of persons who are victimized is increasing exponentially. Multiple risk factors are involved in the CAD prediction. Thus, healthcare centers require a tool to detect CAD at earlier stages. The recent developments in CNN models enable researchers to develop a prediction model for CAD. However, CNN's structure is complex and needs an excellent graphical processing unit (GPU) to process complex images.

Among conventional approaches, analytical angiography is considered one of the most accurate procedures for detecting heart abnormalities. The disadvantages of angiography include the expensive cost, various side effects, and the need for a high level of technological competence [18]. Due to human error, conventional methods often yield inaccurate diagnoses and take longer to complete. In addition, it is a costly and time-consuming method for diagnosing disease and requires considerable processing.

Artificial intelligence (AI) applications have been increasingly included in clinical diagnostic systems during the last three decades to improve their accuracy. Data-driven decision-making using AI algorithms has been increasingly common in the CAD field in recent years [19]. The diagnostic accuracy can be improved by automating and standardizing the interpretation and inference processes. AI-based systems can help speed up decision-making. Healthcare centers can obtain, evaluate, and interpret data from these emerging technologies and facilitate better patient service [20]. The raw data can significantly affect the quality and performance of AI approaches. As a result, extensive collaboration between AI engineers and clinical professionals is required to improve the quality of diagnosis [21]. The recent CAD detection technique is based on images. Faster predictions can be made for clinicians and computer scientists by deleting irrelevant features. The key features representing the crucial part of CAD decide the performance of the AI techniques [22]. Many studies use deep learning (DL) to determine the existence of CAD.

Convolutional neural networks (CNN) are becoming increasingly popular in medical image processing. CNN was initially demonstrated in medical image analysis in the work of [23] for lung nodule diagnosis. Numerous medical imaging techniques are based on this concept [24–27]. Using a pre-trained network as a feature generator and fine-tuning a pre-trained network to categorize medical pictures are two strategies to transmit the information stored in the pre-trained CNNs. Standard networks can be divided into multiple classes as pre-trained medical image analysis models. Kernels with large receptive fields are used in the higher layers near the input, while smaller kernels are used in the deeper levels. Among the networks in this group, AlexNet is the most widely used and has many applications in medical image processing [28–31].

Deep learning networks are advanced AI techniques and have gained popularity in the medical field. The first network in this category was GoogleNet [32–36]. However, there is a shortcoming in the existing methods, such as more computation time and high-end systems. In addition, the performance of the current CNN architectures is limited in terms of accuracy and F-Measure. In addition, literature is scarce related to integrating feature minimization and CAD techniques. Therefore, this study intends to develop a CNN-based classifier to predict CAD with high accuracy. The objective of the study is as follows:

- To build a CNN model to predict CAD from CT images.
- To improve the performance of CNN by reducing the number of features.

The research questions of the proposed study are:
Research Question-1 (RQ1): How to improve the performance of a CAD detection technique?
Research Question-2 (RQ2): How to evaluate the performance of a CAD detection technique?

The structure of the study is organized as follows: Section 2 presents the recent literature related to CNN and CAD. Section 3 outlines the methodology of the proposed

research. Results and discussion are highlighted in Section 4. Finally, Section 5 concludes the study with its future improvement.

2. Literature Review

High-accuracy data-mining techniques can identify risk factors for heart disease. Studies on the diagnosis of CAD can be found in existing studies [1–5]. Artificial immune recognition system (AIRS), K nearest neighbor (KNN), and clinical data were used to develop a system for diagnosing CAD and achieved an accuracy rate of 87%.

The authors [1] developed and evaluated a deep-learning algorithm for diagnosing CAD based on facial photographs. Patients who underwent coronary angiography or CT angiography at nine Chinese locations participated in a multicenter cross-sectional study to train and evaluate a deep CNN to detect CAD using patient facial images. More than 5796 patients were included in the study and were randomly assigned to training and validation groups for algorithm development. According to the findings, a deep-learning algorithm based on facial photographs can help predict CAD.

According to a study [2], the combination of semi-upright and supine stress myocardial perfusion imaging with deep learning can be used to predict the presence of obstructive disease. The total perfusion deficit was calculated using standard gender and camera type limits. A study [3] employed interferometric OCT in cardiology to describe coronary artery tissues, yielding a resolution of between 10 and 20 μm. Using OCT, the authors [3] investigated the various deep learning models for robust tissue characterization to learn the various intracoronary pathological formations induced by Kawasaki disease. A total of 33 historical cases of intracoronary cross-sectional pictures from different pediatric patients with KD are used in the experimentation. The authors analyzed in-depth features generated from three pre-trained convolutional networks, which were then compared. Moreover, voting was conducted to determine the final classification.

The authors [6] used deep-learning analysis of the myocardium of the left ventricle to identify individuals with functionally significant coronary stenosis in rest coronary CT angiography (CCTA). There were 166 participants in the study who had invasive FFR tests and CCTA scans taken sequentially throughout time. Analyses were carried out in stages to identify patients with functionally significant stenosis of the coronary arteries.

Using deep learning, the researchers [7] investigated the accuracy of the automatic prediction of obstructive disease from myocardial perfusion imaging compared to the overall perfusion deficit. Single-photon emission computed tomography may be used to build deep convolutional neural networks that can better predict coronary artery disease in individual patients and individual vessels. Obstructive disease was found in 1018 patients (62%) and 1797 of 4914 (37%) arteries in this study. A larger area under the receiver-operating characteristic curve for illness prediction using deep learning than for total perfusion deficits. Myocardial perfusion imaging can be improved using deep learning compared to existing clinical techniques.

In the study [8], several deep-learning algorithms were used to classify electrocardiogram (ECG) data into CAD, myocardial infarction, and congestive heart failure. In terms of classification, CNNs and LSTMs tend to be the most effective architectures to use. This study built and verified a 16-layer LSTM model using a 10-fold cross-validation procedure. The accuracy of the classification was 98.5%. They claimed their algorithm might be used in hospitals to identify and classify aberrant ECG patterns.

Author [9] proposed an enhanced DenseNet algorithm based on transfer learning techniques for fundus medical imaging. Medical imaging data from the fundus has been the subject of two separate experiments. A DenseNet model can be trained from scratch or fine-tuned using transfer learning. Pre-trained models from a realistic image dataset to fundus medical images are used to improve the model's performance. Fundus medical image categorization accuracy can be improved with this method, which is critical for determining a patient's medical condition.

The study [10] developed and implemented a heterogeneous low-light image-enhancing approach based on DenseNet generative adversarial network. Initially, a generative adversarial network is implemented using the DenseNet framework. The generative adversarial network is employed to learn the feature map from low-light to normal-light images.

To overcome the gradient vanishing problem in deep networks, the DenseNet convolutional neural network with dense connections combines ResNet and Highway's strengths [11,12]. As a result, all network layers can be directly connected through the DenseNet. Each layer of the network is directly related to the next layer. It is important to remember that each subsequent layer's input is derived from the output of all preceding layers. The weak information transmitted in the deep network is the primary cause of the loss of gradients [13]. A more efficient way to reduce gradient disappearance and improve network convergence is to use the dense block design, in which each layer is directly coupled to input and loss [14].

The authors [15] employed a bright-pass filter and logarithmic transformation to improve the quality of an image. Simultaneous reflectance and illumination estimation (SRIE) was given a weighted variational model by the authors [16] to deal with the issue of overly enhanced dark areas. Authors [17] developed low light image enhancement by illumination map estimation (LIME), which simply estimates the illumination component. The reflection component of the image was calculated using local consistency and structural perception restrictions, and the output result was based on this calculation.

The study [18] used the Doppler signal and a neural network to gain the best possible CAD diagnosis. By combining the exercise test data with a support vector machine (SVM), the authors [19] achieved an accuracy of 81.46% in the diagnosis of coronary artery disease (CAD). By employing multiple neural networks, authors [20] achieved an accuracy of 89.01% for CAD diagnosis using the Cleveland dataset [21]. It is possible to forecast artery stenosis disease using various feature selection approaches, including CBA, filter, genetic algorithm, wrapper, and numerical and nominal attribute selection. Also, Ref. [22] uses a new feature creation method to diagnose CAD.

Inception-v3 [24] is an enhanced version of GoogleNet and is applied in medical image analysis. It categorizes knee images by training support vector machines using deep feature extraction from CaffeNets. Adults' retinal fundus pictures were analyzed using a fine-tuned network to detect diabetic retinopathy [24]. Classification results utilizing fine-tuned networks compete with human expert performance [25]. Recent research has focused on applying deep learning techniques to segment retinal optical coherence tomography (OCT) images [26–28]. Combining CNN and graph search methods, OCT retinal images are segmented. Layer border classification probabilities are used in the Cifar-CNN architecture to partition the graph search layer [29,30].

Authors [31] proposed a deep learning technique to quantify and segment intraregional cystoid fluid using fuzzy CNN. Geographic atrophy (GA) segmentation using a deep network is the subject of another study [33]. An automated CAD detector was developed using a CNN with encoder–decoder architecture [34]. In another study, researchers employed GoogleNet to identify retinal diseases in OCT pictures [35].

Several grayscale features collected from echocardiogram pictures of regular and CAD participants were proposed in [36] as a computer-aided diagnosis approach. In [24], ECG data from routine and CAD participants was evaluated for HR signals. Various methods were used to examine the heart rate data, including non-linear analysis, frequency, and time-domain. They found that CAD participants' heart rate signals were less erratic than normal subjects. The recent CNN models are widely applied in CAD diagnostics [36]. In [37], the authors proposed a model for identifying cardiovascular diseases and obtained a prediction accuracy of 96.75%. Ali Md Mamun et al. [38] argued that a simple supervised ML algorithm can predict heart disease with high accuracy. The authors [39] developed a biomedical electrocardiogram (ECG)-based ML technique for detecting heart disease. Jiely Yan et al. [40], proposed a model to predict ion channel peptide from the images. Table 1 outlines the features and limitations of the existing CNN models.

Table 1. Features of the existing literature.

Authors	Methodology	Features	Limitations
Lin. S et al. [1]	Conducted a cross-sectional study of CAD patients for validating CNN-based CAD.	The findings showed that the deep learning algorithm could support physicians in detecting cardiovascular diseases.	The findings are based on the specific location and lack of a benchmark dataset for evaluating the CNN model.
Jingsi Z et al. [10]	Proposed a low-light image enhancement method.	The DenseNet framework has reduced the noise in the images.	Lack of discussion of the application of bright images.
Abdar M et al. [13]	Integrated genetic algorithm and support vector machine for feature extraction.	The outcome showed that N2Genetic-nuSVM showed a better accuracy.	Lack of comparison with the recent techniques.
Wolterink J.M. et al. [20]	A 3D-dilated CNN is developed to predict the radius of an artery from CCTA images.	Results show that the method extracted 92% of clinically relevant coronary artery segments.	Trained with a small dataset. The outcome may be with the size of the dataset.
Papandrianos N. and Papageorgiou E. [21]	Applied CNN model for CAD detection from images.	The method can differentiate the infarction from healthy patients.	The classification accuracy is better. However, there is a lack of benchmark evaluation techniques.
Nishi et al. [27]	Developed an image segmentation technique for predicting CAD.	The outcome highlighted that the method could produce effective results.	The performance is based on a single dataset.
Cho et al. [30]	Proposed an intravascular ultrasound-based algorithm for classifying attenuation and calcified plaques.	The results outlined that the model achieved 98% accuracy.	The model performance is based on the dataset of 598 patients.
Morris S.A. and Lopez K.N. [31]	Developed a detection model for congenital heart disease in the fetus.	The outcome showed that the model's performance is better than the recent models.	The authors evaluated the model using 1326 fetal echocardiograms.
Cheung et al. [36]	Proposed an image segmentation approach using Unet model.	The model achieved 91,320% of dice similarity coefficient.	The lack of discussion of the image quality used in the study.
Bhanu Prakash Doppala et al. [37]	Developed an ensemble model for cardiovascular disease detection.	The model achieves an accuracy of 96.75%.	The model is based on the voting mechanisms, which may lead to a larger computation time.
Ali Md Mamun et al. [38]	Proposed an ML algorithm for heart disease detection.	The outcome shows that the model has achieved a 100% of accuracy with the Kaggle dataset.	There is a lack of experimentation with the model with different datasets.
Khanna, Ashish et al. [39]	Developed an ML technique for heart disease detection from ECG.	Employed regression model to predict heart disease from ECG.	Limited discussion on the model uncertainty.
Yan, Jielu et al. [40]	Proposed an ML technique for predicting ion channel peptides.	The outcome shows that the model achieves high accurate results.	The dataset is relatively small.

3. Research Methodology

According to the research questions, the researchers developed a CNN architecture to predict positive CAD patients from CT images. Figure 1 presents the proposed architecture. Initially, the images are processed to extract the features. The CNN model treats the extracted features, generating output through an activation function. The following part of this section provides the information related to datasets, feature extraction, CNN construction, and evaluation metrics.

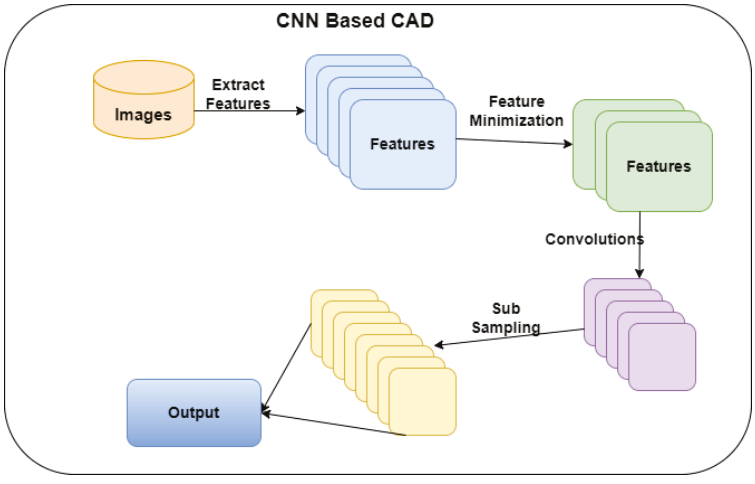


Figure 1. Proposed CNN network for CAD.

In this study, researchers employed two datasets of CT angiography images. The details of the datasets are as follows:

Dataset 1 [4] contains coronary artery image sets of 500 patients. A number of 18 views of the same straightened coronary artery are shown in each mosaic projection view (MPV). The Training–Validation–Test picture sets have a 3/1/1 ratio (300/100/100) with 50% normal and 50% sick cases for each patient in the subset. To improve modeling and dataset balance, 2364 (i.e., 394 × 6) artery pictures were obtained from the 300 training instances. Only 2304 images of the training dataset were augmented: 1. the standard component; 2. all the validation images; and 3. all the testing images. The balance was maintained in the validation dataset by randomly selecting one artery per normal case (50 images) and sick patient (50 images). Figure 2a,b outlines the CT images of positive and negative CAD patients.

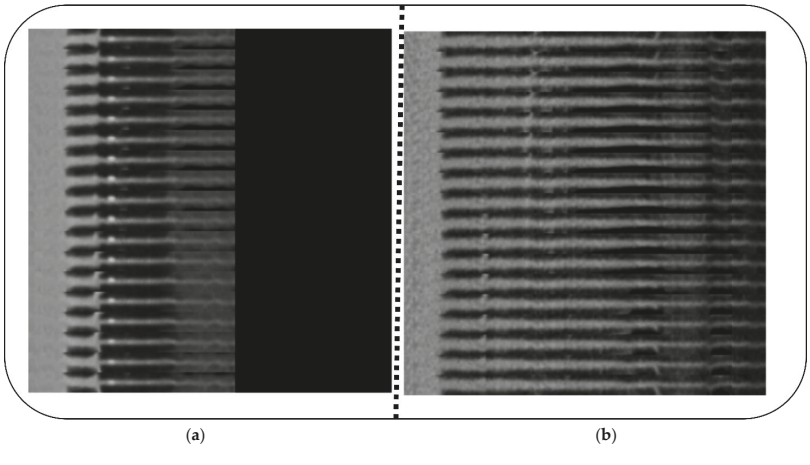


Figure 2. (**a**): Positive individual, (**b**): negative individual.

Dataset 2 [5] consists of CT angiography images of 200 patients. This dataset used images from a multicenter registry of patients who had undergone clinically indicated coronary computed tomography angiography (CCTA). The annotated ground truth included

the ascending and descending aortas (PAA, DA), superior and inferior vena cavae (SVC, IVC), pulmonary artery (PA), coronary sinus (CS), right ventricular wall (RVW), and left atrial wall (LAW). Figure 3 shows the CT images of dataset_2. Table 2 outlines the description of the datasets. Both datasets contain CT images of CAD and Non-CAD patients.

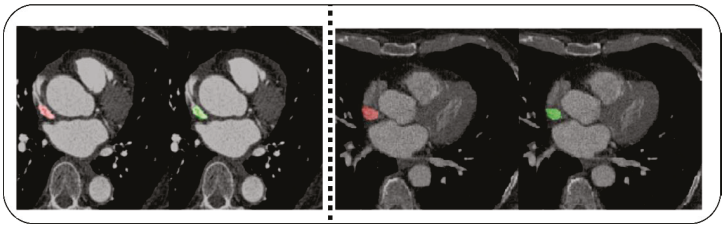


Figure 3. Superior vena cava images of individuals.

Table 2. Description of datasets.

Dataset	Number of Patients	Number of Images	Classification
1	500	2637	2
2	200	716	2

The study applies the following steps for identifying CAD using CNN architecture from datasets:

Step 1: Preprocess images

The CCTA images are processed to fit the feature extraction phase. All images are converted into 600 × 600 pixels. The image size suits the feature extraction process to generate a reduced set of features without losing any valuable data.

Step 2: Feature extraction

The proposed study applies an enhanced features from accelerated segment test (FAST) [6] algorithm for extracting features to support the pooling layer of CNN to produce effective feature maps to answer RQ1. To reduce the processing time of the FAST algorithm, researchers employed the enhanced FAST [5]. Figure 4 showcases the feature extracted from a 4 × 4 image into a 2 × 2 image. In addition, it highlights that the actual image can be reconstructed from a 2 × 2 image to a 4 × 4 image.

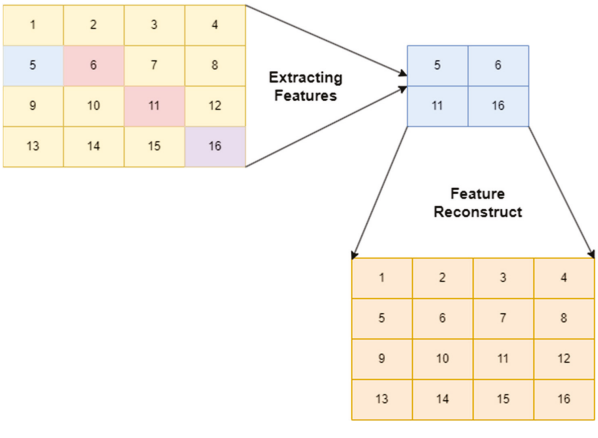


Figure 4. Process of feature extraction.

The extraction process is described as follows:

Let image I of $M_1 \times M_2$ pixels be divided into segments $S_1 \times S_n$. The number of segments is $N_1 \times N_2$, where $N_1 = M_1/S_1$ and $N_2 = M_2/S_n$. The segments are represented in Equation (1).

$$I = \begin{bmatrix} Sd_{1,1} & Sd_{1,2} \dots & Sd_{1,N_n} \\ \vdots & \vdots & \vdots \\ Sd_{N_1,1} & Sd_{N_1,2} \dots & Sd_{N_1,N_n} \end{bmatrix} \quad (1)$$

where $Sd_{x,y}$ referred to the image segment in the x and y direction and is described in Equation (2).

$$Sd_{x,y} = I(i,j) \quad (2)$$

where i and j represent the size of the image segment, Sdx,y.

Both Equations (3) and (4) describe the pixel values of image segments.

$$i = (y-1)M_2, (y-1)M_2 - 1, \dots, yM_2 - 1 \quad (3)$$

$$j = (x-1)M_1, (x-1)M_1 - 1, \dots, yM_1 - 1 \quad (4)$$

The transformation function ensures that the image or segment can be reconstructed to its original form. It supports the proposed method to backtrack the CNN network to fine-tune its performance. The transformation function for each segment is mentioned in Equation (5) as follows:

$$\varphi Md_{x,y} = Z_{S_1} Md_{x,y} Z_{M_2}^T \quad (5)$$

where $\varphi Md_{x,y}$ represents a part of an extracted feature from the image segment, $x = 1, \dots\dots, N_1$, $y = 1, \dots\dots, N_n$ and T represents the transform matrix, $Z_{M1} \in Z_{M_1}^O$, O represents the order of the transformation. The segment can be reconstructed as in Equation (6).

$$Sd_{x,y} = Z_{S_1}^T \varphi Sd_{x,y} Z_{S_n} \quad (6)$$

Sequentially, the process must be repeated $N_1 \times N_n$ times to extract a set of features from the image. Thus, the transform co-efficient of all image segments can be integrated using Equations (7)–(11).

$$\varphi = \begin{bmatrix} Z_{S_1} Sd_{1,1} Z_{S_n}^T & \dots & Z_{S_1} Sd_{1,N_n} Z_{S_2}^T \\ \vdots & \dots & \vdots \\ Z_{S_1} Sd_{N_1,1} Z_{S_n}^T & \dots & Z_{S_1} Sd_{N_1,N_n} Z_{S_2}^T \end{bmatrix} \quad (7)$$

Equations (8) and (9) denote the features F_{S_1} and F_{S_n}, which represent the features that can be constructed using Z_{s1} & Z_{sn}, as follows:

$$F_{S_1} = \begin{bmatrix} Z_{S_1} & O & \dots & O \\ O & Z_{S_1} & \dots & \vdots \\ \ddots & \vdots & \vdots & O \\ O & \dots & O & Z_{S_1} \end{bmatrix} \text{ order of } N_1 \quad (8)$$

$$F_{S_n} = \begin{bmatrix} Z_{S_n} & O & \dots & O \\ O & Z_{S_n} & \dots & \vdots \\ \ddots & \vdots & Z_{S_n} & O \\ O & \dots & O & Z_{S_n} \end{bmatrix} \text{ order of } N_n \quad (9)$$

Equation (10) shows a sample set of features, ∂_{nm}.

$$\sum_{x \in X} \left(F_{S(n,x)} * F_{S(m,x)} \right) = \partial_{nm} \quad (10)$$

Equation (11) defines the reconstruction of the image using the extracted features.

$$I = F_{S_1}^{T} \varphi F_{S_n} \tag{11}$$

Step 3: Processing features

The extracted features $F_{S_1} \ldots\ldots F_{S_n}$ are treated as an input for the proposed CNN. DenseNet ensures the transmission of information between the layers. One of the features of the DenseNet is the direct link between each layer. Thus, a back propagation method can be implemented in DenseNet. The feature extraction process reduces the number of blocks in DenseNet and improves its performance. Therefore, the modified DenseNet contains a smaller number of blocks and parameters. Research studies highlight that the complex network requires a greater number of samples. This study applies DenseNet-161 (K = 48), which includes three block modules. Figure 5 illustrates the proposed DenseNet model. Most CNN models depend on the features to make a decision. Thus, the feature extraction process is crucial in disease detection techniques. The minimal set of features reduces the training time of the CNN model. In addition, the features should support CNN to generate effective results. Researchers applied an edge-detection technique.

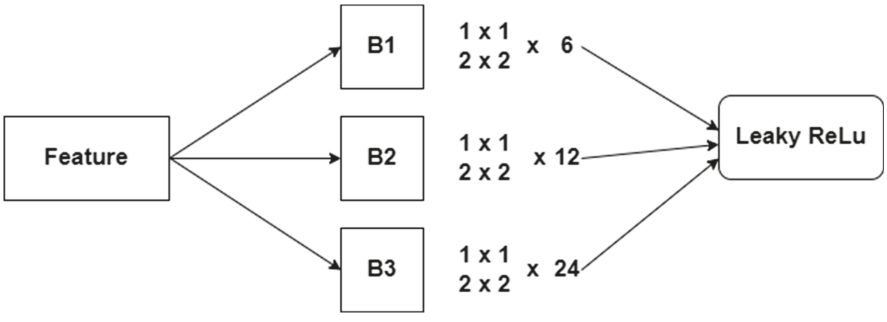


Figure 5. Fine-tuned DenseNet Architecture.

Step 3.1: Pooling layer

Two-dimensional filters are used to integrate the features in the area covered by the two-dimensional filter as it slides over each feature map channel. The dimension of the pooling layer output is in Equation (12):

$$(I_h - f + 1)/l * (I_w - f + 1)/s * I_c \tag{12}$$

where I_h—the height of the feature map, I_w—width of the feature map, I_c—number of channels in the map, f—filter size, l—stride length

Step 3.2: Generating output

Transfer learning is adopted to alter the architecture of DenseNet. Leaky ReLu is used as the activation function. The existing CNN includes are employed. GITHUB portal (https://github.com/titu1994/DenseNet accessed on 7 December 2021) is utilized to implement the existing CNN architecture. The studies [10,18,21] are employed to evaluate the performance of the proposed CNN (PCNN) model. In addition, CNN models, including GoogleNet and Inception V3, are used for performance evaluation. The following form of the sigmoid function is applied for implementing the modified DenseNet. Figure 6 represents the proposed feature extraction for pre-processing the CT images and extracting the valuable features. Furthermore, Figure 7 highlights the proposed CNN technique for predicting CAD from the CT images.

```
Feature_Extraction()
Input: Array of images
Output: Array of features
Initialize n=0, f=0
Assign n = array of images
For (I =1 to n) do
        F = Extract_feature(i)
        If (reconstruct(f)==1) then
                Return f
        Else
                Return -1
        Endif
End for
```

Figure 6. Proposed feature extraction algorithm.

```
Proposed CNN model()
Input: Array of features
Output: Predicting CAD and No CAD
Initialize model by DenseNet with imageNet pre-trained weights.
PCNN_model = Transfer_learning(Pooling layer)
Split_dataset(features)
Initialize output layer(number of neurons, activation function)
If (epoch < 100) then
        Train_model(train_set)
        Compute(loss)
        Compute(Performance metrics)
Else
        Epoch_initialize(min(loss))
        Train_model(train_set)
Endif
Predict = Test_model(test_set)
Compute(Performance metrics)
```

Figure 7. Proposed CNN model.

The study constructs a feed-forward back propagation network. Thus, Leaky ReLu is employed in the study as an activation function in Equation (13) to produce an outcome.

$$f(x) = \max(0, x) \tag{13}$$

Leaky ReLu considers negative value as a minimal linear component of X. The definition of Leaky ReLu is defined as:

Def Leaky_function(I)

If feature(I) < 0:

return 0.01 * f(I)

Else:

return f(I)

Step 4: Evaluation metrics

The study applies the benchmark evaluation metrics, including accuracy, recall, precision, and F-measure, to provide a solution for RQ2. The metrics are computed as shown in Equations (14)–(18):

True positive (TP_{CI}) = predicting a valid positive CAD patient from CT images (CI).

True negative (TN_{CI}) = predicting a valid negative CAD patient from CI.

False positive (FP_{CI}) = predicting a negative CAD patient as positive from CI.

False negative (FN_{CI}) = predicting a positive CAD patient as negative from CI.

$$\text{Recall}= \frac{TP_{CI}}{TP_{CI} + FN_{CI}} \tag{14}$$

$$\text{Precision} = \frac{TP_{CI}}{TP_{CI} + FP_{CI}} \tag{15}$$

$$\text{F} - \text{measure} = \frac{2{*}\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{16}$$

$$\text{Accuracy} = \frac{TP_{CI} + TN_{CI}}{TP_{CI} + TN_{CI} + FP_{CI} + FN_{CI}} \tag{17}$$

$$\text{Specificity} = \frac{TN_{CI}}{TN_{CI} + FP_{CI}} \tag{18}$$

In addition, Matthews correlation coefficient (MCC) (Equation (19)) and Cohen's Kappa (K) (Equation (20)) are employed to ensure the performance of the proposed method.

$$\text{MCC} = \frac{(TP_{CI}{*}TN_{CI}) - (FP_{CI}{*}FN_{CI})}{\sqrt{(TP_{CI} + FP_{CI}) * (TP_{CI} + FN_{CI}) * (TN_{CI} + FP_{CI}) * (TN_{CI} + FN_{CI})}} \tag{19}$$

The minimum MCC is -1, which indicates a wrong prediction, whereas the maximum MCC is +1, which denotes a perfect prediction.

$$\text{K} = \frac{2 * ((TP_{CI}{*}TN_{CI}) - (FP_{CI}{*}FN_{CI}))}{(TP_{CI} + FP_{CI}) * (FP_{CI} + TN_{CI}) * (TP_{CI} + FN_{CI}) * (FN_{CI} + TN_{CI})} \tag{20}$$

MCC and K are class symmetric, reflecting the ML technologies' classification accuracy. Finally, CNN technique computational complexity is presented to find the time and space complexities.

In order to ensure the predictive uncertainty of the proposed CNN (PCNN), the researchers applied standard deviation (SD) and entropy (E). The mathematical expression of the confidence interval (CI) is defined in Equation (21).

$$\text{CI} = a \pm z\frac{\sigma}{\sqrt{N}} \tag{21}$$

where a represents the mean of the predictive distribution of an image $a^{(i)}$, N is the total number of predictions, and z is the critical value of the distribution. The researchers computed CI at 95% confidence. Thus, the value of Z is 1.96.

Finally, the researchers followed E of the prediction to evaluate the uncertainty of the proposed model. It is calculated over the mean predictive distribution. The mathematical expression of E is defined in Equation (22).

$$E(P(y^*|a^*)= -\sum_{i=1}^{C} P(y^*|a^*)\log(P(y^*|a^*) \tag{22}$$

4. Experiment and Results

The PCNN is implemented in Python with Windows 10 Professional platform. The existing algorithms are developed using the GITHUB portal. Both datasets are divided into training and testing sets. Accordingly, the CNN architectures are trained with a relevant training set of dataset_1 and dataset_2.

To evaluate the performance of PCNN, the dataset is utilized using 5-fold cross-validation. Statistical tests, including SD, CI using binary class classification, and E are applied accordingly on the dataset_1 and dataset_2. Table 3 presents the implementation of PCNN during the cross-validation using daaset_1. It highlights that PCNN achieves

more than 98% accuracy, precision, recall, F-measure, and specificity, respectively. Likewise, Table 4 denotes the cross-validation outcome for dataset_2.

Table 3. Performance analysis of PCNN model for dataset_1.

Fold(s)	Accuracy	Precision	Recall	F-Measure	Specificity
1	98.6	97.4	98.4	97.9	98.5
2	98.2	98.2	97.9	98.05	97.8
3	99.1	97.7	98.3	98	98.8
4	99.3	98.6	98.7	98.65	98.8
5	99.6	99.1	99.3	99.2	99.6
Average	98.96	98.2	98.52	98.36	98.7

Table 4. Performance analysis of PCNN model for dataset_2.

Fold(s)	Accuracy	Precision	Recall	F-Measure	Specificity
1	98.4	97.8	98.2	98	98.1
2	97.8	99.3	99.1	99.2	99.3
3	99.1	98.7	98.7	98.7	98.6
4	98.9	98.2	98.6	98.4	98.2
5	99.1	99.3	98.7	99	98.9
Average	98.66	98.66	98.66	98.66	98.62

4.1. Uncertainty Estimation

In this study, the researchers apply Monte Carlo dropout (MC dropout) to compute the model uncertainty. The dropout value ensures that the predictive distribution is not diverse, and CI is insignificant. The researchers experimentally found that the MC dropout value of 0.379 is optimal for this model. The predictive distribution is obtained by evaluating PCNN 200 times for each image. Furthermore, model uncertainty is computed using CI, SD, and E.

Tables 5 and 6 highlight the model uncertainty for dataset_1 and dataset_2, respectively. The proposed model achieved a low entropy and SD for both datasets. It can be observed in Tables 5 and 6 that the average CI of [98.55–98.61] and [98.45–98.51] for dataset_1 and dataset_2 indicate the proposed model has high confidence and minimum variance in its outcome.

Table 5. Model uncertainty analysis outcome for dataset_1.

Fold(s)	CI (%) @95%	SD	Entropy
1	[97.92–97.99]	0.0012	0.0049
2	[98.12–98.19]	0.0019	0.0329
3	[98.79–98.87]	0.0021	0.0319
4	[98.84–98.91]	0.0020	0.0281
5	[99.08–99.11]	0.0017	0.0091
Average	[98.55–98.61]	0.0017	0.0213

Table 6. Model uncertainty analysis outcome for dataset_1.

Fold(s)	CI (%) @95%	SD	Entropy
1	[98.11–98.18]	0.0021	0.0041
2	[97.41–97.49]	0.0018	0.0312
3	[98.42–98.46]	0.0014	0.0187
4	[99.12–99.17]	0.0011	0.0093
5	[99.21–99.26]	0.0009	0.0089
Average	[98.45–98.51]	0.0014	0.0144

Table 7 highlights the performance measures of dataset_1. Among the CNN architectures, PCNN scored a superior accuracy, precision, recall, and specificity of 98.96, 98.2, 98.52, 98.36, and 98.7, respectively. The performance of the Banerjee model [18] is lower than the other CNN architectures. PCNN performs better than the existing CNN models for CAD prediction. Dataset_1 contains a greater number of images. The mapping of features made the CNN architectures generate more features. However, the feature extraction process of the proposed method enabled PCNN to produce a smaller number of features and maintain a better performance than the existing architectures. Figure 8 represents the comparative analysis outcome of CNN. It is evident from Figure 8 that the performance of PCNN is higher than the current architectures.

Table 7. Comparative analysis outcome of CNN model for dataset_1.

Methods/ Measures	Accuracy	Precision	Recall	F-Measure	Specificity
Jingsi model [10]	96.7	96.2	96.7	96.45	97.65
GoogleNet	96.9	97.1	97.4	97.25	96.5
Inception V3	97.8	96.7	96.1	96.4	96.2
Banerjee model [18]	98.1	97.3	97.5	97.4	97.57
Papandrianos model [21]	98.3	97.6	97.1	97.35	97.69
PCNN	98.96	98.2	98.52	98.36	98.7

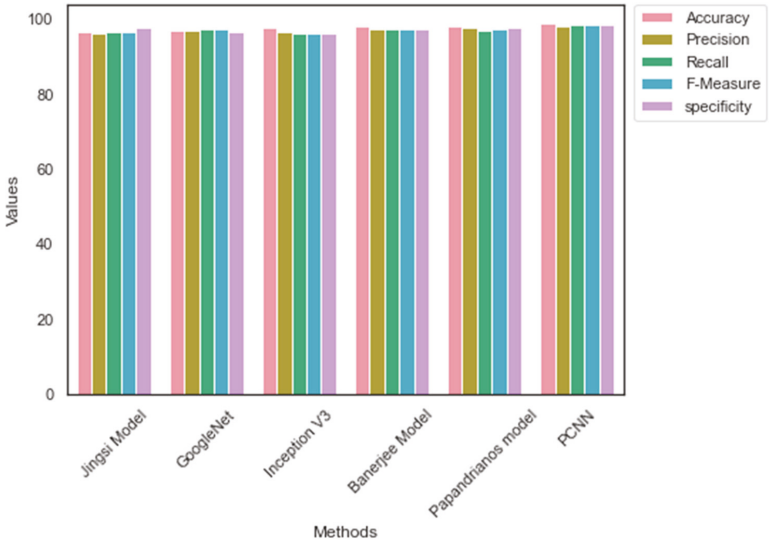


Figure 8. Comparative analysis outcome: Dataset_1.

Likewise, Table 8 outlines the performance of CNN architectures with Dataset_2. The value of accuracy, precision, recall, F-measure, and specificity is 98.96, 98.2, 98.52, 98.36, and 98.7, accordingly. However, GoogleNet has scored low accuracy, precision, recall, F-measure, and specificity of 97.1, 96.7, 97.1, 96.9, and 96.4, respectively. The absence of temporary memory is one of the limitations of the Banerjee model that reduces its predicting performance. In addition, the outcome of Tables 5 and 6 suggest that the performance of PCNN is higher than the existing CNN architectures. Figure 9 shows the relevant graph of Table 6.

Table 8. Comparative analysis outcome of CNN model for dataset_2.

Methods/ Measures	Accuracy	Precision	Recall	F-Measure	Specificity
Jingsi model	96.3	95.8	96.7	96.25	97.2
GoogleNet	97.1	96.7	97.1	96.9	96.4
Inception V3	97.6	97.2	96.8	97	97.3
Banerjee model	98.1	97.6	97.5	97.55	97.1
Papandrianos model	98.3	98.2	97.9	98.05	97.8
PCNN	98.96	98.2	98.52	98.36	98.7

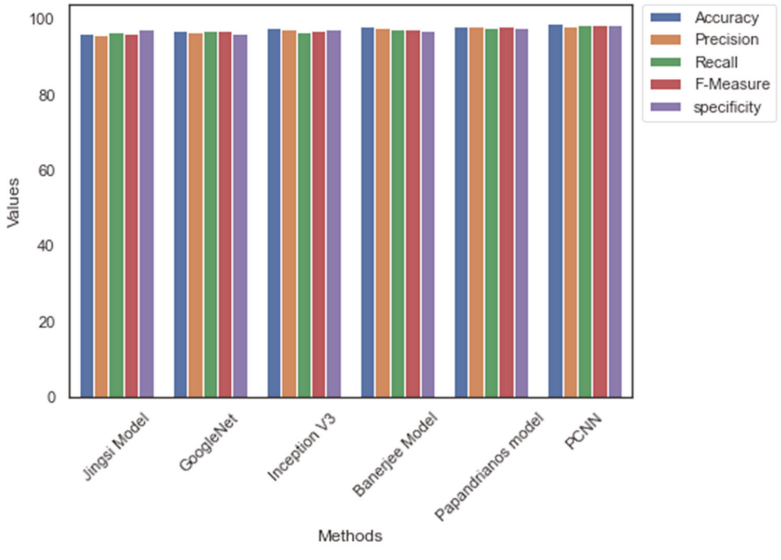


Figure 9. Comparative analysis outcome: Dataset_1.

In addition to the initial comparative analysis, the researcher applied MCC and Kappa to evaluate the performance of PCNN. Figures 10 and 11 reveal that PCNN achieved a superior MCC and K score compared to the existing models.

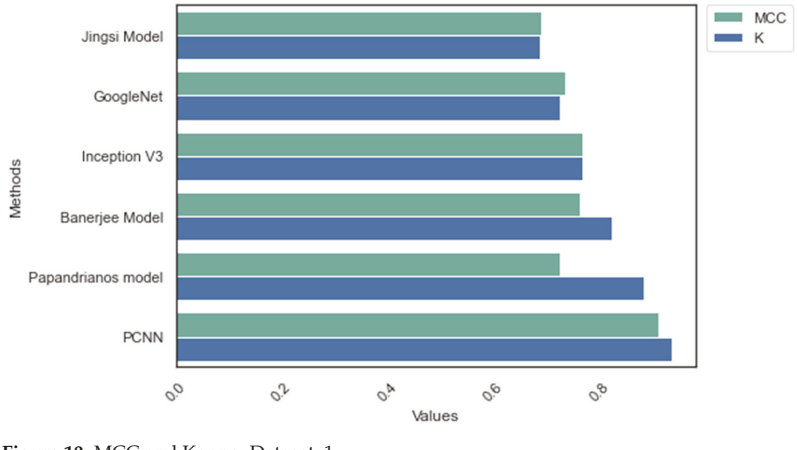


Figure 10. MCC and Kappa: Dataset_1.

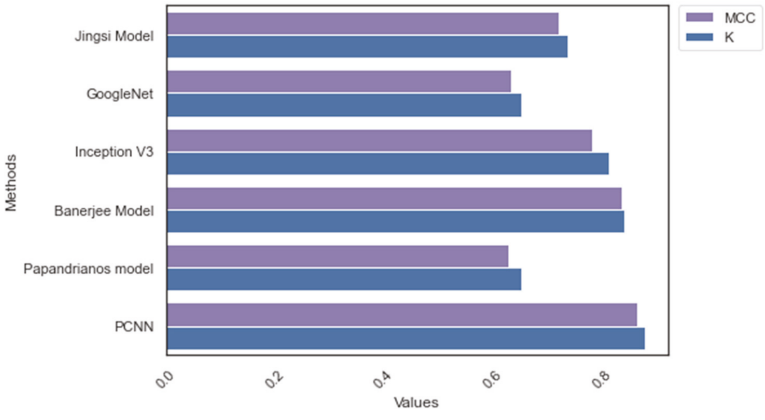


Figure 11. MCC and Kappa: Dataset_2.

Table 9 outlines the memory size and computing time during the training phase. PCNN consumes 121.45 MB and 118.45 MB for Dataset_1 and Dataset_2, accordingly. The computing time of PCNN is 99.32 min and 99.21 min, respectively. The computing time of PCNN is superior to the existing CNN with less memory. Figure 12 highlights CNN's space and computation time for both Dataset_1 and Dataset_2.

Table 9. Memory sizes of CNN for Dataset_1 and Dataset_2.

Methods/Datasets	Dataset_1 (MB)	Dataset_2 (MB)	Dataset_1 Time (Minutes)	Dataset_2 Time (Minutes)
Jingsi model	279.21	189.32	105.26	101.25
GoogleNet	175.69	159.27	102.26	101.36
Inception V3	138.14	142.58	134.56	129.71
Banerjee model	128.54	143.96	116.32	107.25
Papandrianos model	129.65	137.89	101.45	103.59
PCNN	119.25	124.26	100.56	98.89

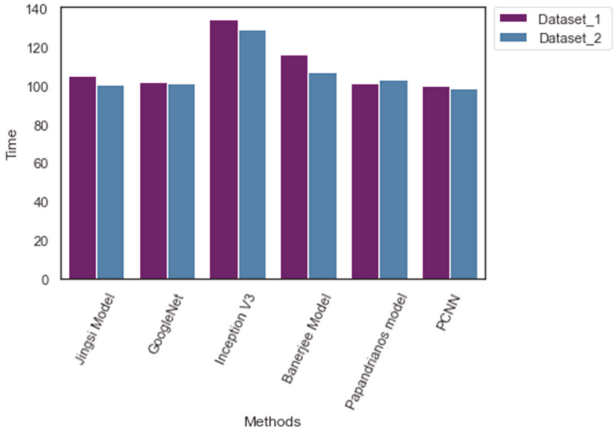


Figure 12. Computation time of CNN models.

Table 10 outlines the error rate of the CNN architectures during the testing phase. The error rate of PCNN is 15.1 and 13.9 for Dataset_1 and Dataset_2, respectively. Nevertheless, Jingsi model scores 20.5 and 19.6, which is higher than other CNN models. The outcome

emphasizes the efficiency of the feature extraction process of PCNN. Figure 13 illustrates the error rate of CNN models.

Table 10. Error rates of CNN for Dataset_1 and Dataset_2.

Methods/Measures	Dataset_1 (%)	Dataset_2 (%)
Jingsi model	20.5	19.6
GoogleNet	19.4	17.3
Inception V3	18.94	17.1
Banerjee model	17.3	16.4
Papandrianos model	16.9	15.7
PCNN	15.1	13.9

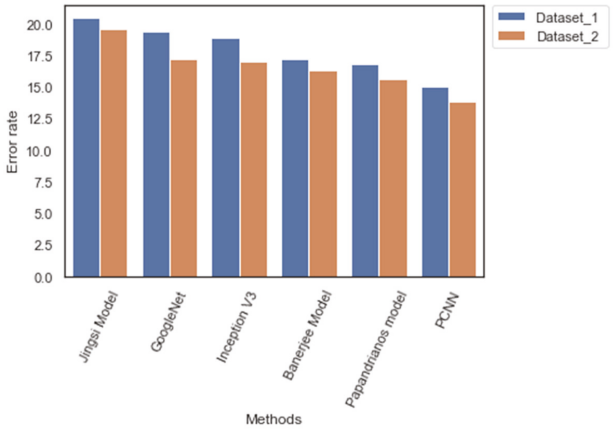


Figure 13. Error rates of CNN models.

Figure 14 represents the receiver operating characteristic (ROC) and precision–recall (PR) curve for dataset_1 during the testing phase. It shows that PCNN achieves a better Area under the ROC curve (AUC) for CAD and No CAD classification, respectively.

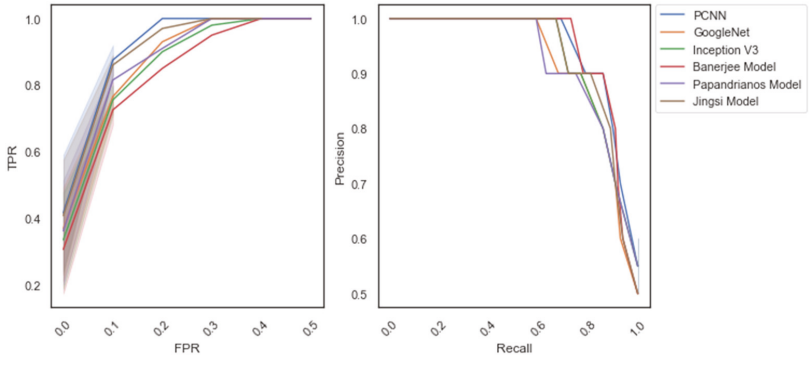


Figure 14. Receiver operating characteristic (ROC) and precision–recall curve: dataset_1.

Similarly, Figure 15 reflects the ROC and PR curve for dataset_2. It outlines that PCNN achieves a better ROC AUC score of 0.93. Furthermore, the AUC score of the PR curve (0.91) indicates that PCNN predicts CAD better than the existing models.

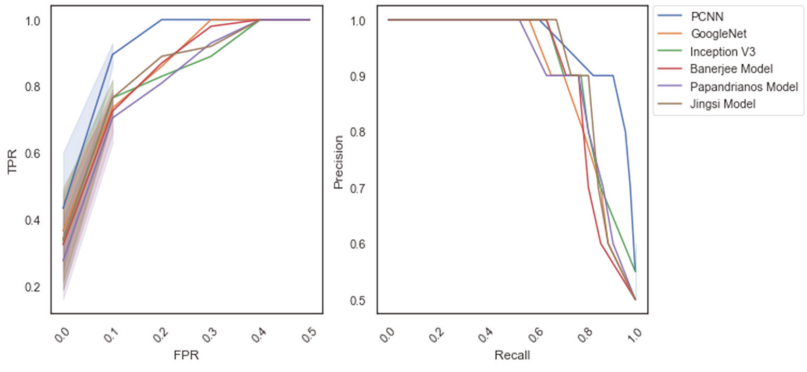


Figure 15. Receiver operating characteristic (ROC) and precision–recall curve: dataset_2.

Table 11 highlights the computational complexities of CNN models for Dataset_1. It is evident from the outcome that PCNN requires a smaller number of parameters (4.3 M), learning rate (1×10^{-4}), number of flops (563 M), and computation time (1.92 s).

Table 11. Computational complexities of CNN for Dataset_1.

Methods/Measures	Number of Parameters	Learning Rate	Number of Flops	Testing Time (s)
Jingsi model	5.1 M	1×10^{-3}	565 M	2.5
GoogleNet	6.7 M	1×10^{-3}	624 M	2.36
Inception V3	7.4 M	1×10^{-4}	594 M	2.7
Banerjee model	14.6 M	1×10^{-3}	1421 M	2.3
Papandrianos model	11.2 M	1×10^{-2}	1530 M	2.1
PCNN	4.3 M	1×10^{-4}	563 M	1.92

Likewise, Table 12 reflects the outcome for Dataset_2. It shows that PCNN generates an output with fewer parameters, flops, and learning rates than the existing CNN models.

Table 12. Computational complexities of CNN for Dataset_2.

Methods/Measures	Number of Parameters	Learning Rate	Number of Flops	Computation Time (s)
Jingsi model	4.3 M	1×10^{-3}	436 M	1.91
GoogleNet	5.6 M	1×10^{-3}	512 M	1.72
Inception V3	6.3 M	1×10^{-5}	402 M	1.86
Banerjee model	9.4 M	1×10^{-4}	921 M	1.98
Papandrianos model	10.3 M	1×10^{-3}	430 M	1.36
PCNN	3.7 M	1×10^{-5}	403 M	1.15

4.2. Clinical Insights and Limitations

PCNN generates outcomes that are superior to the existing CNN models. It can be employed in real-time applications to support physicians in diagnosing CAD. In addition, it can be integrated with Internet of Things devices to support healthcare centers in identifying CAD at an earlier stage. The feature extraction and the pooling layer of PCNN can detect CAD from complex CT images. The dropout layer reduces the neurons to avoid limitations such as overfitting and underfitting. PCNN applies loss function to compute the kernels and weights of the model. It optimizes the model's performance and generates a meaningful outcome.

PCNN produces an effective result and supports CAD diagnosing process. However, a few limitations need to be addressed in future studies. The multiple layers of CNN

increase the training time and require a better graphical processing unit. The imbalanced dataset may reduce the performance of the proposed method. The researcher introduced the concept of temporary storage to hold the intermediate results.

Nonetheless, there is a possibility of losing information due to multiple features. The lack of co-ordinate frames may lead to the adversarial visualization of images. The feature selection process can improve the images' internal representation. Finally, the structure of PCNN requires a considerable amount of data to produce an exciting result. To maintain the better performance, data pre-processing is necessary to handle image rotation and scaling tasks.

5. Conclusions

This study developed a CNN model for predicting CAD from CT images. The existing CNN architectures require a high-end hardware configuration for processing complex images. A feature extraction technique is employed to support the proposed CNN model. The proposed method modifies the existing DenseNet architecture in order to implement a feed-forward back-propagation network. Two benchmark datasets are used for the performance evaluation. The experiment analysis's outcome highlights the superior performance of the proposed CNN model in terms of accuracy, precision, recall, F-measure, and specificity. Moreover, the proposed CNN's memory consumption and computation time during the training phase are lower than the existing CNNs. In addition, ROC and PR curve analysis suggest that the proposed method can predict CAD with a lower false positive rate with higher prediction accuracy. Thus, the proposed method can support the physician in detecting and preventing CAD patients. In the future, the proposed model can be implemented to predict CAD from electronic health records.

Author Contributions: Conceptualization, A.F.A., A.R.W.S. and T.A.A.; Data curation, A.F.A. and A.R.W.S.; Formal analysis, A.R.W.S.; Investigation, A.R.W.S.; Methodology, A.F.A. and A.R.W.S.; Project administration, A.F.A. and A.R.W.S.; Resources, A.F.A. and A.R.W.S.; Software, A.R.W.S.; Validation, A.R.W.S.; Visualization, T.A.A.; Writing—original draft, A.F.A. and A.R.W.S.; Writing—review & editing, T.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. GRANT843].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

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

References

1. Lin, S.; Li, Z.; Fu, B.; Chen, S.; Li, X.; Wang, Y.; Wang, X.; Lv, B.; Xu, B.; Song, X.; et al. Feasibility of using deep learning to detect coronary artery disease based on facial photo. *Eur. Heart J.* **2020**, *41*, 4400–4411. [PubMed]
2. Betancur, J.; Hu, L.H.; Commandeur, F.; Sharir, T.; Einstein, A.J.; Fish, M.B.; Ruddy, T.D.; Kaufmann, P.A.; Sinusas, A.J.; Miller, E.J.; et al. Deep learning analysis of upright-supine high-efficiency SPECT myocardial perfusion imaging for prediction of obstructive coronary artery disease: A multicenter study. *J. Nucl. Med.* **2019**, *60*, 664–670. [CrossRef] [PubMed]
3. Abdolmanafi, A.; Duong, L.; Dahdah, N.; Adib, I.R.; Cheriet, F. Characterization of coronary artery pathological formations from OCT imaging using deep learning. *Biomed. Opt. Express* **2018**, *9*, 4936–4960. [CrossRef]
4. Demirer, M.; Gupta, V.; Bigelow, M.; Erdal, B.; Prevedello, L.; White, R. Image Dataset for a CNN Algorithm Development to Detect Coronary Atherosclerosis in Coronary CT Angiography. Mendeley Data, V1. Available online: https://data.mendeley.com/datasets/fk6rys63h9/1 (accessed on 2 November 2021).
5. Hong, Y.; Commandeur, F.; Cadet, S.; Goeller, M.; Doris, M.K.; Chen, X.; Kwiecinski, J.; Berman, D.S.; Slomka, P.J.; Chang, H.J.; et al. Deep learning-based stenosis quantification from coronary CT angiography. *Proc. SPIE Int. Soc. Opt. Eng.* **2019**, *10949*, 109492I.
6. Zreik, M.; Lessmann, N.; van Hamersvelt, R.W.; Wolterink, J.M.; Voskuil, M.; Viergever, M.A.; Leiner, T.; Išgum, I. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. *Med. Image Anal.* **2018**, *1*, 72–85. [CrossRef]

7. Lih, O.S.; Jahmunah, V.; San, T.R.; Ciaccio, E.J.; Yamakawa, T.; Tanabe, M.; Kobayashi, M.; Faust, O.; Acharya, U.R. Comprehensive electrocardiographic diagnosis based on deep learning. *Artif. Intell. Med.* **2020**, *103*, 101789. [CrossRef]
8. Hampe, N.; Wolterink, J.M.; Van Velzen, S.G.M.; Leiner, T.; Išgum, I. Machine Learning for Assessment of Coronary Artery Disease in Cardiac CT: A Survey. *Front. Cardiovasc. Med.* **2019**, *6*, 172. [CrossRef]
9. Xu, X.; Lin, J.; Tao, Y.; Wang, X. An Improved DenseNet Method Based on Transfer Learning for Fundus Medical Images. In Proceedings of the 7th International Conference on Digital Home (ICDH), Guilin, China, 30 November–1 December 2018; pp. 137–140.
10. Zhang, J.; Wu, C.; Yu, X.; Lei, X. A Novel DenseNet Generative Adversarial Network for Heterogenous Low-Light Image Enhancement. *Front. Neurorobotics* **2021**, *15*, 700011. [CrossRef]
11. Wang, Z.Q.; Zhou, Y.J.; Zhao, Y.X.; Shi, D.M.; Liu, Y.Y.; Liu, W.; Liu, X.L.; Li, Y.P. Diagnostic accuracy of a deep learning approach to calculate FFR from coronary CT angiography. *J. Geriatr. Cardiol.* **2019**, *16*, 42–48.
12. Zreik, M.; van Hamersvelt, R.W.; Khalili, N.; Wolterink, J.M.; Voskuil, M.; Viergever, M.A. Deep learning analysis of coronary arteries in cardiac CT angiography for detection of patients requiring invasive coronary angiography. *IEEE Trans. Med. Imaging* **2019**, *39*, 1545–1557. [CrossRef]
13. Abdar, M.; Książek, W.; Acharya, U.R.; Tan, R.; Makarenkov, V.; Pławiak, P. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Comput. Methods Programs Biomed.* **2019**, *179*, 104992. [PubMed]
14. Huang, W.; Huang, L.; Lin, Z.; Huang, S.; Chi, Y.; Zhou, J.; Zhang, J.; Tan, R.S.; Zhong, L. Coronary artery segmentation by deep learning neural networks on computed tomographic coronary angiographic images. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, Hawaii, 18–21 July 2018; pp. 608–611.
15. Tatsugami, F.; Higaki, T.; Nakamura, Y.; Yu, Z.; Zhou, J.; Lu, Y.; Fujioka, C.; Kitagawa, T.; Kihara, Y.; Iida, M.; et al. Deep learning–based image restoration algorithm for coronary CT angiography. *Eur. Radiol.* **2019**, *29*, 5322–5329. [CrossRef] [PubMed]
16. Yang, S.; Kweon, J.; Roh, J.H.; Lee, J.H.; Kang, H.; Park, L.J.; Kim, D.J.; Yang, H.; Hur, J.; Kang, D.Y.; et al. Deep learning segmentation of major vessels in X-ray coronary angiography. *Sci. Rep.* **2019**, *9*, 16897. [PubMed]
17. Cardiovascular Diseases. 2021. Available online: https://www.who.int/health-topics/cardiovascular-diseases (accessed on 1 November 2021).
18. Banerjee, R.; Ghose, A.; Mandana, K.M. A hybrid CNN-LSTM architecture for detection of coronary artery disease from ECG. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
19. Zreik, M.; van Hamersvelt, R.W.; Wolterink, J.M.; Leiner, T.; Viergever, M.A.; Isgum, I. A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography. *IEEE Trans. Med. Imaging* **2018**, *38*, 1588–1598.
20. Wolterink, J.M.; van Hamersvelt, R.W.; Viergever, M.A.; Leiner, T.; Išgum, I. Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier. *Med. Image Anal.* **2019**, *51*, 46–60.
21. Papandrianos, N.; Papageorgiou, E. Automatic Diagnosis of Coronary Artery Disease in SPECT Myocardial Perfusion Imaging Employing Deep Learning. *Appl. Sci.* **2021**, *11*, 6362. [CrossRef]
22. Khan Mamun, M.M.R.; Alouani, A. FA-1D-CNN Implementation to Improve Diagnosis of Heart Disease Risk Level. In Proceedings of the 6th World Congress on Engineering and Computer Systems and Sciences, Virtual Conference, 13–15 August 2020; pp. 122-1–122-9.
23. Sharma, M.; Acharya, U.R. A new method to identify coronary artery disease with ECG signals and time-Frequency concentrated antisymmetric biorthogonal wavelet filter bank. *Pattern Recognit. Lett.* **2019**, *125*, 235–240. [CrossRef]
24. Alizadehsani, R.; Abdar, M.; Roshanzamir, M.; Khosravi, A.; Kebria, P.M.; Khozeimeh, F.; Nahavandi, S.; Sarrafzadegan, N.; Acharya, U.R. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Comput. Biol. Med.* **2019**, *111*, 103346. [CrossRef]
25. Gülsün, M.A.; Funka-Lea, G.; Sharma, P.; Rapaka, S.; Zheng, Y. Coronary centerline extraction via optimal flow paths and CNN path pruning. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; pp. 317–325.
26. Liu, X.; Mo, X.; Zhang, H.; Yang, G.; Shi, C.; Hau, W.K. A 2-year investigation of the impact of the computed tomography–derived fractional flow reserve calculated using a deep learning algorithm on routine decision-making for coronary artery disease management. *Eur. Radiol.* **2021**, *31*, 7039–7046. [CrossRef]
27. Nishi, T.; Yamashita, R.; Imura, S.; Tateishi, K.; Kitahara, H.; Kobayashi, Y.; Yock, P.G.; Fitzgerald, P.J.; Honda, Y. Deep learning-based intravascular ultrasound segmentation for the assessment of coronary artery disease. *Int. J. Cardiol.* **2021**, *333*, 55–59. [CrossRef]
28. Liu, C.Y.; Tang, C.X.; Zhang, X.L.; Chen, S.; Xie, Y.; Zhang, X.Y.; Qiao, H.Y.; Zhou, C.S.; Xu, P.P.; Lu, M.J.; et al. Deep learning powered coronary CT angiography for detecting obstructive coronary artery disease: The effect of reader experience, calcification and image quality. *Eur. J. Radiol.* **2021**, *142*, 109835. [PubMed]
29. Lin, A.; Kolossváry, M.; Motwani, M.; Išgum, I.; Maurovich-Horvat, P.; Slomka, P.J.; Dey, D. Artificial Intelligence in Cardiovascular Imaging for Risk Stratification in Coronary Artery Disease. *Radiol. Cardiothorac. Imaging* **2021**, *3*, e200512. [CrossRef] [PubMed]
30. Cho, H.; Kang, S.; Min, H.; Lee, J.; Kim, W.; Kang, S.H.; Kang, D.; Lee, P.H.; Ahn, J.; Park, D.; et al. Intravascular ultrasound-based deep learning for plaque characterization in coronary artery disease. *Atherosclerosis* **2021**, *324*, 69–75. [PubMed]

31. Alizadehsani, R.; Khosravi, A.; Roshanzamir, M.; Abdar, M.; Sarrafzadegan, N.; Shafie, D.; Khozeimeh, F.; Shoeibi, A.; Nahavandi, S.; Panahiazar, M.; et al. Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020. *Comput. Biol. Med.* **2021**, *128*, 104095. [PubMed]
32. Rim, T.H.; Lee, C.J.; Tham, Y.; Cheung, N.; Yu, M.; Lee, G.; Kim, Y.; Ting, D.S.W.; Chong, C.C.Y.; Choi, Y.S.; et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *Lancet Digit. Health* **2021**, *3*, e306–e316. [CrossRef]
33. Morris, S.A.; Lopez, K.N. Deep learning for detecting congenital heart disease in the fetus. *Nat. Med.* **2021**, *27*, 764–765.
34. Cheung, W.K.; Bell, R.; Nair, A.; Menezes, L.J.; Patel, R.; Wan, S.; Chou, K.; Chen, J. A computationally efficient approach to segmentation of the aorta and coronary arteries using deep learning. *IEEE Access* **2021**, *9*, 108873–108888.
35. Li, G.; Wang, H.; Zhang, M.; Tupin, S.; Qiao, A.; Liu, Y.; Ohta, M.; Anzai, H. Prediction of 3D Cardiovascular hemodynamics before and after coronary artery bypass surgery via deep learning. *Commun. Biol.* **2021**, *4*, 99.
36. Krittanawong, C.; Virk, H.U.H.; Kumar, A.; Aydar, M.; Wang, Z.; Stewart, M.P.; Halperin, J.L. Machine learning and deep learning to predict mortality in patients with spontaneous coronary artery dissection. *Sci. Rep.* **2021**, *11*, 1–10. [CrossRef]
37. Doppala, B.P.; Bhattacharyya, D.; Janarthanan, M.; Baik, N. A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using ensemble Techniques. *J. Heal. Eng.* **2022**, *2022*, 2585235.
38. Ali, M.M.; Paul, B.K.; Ahmed, K.; Bui, F.M.; Quinn, J.M.W.; Moni, M.A. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Comput. Biol. Med.* **2021**, *136*, 104672. [CrossRef] [PubMed]
39. Khanna, A.; Selvaraj, P.; Gupta, D.; Sheikh, T.H.; Pareek, P.K.; Shankar, V. Internet of things and deep learning enabled healthcare disease diagnosis using biomedical electrocardiogram signals. *Expert Syst.* **2021**, e12864. [CrossRef]
40. Yan, J.; Zhang, B.; Zhou, M.; Kwok, H.F.; Siu, S.W.I. Multi-Branch-CNN: Classification of ion channel interacting peptides using multi-branch convolutional neural network. *Comput. Biol. Med.* **2022**, *147*, 105717. [CrossRef] [PubMed]

diagnostics

MDPI

Article

Deep Transfer Learning for the Multilabel Classification of Chest X-ray Images

Guan-Hua Huang [1,*], Qi-Jia Fu [1], Ming-Zhang Gu [1], Nan-Han Lu [2,3,4], Kuo-Ying Liu [5] and Tai-Been Chen [1,4]

[1] Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan; qijia444@gmail.com (Q.-J.F.); eric956412@gmail.com (M.-Z.G.); ctb@isu.edu.tw (T.-B.C.)
[2] Department of Pharmacy, Tajen University, Pingtung City 90741, Taiwan; ed103911@edah.org.tw
[3] Department of Radiology, E-Da Hospital, I-Shou University, Kaohsiung City 82445, Taiwan
[4] Department of Medical Imaging and Radiological Science, I-Shou University, Kaohsiung City 82445, Taiwan
[5] Department of Radiology, E-Da Cancer Hospital, I-Shou University, Kaohsiung City 82445, Taiwan; ed102500@edah.org.tw
* Correspondence: ghuang@nycu.edu.tw; Tel.: +886-3-513-1334

Abstract: Chest X-ray (CXR) is widely used to diagnose conditions affecting the chest, its contents, and its nearby structures. In this study, we used a private data set containing 1630 CXR images with disease labels; most of the images were disease-free, but the others contained multiple sites of abnormalities. Here, we used deep convolutional neural network (CNN) models to extract feature representations and to identify possible diseases in these images. We also used transfer learning combined with large open-source image data sets to resolve the problems of insufficient training data and optimize the classification model. The effects of different approaches of reusing pretrained weights (model finetuning and layer transfer), source data sets of different sizes and similarity levels to the target data (ImageNet, ChestX-ray, and CheXpert), methods integrating source data sets into transfer learning (initiating, concatenating, and co-training), and backbone CNN models (ResNet50 and DenseNet121) on transfer learning were also assessed. The results demonstrated that transfer learning applied with the model finetuning approach typically afforded better prediction models. When only one source data set was adopted, ChestX-ray performed better than CheXpert; however, after ImageNet initials were attached, CheXpert performed better. ResNet50 performed better in initiating transfer learning, whereas DenseNet121 performed better in concatenating and co-training transfer learning. Transfer learning with multiple source data sets was preferable to that with a source data set. Overall, transfer learning can further enhance prediction capabilities and reduce computing costs for CXR images.

Keywords: convolutional neural network; deep learning; source data set; supervised classification

Citation: Huang, G.-H.; Fu, Q.-J.; Gu, M.-Z.; Lu, N.-H.; Liu, K.-Y.; Chen, T.-B. Deep Transfer Learning for the Multilabel Classification of Chest X-ray Images. *Diagnostics* **2022**, *12*, 1457. https://doi.org/10.3390/diagnostics12061457

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 25 May 2022
Accepted: 10 June 2022
Published: 13 June 2022

1. Introduction

A chest X-ray (CXR), which is generated by exposing the chest to a small dose of ionizing radiation, is a projection radiograph of the chest used for imaging subtle lesions and the density of human tissues. It is commonly used for visualizing the condition of the thoracic cage, chest cavity, lung tissue, mediastinum, and heart. It can thus facilitate the diagnosis of common thorax diseases, including aortic sclerosis or calcification, arterial curvature, abnormal lung fields, anomalous lung patterns, spinal lesions, intercostal pleural thickening, and cardiac hypertrophy.

Computer vision technology and hardware computing capabilities have progressed considerably. Considering the overload of medical resources and the high demand for medical image analysis, the development of computer-aided diagnosis systems with a high diagnostic efficiency and accuracy is warranted. CXR images containing large amounts of physiological data can aid data-hungry deep learning paradigms in the construction of valuable intelligent auxiliary systems. Deep learning is a branch of machine learning;

through linear or non-linear transform from multiple layers, deep learning can automatically extract sufficient and representative features from a data set. In traditional machine learning, features are usually extracted using handcrafted rules, which are created by relevant domain experts. After the data characteristics are understood, useful and effective features can be produced. However, the ability to automatically extract features from deep learning can reduce the time spent by experts in feature engineering. Therefore, deep learning may afford an excellent performance in applications where machines may have failed in the past.

Several studies have used deep learning for CXR analysis, particularly for image classification. Most of these studies have trained deep learning models by using well-designed convolutional neural network (CNN) architectures such as VGG [1], GoogleNet [2], ResNet [3], and DenseNet [4]. CNN architecture depth, data augmentation, input preprocessing methods, image size, and pretraining schemes can affect the performance of a deep learning model [5]. No standardized design methodology for improving deep learning model performance has been reported thus far. Most of the relevant studies have focused on comparing the performance of multiple design methodologies for a specific task, rather than reporting novel methodologies [6–8]. Some studies have achieved methodological novelty by utilizing methods that can aid in improving model performance. For example, Hu et al. [9] and Wu et al. [10] used an extreme learning machine (ELM) to replace the conventional fully connected layer in a deep CNN for real-time analysis and applied a Chimp optimization algorithm or sine-cosine algorithm to ameliorate the ELM's ill-conditioning and nonoptimal problems. Wang et al. [11] trained CNN models by using the whale optimization algorithm, which can resolve difficulties related to requiring a considerable amount of manual parameter tuning and parallelizing the training process of traditional gradient descent-based approaches. Khishe et al. [12] proposed an efficient biogeography-based optimization approach for automatically finetuning model hyperparameters (e.g., number of output channels, convolution kernel size, layer type, training algorithm's learning rate, epoch, and batch size), which are typically selected manually. CXR images are commonly classified as normal or abnormal in the literature [7]. Although CXR can be used to detect multiple diseases of the thorax, few methods have been proposed for classifying multiple disease labels [13].

Applying deep learning methods to CXR image analysis may have promising applications. However, the advancement of automatic image analysis in hindered by several underlying limitations. The main limitation is the lack of large-scale CXR datasets. Although the number of parameters required for deep learning models is considerably large, CXR image training data are limited; this can cause model overfitting. Compared with ordinary images, collecting and labeling CXR images can be difficult and cost intensive. CXR images generated using various instruments with different settings and in different environments cannot be naively analyzed together because this can lead to various errors. Transfer learning [14], which uses the knowledge learned from one task as a starting point for related tasks, can aid in making the best use of different CXR databases. Transfer learning mainly involves using a large amount of open-source data to make up for target data shortages so as to achieve improved performance in model fitting.

In this study, we used one private target data set—1630 chest radiographs provided by the E-Da Hospital, I-Shou University, Taiwan—and three open-source datasets—the ImageNet dataset [15], ChestX-ray dataset [16,17] (with >100,000 chest radiograph images provided by the National Institutes of Health (NIH)), and CheXpert dataset [18] (collected by the Stanford ML Group, comprising nearly 220,000 chest radiographs). The images from the private target data set were labeled by radiology specialists as either being disease-free or containing multiple sites of abnormalities—representing a typical multilabel classification problem. Although ImageNet is sufficiently large for deep learning, most of its 14 million images were dissimilar to those in our private data set; therefore, this data set may not have been able to provide accurate feature representations to classify images in our

private data set. ChestX-ray and CheXpert, although modest in size, are more similar to our private data set, and thus, they may be able to support target data training more efficiently.

Transfer learning has been widely applied in CXR deep learning analysis. Most studies have first trained deep CNN models on the large ImageNet dataset for natural image classification, followed by the use of trained weights for initialization to retrain all layers or only retrain the final (fully connected) layer for target CXR image classification [8]. The performance of transfer learning might be affected by several factors such as the sizes of the source and target data set, similarity between these data sets, retraining of all or partial layers in the target task, and CNN architecture [19]. For natural image classification, Azizpour et al. [19] and Cui et al. [20] have reported best practices on how these factors should be set in generic and domain-specific tasks, respectively. However, only a few studies have focused on the factors that affect the transferability of medical image analysis approaches. Tajbakhsh et al. [21] considered four different medical imaging applications and demonstrated that CNNs pretrained on ImageNet performed better and were more robust to the size of the target training data than the CNNs trained from scratch. Gozes and Greenspan [22] pretrained their model on the ChestX-ray data set and demonstrated that the pretrained weights enabled the model to exhibit improved predictions on small-scale CXR data sets compared with the performance of a model pretrained on ImageNet.

To maximize the performance of transfer learning for CXR image classification, the effects of different transfer learning characteristics in medical image analysis must be systematically investigated. Because of the substantial differences between natural and medical images, we could not apply the knowledge learned for natural images in previous studies [19,20] to our current CXR analysis. Accordingly, our study focused on the multilabel classification of CXR images, which is a crucial topic that warrants research. We thoroughly investigated the aforementioned transferability factors and included the ImageNet, ChestX-ray, and CheXpert data sets as the source data. Previous studies have typically used one type of source data at a time for transfer learning. In this study, we developed new approaches for integrating different source data sets practically to eventually obtain novel powerful source data sets. Our results may aid in devising best practices for the efficient use of different types of data sets to alleviate the insufficiency of training data and enhance the performance of deep learning models in the medical field.

2. Materials

The target data set analyzed in this study contained the CXR images from the E-Da Hospital, I-Shou University, Taiwan. A deep learning model was trained to classify these images as being disease-free or containing multiple sites of abnormalities. We selected three source datasets with different sizes to pretrain our deep learning model and to improve its performance for the target data. Of all our source data sets included here, ImageNet contains the largest amount of data, followed by CheXpert and then ChestX-ray. Although ImageNet is the largest in size, its data had less similarity to the target data than the other two data sets had. Table 1 lists the basic characteristics of the data sets used.

Table 1. Characteristics of the included data sets.

Category	Name	Label Category	Size	Feature
Source data	ImageNet	20,000 +	14 million +	Large and diverse
Source data	CheXpert	14 [1]	224,316	Similar to the target data
Source data	ChestX-ray	14 [2]	112,010 [4]	Similar to the target data
Target data	E-Da chest X-ray	8 [3]	1630	Small but important

[1] Thirteen common thoracic disease labels and a "no finding" label (indicating the absence of any disease). [2] Thirteen disease labels and a normal label. [3] Seven disease labels and a normal label. [4] One hundred and ten images labeled as "hernia" in the original dataset were discarded in this analysis due to the small sample size.

2.1. Target Data Set

The target data set contained CXR images that were collected from patients who received a CXR between January 2008 and December 2018; the images were stored in the DICOM (Digital Imaging and Communications in Medicine) format. These images were retrospectively extracted from the archiving and communication system (PACS) of E-Da Hospital. Patients' gender, age, and diagnostic reports from radiology specialists were also provided. Images were excluded if their quality and the corresponding interpretation of diagnostic reports were unclear. Images of minors (aged < 18 years) were also excluded from this study. This clinical study was approved by the Institutional Review Board of E-Da Hospital. All patients signed written informed consent before participating.

The image size ranged between 1824 and 2688 pixels in length and 1536 and 2680 pixels in width. The image resolution was 0.16 mm per pixel. When we experimentally analyzed images resized at 1024 × 1024, the results were nonsignificant, and the process was cost intensive. Therefore, we resized all images to 512 × 512 pixels for analysis.

We first removed duplicate and outlier images. We also discarded five images uniquely labeled as "heart pacemaker placement" due to their inconsistent disease property. The analyzed data set comprised 1630 images with 1 normal label and 17 diseases labels, which had been integrated into 8 categories (including normal) with guidance from physicians. Of these images, 1485 had a single label and 145 had multiple labels. Table 2 lists the numbers of images that contained certain labels in the data set.

Table 2. Numbers and labels of images in the target data set.

Category	Sample Size	Subcategory	Sample Size
normal	1212	normal	1212
aortic sclerosis/calcification	90	aortic arch atherosclerotic plaque	90
arterial curvature	93	tortuous aorta/thoracic aortic ectasia	93
abnormal lung fields	33	shadows of pulmonary nodules	13
		consolidations or lung cavities in the upper lobes	5
		pulmonary fibrosis	15
increased lung patterns	153	atelectasis/focal consolidation	109
		enlarged hilar shadow	44
spinal lesions	148	degenerative joint disease of the thoracic spine	75
		scoliosis	73
cardiomegaly	41	cardiomegaly	41
intercostal pleural thickening	36	intercostal pleural thickening	36

2.2. Source Data Sets

2.2.1. ImageNet

ImageNet is a large visual database designed for visual object recognition. It contains more than 14 million images that have been hand-labeled to indicate their object category, of which there are >20,000. The data set size is approximately 1 TB. The Python package Keras provides the pretrained weights for various networks, which eliminates the requirement to training them from scratch.

2.2.2. CheXpert

CheXpert is a large CXR image data set collected by the Stanford ML Group; it comprises 224,316 chest radiographs labeled to indicate the presence of 13 common thoracic diseases or labeled "no finding" to indicate the absence of all diseases [18]. Natural language processing (NLP) is used to extract observations from radiology reports, and this extracted information serves as the basis of labeling. The training labels in the data

set for each category are 0 (negative), 1 (positive), or *u* (uncertain). Different approaches for using uncertainty labels during model training may lead to differences in network performance. In this study, we followed the results from the original paper for dealing with the uncertainty label in relation to five diseases: atelectasis, edema, pleural effusion, cardiomegaly, and consolidation. In other words, we reconstructed a five-dimensional label vector for the five aforementioned diseases—where *u* was replaced with 1 for the first three diseases and with 0 for the final two—and then we applied it as the label for five-class multilabel classification. In other words, we transformed the pretraining task into a task for classifying whether the images had these five diseases.

In this study, we resized the images to 512 × 512 pixels to save memory. Moreover, some of the images from the data set were gray-scale images (where they only had one channel), whereas the other images had four channels. To feed them into the available pretrained models through a three-channel input, we replicated the one-channel images three times but removed the fourth channel of the four-channel images.

2.2.3. ChestX-ray

ChestX-ray is an open-source data set compiled by the National Institutes of Health; it comprises 112,120 chest radiographs with 1 normal label and 14 disease labels [16,17]. Similar to CheXpert, ChestX-ray uses NLP for labeling, but it does not have an uncertainty label. We stacked and removed the channels of the images to fit them into the available pretrained networks as we did for CheXpert. We also resized the original images to 512 × 512 pixels for fast processing and excluded the disease label "hernia" (and thus deleted 110 images) from the analysis because its sample size was relatively small.

3. Methods

This analysis aimed to predict common thorax diseases in patients through their CXR images. The images were labeled as being disease-free or containing multiple sites of abnormalities, representing a typical multilabel classification problem. Deep learning architectures were used to automatically extract features and build a classifier, and transfer learning was conducted to combine information from different datasets to improve model performance. We trained deep learning models using the well-designed convolutional neural network (CNN) architectures ResNet50 [3] and DenseNet121 [4]. Implementing transfer learning in the CNN involved reusing the first several layers of the network for classifying images in the open-source data sets ImageNet, CheXpert, and ChestX-ray (the source task). Pretrained weights from these layers served as the starting points for classifying the CXR images from E-Da Hospital (the target task). By combining three source data sets in various manners, we obtained several different sets of pretrained weights. To transfer the pretrained weights to the target task, we either used the pretrained weights as the initial values and retrained the model from the scratch or fixed the weights of some early layers at the corresponding pretrained weights and reconstructed the others.

Image augmentation techniques were applied to artificially create variations in the existing images; this expanded the training data set to represent a comprehensive set of possible images. Our target dataset was imbalanced: the number of images containing each disease label was unequal. Deep learning algorithms can be biased toward the majority labels and fail to detect the minority labels. To prevent imbalanced data, sample weighting in the loss function was used. Model performance was evaluated using various metrics and stratified five-fold cross-validation. Our analytic approach is illustrated in Figure 1. The programming language Python [23] was used to implement these methods.

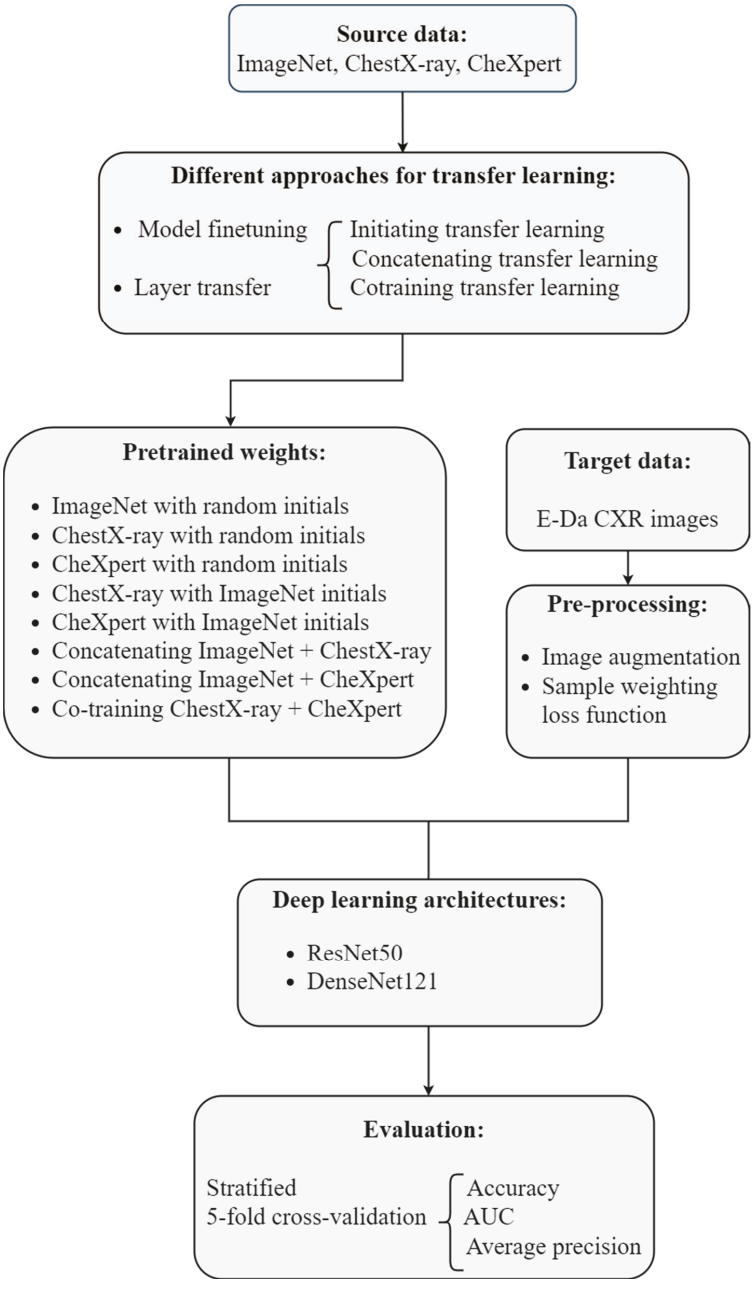


Figure 1. Flow chart of our deep transfer learning approach for the multilabel classification of the chest X-ray images.

3.1. Multilabel Classification

Image classification aims at building a model that maps the input of the ith image $\boldsymbol{x}_i$ to a label vector $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iK})$, where $y_{ik} = 0$ or 1, $k = 1, \cdots, K$ is the indicator for the kth disease class. Multiclass classification involves classifying an image into one of multiple

classes; in other words, the label vector of multiclass classification has only a single element equal to 1. However, in multilabel classification, more than one class may be assigned to the image, with the label vector possibly having 0 or 1 in each element.

CXR images in our target data set were labeled for seven diseases, and multiple diseases were often identified in one image. The target task was multilabel classification and defined a seven-dimensional label vector with an all-zero vector $(0, 0, 0, 0, 0, 0, 0)$ representing a normal status (where none of the seven diseases was detected). We treated multilabel classification as a multiple binary classification problem; thus, the loss function was the sum of multiple binary cross-entropies. However, our data were seriously imbalanced: the elements of the label vectors were almost equal to 0. The number of 0 s was much larger than that of 1 s; this could have misled our model toward predicting 0 s if the usual loss function was used. Therefore, we adjusted the common binary cross-entropy with weights that considered the proportions of 0 s and 1 s in the same sampling batch.

Suppose that the entire training set is divided into J batches, with each batch size being M. Let $\boldsymbol{x}_{jm}$ be the input for the mth image in the jth batch and $\boldsymbol{y}_{jm} = (y_{jm1}, \cdots, y_{jmK})$ be its label vector. Then, the proposed weighted binary cross-entropy (WBCE) loss is defined as:

$$L_{\text{WBCE}} = \sum_{j=1}^{J} \left(\sum_{m=1}^{M} \left\{ \beta_{Pj} \sum_{k:\ y_{jmk}=1} \left[-\ln\left(\sigma\left(f_k\left(\boldsymbol{x}_{jm}\right)\right)\right) \right] + \beta_{Nj} \sum_{k:\ y_{jmk}=0} \left[-\ln\left(1-\sigma\left(f_k\left(\boldsymbol{x}_{jm}\right)\right)\right) \right] \right\} \right), \tag{1}$$

where $f_k\left(\boldsymbol{x}_{jm}\right)$ is $\boldsymbol{x}_{jm}$'s kth input for the final fully-connected layer, $\sigma(z) = e^z/(1+e^z)$ is the sigmoid function, and β_{Pj} is set to $\frac{|P_j|+|N_j|}{|P_j|}$, whereas β_{Nj} is set to $\frac{|P_j|+|N_j|}{|N_j|}$, with $|P_j|$ and $|N_j|$ being the number of 1 s and 0 s in the label vectors of the jth batch, respectively.

3.2. Image Augmentation

Training deep CNNs requires a considerable amount of data, but the sample size of medical images is typically not sufficiently large. Image augmentation is a powerful technique for generating images using a combination of affine transformations—such as shift, zoom in-zoom out, rotate, flip, distort, and shade with a hue—which can feed more images into the neural networks and exploit information in the original data more fully. We augmented our training images by using the ImageDataGenerator function in the Keras Python library. "Online" augmentation was applied; here, we applied the image augmentation techniques in mini-batches and then fed them to the model. The model with online augmentation was presented with different images at each epoch; this aided the model in generalizing, and the model did not need to save the augmented images on the disk, which reduced the computing burden.

3.3. Deep CNNs

Image classification using deep CNNs has an outstanding performance compared with traditional machine learning approaches. CNN is a deep learning method in which a series of layers is constructed to progressively extract higher-level features from the raw input. In a CNN, the typical layers comprise the input layer (which imports image data for model training), the convolution layer (which uses various filters to automatically learn representation of features at different levels), the pooling layer (which selects the most prominent features to reduce the dimension of subsequent layers), and the fully connected layer (which flattens the matrix feature maps into a single vector for label prediction).

The CNN architecture refers to the overall structure of the network: the types of layers it should have, the number of units each layer type should contain, and the manner in which these units should be connected to each other. CNN architectures such as VGG [1], GoogleNet [2], ResNet [3], and DenseNet [4] have been widely used. We included two high-performing CNN models: 50-layer ResNet (ResNet50) and 121-layer DenseNet (DenseNet121). ResNet50 has fewer filters and a lower complexity than VGG nets, and DenseNet121 requires fewer parameters than ResNet50 does.

3.3.1. ResNet

In theory, the deeper the network model, the better its results. Nevertheless, the degradation problem may arise: as the network becomes deeper, the model's accuracy may become saturated or even decrease. This problem is different from overfitting because of increased training errors. ResNet made a historical breakthrough in deep neural networks by solving the degradation problem through residual learning.

In this study, we established a deeper model by stacking new layers on a shallower architecture. Let $\boldsymbol{x}$ denote the output of the shallow part of the model and $H(\boldsymbol{x})$ denote the output of the deeper model. No higher training error should be obtained if the added layers are for identity mapping. Rather than expecting the stack of layers to learn identity mapping, ResNet argues that it is easier to let these layers fit a residual mapping of $F(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}$ to zero and recast the output of the deeper model as $\hat{F}(\boldsymbol{x}) + \boldsymbol{x}$, where $\hat{F}(\boldsymbol{x})$ is the fitted residual. Generally, $\hat{F}(\boldsymbol{x})$ will not be zero; therefore, the stacking layers can still learn new features and demonstrate an improved performance.

The formulation of $\hat{F}(\boldsymbol{x}) + \boldsymbol{x}$ can be realized using feedforward neural networks with a shortcut connection, which skips some of the layers in the neural network and feeds the output of one layer as the input to the next layers. A series of shortcut connections (residual blocks) forms ResNet. This study applied the deeper ResNet50 that contained 50 layers and used a stack of three layers with 1×1, 3×3, and 1×1 convolutions as the building residual block. This three-layer residual block adopted a bottleneck design to improve computational efficiency, where the 1×1 layers were responsible for reducing and then increasing (restoring) the dimensions, leaving the 3×3 layer as a bottleneck with smaller input and output dimensions [3]. In ResNet50, batch normalization (BN) [24] is adopted immediately after each convolution and before ReLU activation, and global average pooling (GAP) [25] is performed to form the final fully connected layer.

3.3.2. DenseNet

As CNNs become increasingly deep, information about the input passes through many layers and can vanish by the time it reaches the end of the network. Different approaches, which vary in network topology and training procedure, create short paths from early layers to later layers to address this problem. DenseNet proposes an architecture that distills this insight into a simple connectivity pattern of dense blocks and transition layers. A dense block is a module containing many layers connected densely with feature maps of the same size. In a dense block, each layer obtains additional inputs from all preceding layers, and it passes on its own feature maps to all the subsequent layers. The transition layer connects two adjacent dense blocks, and it reduces the size of the feature map through pooling. Compared with ResNet that connects layers through element-level addition, layers in DenseNet are connected by concatenating them at the channel level.

In a dense block, the convolution for each layer produces k feature maps, which denote k channels in the output. k is a hyperparameter known as the growth rate of DenseNet, which is usually set to be small. Under the assumption that the initial number of channels is k_0, the number of input channels in the $\updownarrow$th layer is $k_0 + k(\updownarrow - 1)$. As the number of layers $\updownarrow$ increases, the input can be extremely large even if k is small. Because the input of the latter layer grows quickly, the bottleneck design is introduced into the dense block to reduce the burden of calculation, where a 1×1 convolution and then a 3×3 convolution are applied to each layer to generate the output. Similar to ResNet, DenseNet uses a composite of three consecutive operations for each convolution: BN + ReLU + convolution.

3.3.3. Implementation

We used the Python package Keras to implement ResNet50 and DenseNet121. Optimizer Adam [26] with a mini-batch size of 16 and an epoch number of 30 was used. The learning rate started from 0.0001 and was divided by 10 when the validation loss did not decrease in 10 epochs. DenseNet121 set the growth rate to $k = 12$.

3.4. Transfer Learning

In practice, we usually do not have sufficient data to train a deep and complex network such as ResNet or DenseNet. Techniques such as image argumentation are insufficient for resolving this problem. The lack of data can cause overfitting, which can make our trained model overly rely on a particular data set and then fail to fit to additional data. This problem can be resolved using transfer learning [14]. The main concepts of transfer learning are to first train the network with sufficiently large source data and then transfer the structure and weights from this trained network to predict the target data.

In the current study, we used the data from the ImageNet [15], ChestX-ray [16,17], and CheXpert [18] data sets as the source data and the CXR images from the E-Da Hospital as the target data. ImageNet contains more than 14 million natural images and thus is sufficiently large for a deep learning application; however, it does not contain images with similarity to medical images, and therefore, it may fail to provide useful feature representations to classify our target data. ChestX-ray and CheXpert—consisting of about 100,000 and 220,000 CXR images, respectively—are modest in size but are more similar to our target data set.

Implementing transfer learning in the CNN involves reusing the parameter estimators pretrained on the source data when fitting the target data. Model finetuning is a method in which these pretrained parameter estimators are used as the initial values and finetuned to fit the target data. In the layer transfer approach, target data fitting preserves some of the layers from a pretrained model and reconstructs the others. We thus adopted model finetuning or layer transfer to reuse the parameter estimators pretrained on the source data (from ImageNet, ChestX-ray, or CheXpert).

We next tried to combine these source data sets to obtain new powerful source data sets for transfer learning.

3.4.1. Initiating Transfer Learning

In Keras, we can implement transfer learning with initial weights selected randomly or from ImageNet pretraining. ImageNet pretraining may result in robust parameter estimation due to the diversity and richness of the data set, and the pretraining of the similar data sets ChestX-ray and CheXpert may accelerate the convergence of the model. Therefore, we adopted five different pretrained weights in modeling our target data: ImageNet pretraining with randomly selected initials, ChestX-ray pretraining with random or ImageNet pretraining initials, and CheXpert pretraining with random or ImageNet pretraining initials.

3.4.2. Concatenating Transfer Learning

In this approach, we first constructed two ResNet backbone models by using different pretrained weights: one from ImageNet pretraining with random initials (IR) and the other from ChestX-ray pretraining with random initials (CR). Thereafter, we concatenated the outputs of the final convolution layer from two models and then used the GAP to reduce the dimension, and finally applied the fully connected layer to generate the final prediction (Figure 2). The ResNet model concatenating IR and CheXpert pretraining with random initials (XR) was obtained in a similar manner. The DenseNet backbone models concatenating IR and CR and concatenating IR and XR could also be obtained. Models using different pretrained weights may extract different features. In contrast to the first approach aimed at extracting features from a single domain, this approach could expand features from two different domains. However, compared with the first approach, this approach used twice the memory and time to store and upgrade parameters.

3.4.3. Co-Training Transfer Learning

In this approach, we aimed to combine two source data sets, namely ChestX-ray and CheXpert, to a larger CXR data set to serve as source data for medical image classification tasks. ChestX-ray and CheXpert cannot be combined directly: although both the data

sets contain only CXR images, they use different class definitions. To resolve this issue, we applied a co-training approach. First, we fed images from two source data sets to jointly train convolutional layers but to also connect to different fully connected layers to predict their corresponding classes. In other words, they shared the parameters from the convolution layers but not those after the convolution layers. Finally, we reserved these shared convolutional weights as the pretrained weights in target data model fitting. The reason that we did not cotrain ImageNet with ChestX-ray (or CheXpert) was that merging different domain data may mislead the model and reduce its efficiency. The co-training approach is illustrated in Figure 3.

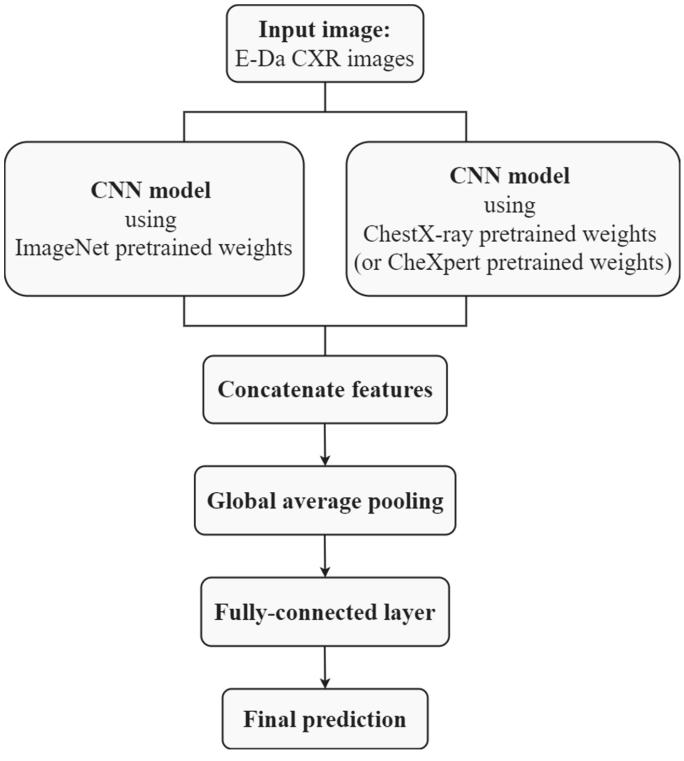


Figure 2. Flow chart of concatenating transfer learning.

3.5. Evaluation

3.5.1. Stratified *K*-Fold Cross-Validation

In *K*-fold cross-validation, first, the data are shuffled to ensure the randomness when the data are split. Second, whole data are split into *K* groups, where *K* is usually set to be 5 or 10 based on the sample size. Third, one of groups is considered "test data," and the others are considered "training data" with a total of *K* different test-training combinations (*K* cross-validation rounds). Finally, the "training data" are further divided into "training" and "validation" sets, which are used to train the model parameters and to instantly evaluate the performance of various hyper-parameters, respectively.

In stratified *K*-fold cross-validation, every group must have the same class distribution. This method has great applicability for multiclass classification, where each sample belongs to only one class. In a multilabel task, every sample may have multiple class categories. To appropriately perform stratified *K*-fold cross-validation for a multilabel classification task, performing iterative stratification is essential [27]; it can be implemented using Python's scikit-learn-compatible package MultilabelStratifiedKFold.

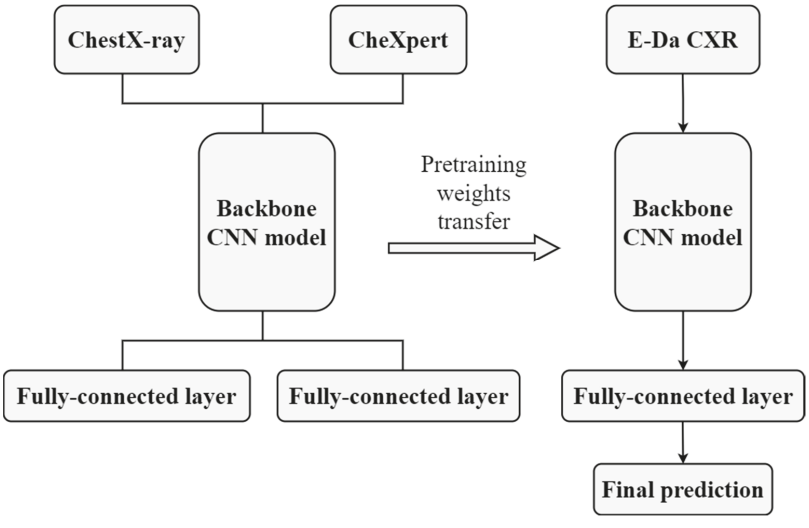


Figure 3. Flow chart of co-training transfer learning.

3.5.2. Metrics

For an imbalanced data set, good accuracy does not indicate that a classifier is reliable. A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier (with vs. without the label) at different thresholds, where the x-axis represents the false positive rate, and the y-axis represents the true positive rate. The area under the ROC curve (AUC) is used to summarize the ROC curve. A precision (PR) curve is another type of plot that evaluates the performance of a model, with the x-axis being the recall and the y-axis being the precision. The area under the PR curve is the average precision (AP). Moreover, training accuracy represents the average accuracy of the training sets during the five cross-validation rounds. Test accuracy, AUC, and AP refer to the accuracy, AUC, and AP of the five test groups, respectively. In multilabel classification, we can treat the problem as a combination of multiple binary classifications; thus, we can evaluate every label individually with binary metrics and obtain the average of these binary metrics (i.e., the mean metric). In this study, we calculated the training accuracy, test accuracy, test AUC, and test AP for each individual label and the mean training accuracy, test accuracy, test AUC, and test AP for all labels.

4. Results

This section presents the results of all the experiments performed in the current study. Table 3 presents a summary of our experimental configurations and parameter settings.

4.1. Layer Transfer versus Model Finetuning

To using the layer transfer approach, we fixed some lower layers in the network at weights from the source data and retrained the remaining layers on the target data. We first froze 10, 22, 40, and 49 layers in ResNet50, which corresponded to the end of the first, second, third, and fourth residual blocks. We used weights pretrained on ImageNet with random initials, ChestX-ray with random initials, and ChestX-ray with ImageNet pretraining initials.

The mean test accuracies and AUCs from the layer transfer or model finetuning on different pretrained weights are listed in Table 4. When adopting the layer transfer approach, it is not recommanded to modify only the final layer, which is equivalent to freezing the first 49 layers in ResNet50. Here, for pretrained weights from ChestX-ray with random or ImageNet pretraining initials, freezing more layers typically led to a significantly

higher accuracy but a somewhat worse AUC. For the ImageNet pretrained weight, freezing more layers resulted in a lower accuracy and AUC, although the difference was negligible. These results indicated that the ideal number of frozen layers depends on the similarity between the source and target data. When the target data are distinct from the source data, more layers may need to be unfrozen.

Table 3. Summary of experimental configurations and parameter settings.

Configuration	Setting
Target data set	CXR images from the E-Da Hospital, Taiwan
Target task	Multilabel classification of CXR images
Loss function	Weighted binary cross entropy
Image augmentation	The ImageDataGenerator function in the Keras Python library for online augmentation
Deep CNN modeling	
Backbone architecture	The Python package Keras to implement 50-layer ResNet and 121-layer DenseNet
Optimizer	Adam
Mini-batch size	16
Epoch number	30
Learning rate	Start from 0.0001 and is divided by 10 when the validation loss does not decrease in 10 epochs
Growth rate in DenseNet	12
Transfer learning	
Source data set	ImageNet, ChestX-ray, and CheXpert
Reuse pretrained weights	Model finetuning and layer transfer
Combine source data sets	Initiating, concatenating, and co-training
Evaluation	
Data splitting	The Python package MultilabelStratifiedKFold for stratified *K*-fold cross-validation
Metrics	Accuracy, AUC, and AP

Table 4. Mean test accuracies and AUCs of different approaches of reusing pretrained weights in ResNet50 [1].

	Pretrained Weight							
	ImageNet with Random Initials		ChestX-ray with Random Initials		ChestX-ray with ImageNet Initials		Random Initials [2]	
Method	**AC** [4]	**AUC**	**AC**	**AUC**	**AC**	**AUC**	**AC**	**AUC**
Layer transfer [3]							0.775	0.765
RB_1	**0.841**	0.487	0.557	**0.519**	0.461	**0.512**		
RB_2	0.828	0.488	0.748	0.514	0.749	0.511		
RB_3	0.814	0.469	**0.871**	0.496	**0.851**	0.505		
RB_4	0.771	**0.520**	0.512	0.498	0.646	0.499		
Model finetuning	**0.873**	**0.796**	**0.811**	**0.831**	**0.789**	**0.827**		

[1] Bold numbers indicate the top two approaches in each metric. [2] The model without transfer learning. [3] RB_1 = Pretrained weights on the first 10 layers, which correspond to the end of the first residual block; RB_2 = Pretrained weights on the first 22 layers, which correspond to the end of the second residual block; RB_3 = Pretrained weights on the first 40 layers, which correspond to the end of the third residual block; RB_4 = Pretrained weights on the first 49 layers, which correspond to the end of the fourth residual block. [4] AC = Accuracy.

Compared with the layer transfer approach, model finetuning afforded a higher accuracy when the ImageNet pretrained weight was set directly and provided considerably larger AUCs for all types of pretrained weights. The model finetuning approach appeared to an attractive option for focusing on prediction accuracy related to every disease label without to sacrificing the minority label.

4.2. Effects of Transfer Learning

After transfer learning was applied with model finetuning and pretrained weights from ImageNet, ResNet50′s mean test accuracy increased by 11% and the mean test AUC increased by 4% compared with when transfer learning was not applied (Table 4). After

this transfer learning approach was used to train DenseNet121, the mean test accuracy and mean test AUC were 0.799 and 0.803, respectively. However they decreased to 0.657 and 0.736, respectively, after DenseNet121 was trained without transfer learning. With transfer learning, DensNet121's accuracy increased by 18% and the AUC increased by 8%. Model finetuning-based transfer learning could improve the model fit and appeared to have a greater influence on models with a more complex structure (i.e., those with a greater network depth). These performance improvements remained when model finetuning with pretrained weights from ChestX-ray (with random or ImageNet pretraining initials) was applied.

The layer transfer approach did not always improve model performance (Table 4). Even with the best selected number of frozen layers, models with transfer learning had worse AUCs than those without transfer learning.

4.3. Comparison of Various Transfer Learning Approaches

The results from Sections 4.1 and 4.2 demonstrated that compared with layer transfer, model finetuning—re-training the whole model with pretrained weights as initial values—provided a better prediction model. We thus adopted model finetuning for subsequent analyses.

This section presents the results from initiating, concatenating, and co-training transfer learning methods, which use different approaches to combine several source data sets to provide an improved model performance. This performance, evaluated using mean accuracy, AUC, and AP on training and test data, is presented in Table 5. In the subsequent subsections, we compare this performance from the perspective of backbone models, source data sets, and combined methods.

Table 5. Training and test performance of various transfer learning approaches [1].

Backbone Model	Transfer Learning Approach [2]	Mean Training Accuracy	Mean Test AUC	Mean Test AP
	Initiating transfer learning			
ResNet50	IR	**0.935**	0.796	**0.214**
ResNet50	CR	0.879	**0.831**	0.209
ResNet50	CI	0.868	**0.827**	0.206
ResNet50	XR	0.819	0.806	0.191
ResNet50	XI	0.852	**0.831**	**0.213**
DenseNet121	IR	0.916	0.803	0.204
DenseNet121	CR	0.807	0.800	0.171
DenseNet121	CI	0.784	0.779	0.179
DenseNet121	XR	0.799	0.781	0.169
DenseNet121	XI	0.864	0.826	**0.221**
	Concatenating transfer learning			
ResNet50	I + C	**0.935**	0.780	0.207
ResNet50	I + X	0.930	0.776	0.190
DenseNet121	I + C	**0.935**	0.802	0.210
DenseNet121	I + X	0.914	0.813	0.210
	Co-training transfer learning			
ResNet50	C ∪ X	0.855	0.790	0.191
DenseNet121	C ∪ X	0.775	0.826	0.210

[1] Bold numbers indicate the top three approaches in each metric. [2] IR = ImageNet pretraining with random initials, CR = ChestX-ray pretraining with random initials, CI = ChestX-ray pretraining with ImageNet pretraining initials, XR = CheXpert pretraining with random initials, XI = CheXpert pretraining with ImageNet pretraining initials, I + C = Concatenating ImageNet + ChestX-ray, I + X = Concatenating ImageNet + CheXpert, C ∪ X = Co-training ChestX-ray + CheXpert.

4.3.1. Backbone Model Comparison

In this study, we used ResNet50 and DenseNet121 as the backbone model for transfer learning. Compared with DenseNet121, ResNet50 required less time to train a model, but it had more parameters to save. To summarize the two backbone models' performance levels,

the radar plots for mean test AUCs and APs from various transfer learning approaches were created (Figures S1 and S2). In initiating transfer learning, ResNet50 outperformed DenseNet121. By contrast, DenseNet121 performed better in concatenating and co-training transfer learning. Therefore, a model with a relatively complex structure may provide more accurate results if a larger, more diverse source data set created by combining different data sets is applied.

4.3.2. Source Data Comparison

In this study, we used data from three sources—ImageNet, ChestX-ray, and CheXpert—to perform transfer learning. Each has its own strengths and limitations. Although it is nearly 100 times the size of other two data sets, ImageNet's data had the weakest association with our target data. Although CheXpert is twice the size of ChestX-ray, it contains uncertainty labels, which might increase the difficulty of the training process. ChestX-ray is the smallest data set among those included in this study; nevertheless, it demonstrated the stronger connection to our target data and the most precise labeling. To compare the performance of the different source data, we collected the results from ResNet50 with initial weights from IR, CR, and XR. Their PR and ROC curves for each disease label on test data are presented in Figures S3–S5. DenseNet121's corresponding PR and ROC curves are presented in Figures S6–S8.

For both the backbone models, the use of ChestX-ray as source data led to a better performance than using CheXpert (Table 5). Even though ChestX-ray is only half the size of CheXpert, its accurate labeling made up for the lack of data. Although ImageNet had the best performance in the training process, its performance in the test process was worse than that of ChestX-ray when it was fit in the ResNet50 model (Table 5). Therefore, the use of ImageNet as the source data possibly led to overfitting. In general, the ChestX-ray data had the best performance when only one single-source data set was adopted as in previous studies.

4.3.3. Comparison of Combined Methods

We used three methods to combine source data sets. The initiating transfer learning approach adopted ImageNet initials when training pretrained weights from ChestX-ray or CheXpert data to ensure a robust estimation of these pretrained weights. Concatenating transfer learning aimed at collecting the features obtained using distinct source data sets (e.g., ImageNet and ChestX-ray) to expand the features' coverage. Finally, co-training transfer learning combined two similar source data sets (e.g., ChestX-ray and CheXpert) to train pretrained weights, with the assumption that a large data size can improve model performance.

The combining methods' PR and ROC curves for an individual disease label on test data are shown in Figures S9–S18. In summary, first, for both backbone models, using ChestX-ray with random initials as the source data led to a better performance than using ChestX-ray with ImageNet initials, and using CheXpert with ImageNet initials as source data led to a better performance than using CheXpert with random initials. ImageNet initials ensured a robust estimation of pretrained weights in CheXpert but not in ChestX-ray. Second, concatenating transfer learning provided the highest training accuracy, but it did not achieve the best performance in the test process, possibly because of the overloading of the model with too many parameters (i.e., with twice the number of parameters), causing overfitting. Third, initiating transfer learning was the most suitable for ResNet50, whereas concatenating and co-training transfer learning were the most suitable for DenseNet121. Notably, under-fitting may have arisen for co-training transfer learning in DenseNet121: more epochs in training can overcome this issue. Fourth, DenseNet121 that involved transfer learning with various source data sets performed better than that with a single-source data set. ResNet50 provided similar results; however, the effect was not as significant as that in DenseNet121.

5. Discussion

When reusing pretrained weights in transfer learning, the approach that re-trains the whole model with pretrained weights as initial values (i.e., the model finetuning approach) typically afforded excellent results but required many more computational resources compared with the other approaches. The layer transfer approach, which freezes some layers on pretrained weights, demonstrated advantages over the model finetuning approach by allowing larger batch sizes and requiring a shorter run time and less GPU/CPU memory. For building the most cost-effective model by using the layer transfer approach, the number of frozen layers should be determined accurately. Our results demonstrated that the higher the similarity between the source and target data, the larger the allowed number of frozen layers should be. However, none of the included approaches were universally applicable; selecting the most appropriate approach would require the assessment of its benefits and costs on the basis of specific goals and available resources.

When adopting only one single-source dataset, ChestX-ray demonstrated a better performance than CheXpert. ChestX-ray is only half the size of CheXpert; nonetheless, its accurate labeling was found to make up for its smaller size. However, after ImageNet initials were attached, CheXpert outperformed ChestX-ray. As such, ImageNet appeared to enhance the data volume and variety of these data sets.

ResNet50 was suitable for initiating transfer learning, whereas DenseNet121 performed better in concatenating and co-training transfer learning. Transfer learning combined with various source data sets was also preferable with the use of a single-source data set; however, DenseNet121 led to greater benefits than ResNet50 did. Compound weights from several source data sets may be superior to single weights because they contain additional information offered by another data set. Nevertheless, a more complex transfer process may produce more noise. Compared with ResNet50, DenseNet121 was more complex, and its dense block mechanism could process more data and absorb more information; consequently, DenseNet121 is more suitable than BesNet50 for integrating source data sets.

Few studies have focused on the factors that affect the performance of transfer learning in medical image analysis. The strength of the present study lies in its systematic approach to investigating the transferability of CXR image analytic approaches. Nevertheless, this study had several limitations that should be resolved in future studies. First, our experiments were based on the 50-layer ResNet and 121-layer DenseNet architectures, and the derived weights were estimated using the gradient descent-based optimizer Adam [26] under the consideration of manually selected hyperparameter values. ResNet and DenseNet are both widely used for conducting deep learning analyses on CXR images; nevertheless, they differ in several aspects, and such differences can be leveraged to investigate the impact of the CNN architecture on transfer learning. Alternatively, new architectures—such as EfficientNet [28] and CoAtNet [29], which have shown a high performance in challenging computer vision tasks—could be used for analysis. Moreover, to enhance the efficiency of model parameter estimation, Adam may be replaced with recent optimizers such as Chimp [9] and Whale [11], and biogeography-based optimization can be applied to automatically finetune model hyperparameters [12]. Second, CXR images can be taken in posteroanterior, anteroposterior, and lateral views. In this study, we included only one target data set in which all CXR images were in the posteroanterior view; this may have limited the real-world applicability of our findings. Moreover, we included only image classification as the target task. However, CXR images can be used for several other types of deep learning tasks such as segmentation, localization, and image generation [5]. Accordingly, future studies could use CXR images taken in different positions and could consider a wide range of deep learning tasks.

6. Conclusions

In this study, we conducted a thorough investigation of the effects of various transfer learning approaches on deep CNN models for the multilabel classification of CXR images.

Our target data set, collected through general clinical pipelines, contained 1630 chest radiographs with 17 clinically common disease labels. Transfer learning methods that reused pretrained weights through model finetuning and layer transfer were examined. We considered three source data sets with different sizes and different levels of similarity to our target data and assessed their effect on transfer learning effectiveness. These source data sets could be incorporated into transfer learning individually or in combination. We also proposed initiating, concatenating, or co-training different source data sets for joint transfer learning and used two backbone CNN models with different network architectures to adopt the aforementioned transfer learning approaches.

Several substantial findings were obtained. The results demonstrated that transfer learning could improve the model fit. Compared with the layer transfer approach, the model finetuning approach typically afforded better prediction models. When only one single-source data set was adopted as in previous studies, ChestX-ray outperformed ImageNet and CheXpert. However, CheXpert with ImageNet initials attached performed better than ChestX-ray with ImageNet initials attached. ResNet50 performed better in initiating transfer learning, whereas DenseNet121 performed better in concatenating and co-training transfer learning. Transfer learning with multiple source data sets was preferable to that with a single-source data set.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics12061457/s1, Figure S1: Radar plots for mean test AUCs from various transfer learning approaches; Figure S2: Radar plots for mean test APs from various transfer learning approaches; Figure S3: PR and ROC curves for each disease label in test data with initial weights from ImageNet pretraining with random initials in ResNet50; Figure S4: PR and ROC curves for each disease label in test data with initial weights from ChestX-ray pretraining with random initials in ResNet50; Figure S5: PR and ROC curves for each disease label in test data with initial weights from CheXpert pretraining with random initials in ResNet50; Figure S6: PR and ROC curves for each disease label in test data with initial weights from ImageNet pretraining with random initials in DenseNet121; Figure S7: PR and ROC curves for each disease label in test data with initial weights from ChestX-ray pretraining with random initials in DenseNet121; Figure S8: PR and ROC curves for each disease label in test data with initial weights from CheXpert pretraining with random initials in DenseNet121; Figure S9: PR and ROC curves for each disease label in test data for initiating transfer learning with pretrained weights from ChestX-ray adopting ImageNet initials in ResNet50; Figure S10: PR and ROC curves for each disease label in test data for initiating transfer learning with pretrained weights from CheXpert adopting ImageNet initials in ResNet50; Figure S11: PR and ROC curves for each disease label in test data for initiating transfer learning with pretrained weights from ChestX-ray adopting ImageNet initials in DenseNet121; Figure S12: PR and ROC curves for each disease label in test data for initiating transfer learning with pretrained weights from CheXpert adopting ImageNet initials in DenseNet121; Figure S13: PR and ROC curves for each disease label in test data for concatenating transfer learning of ImageNet + ChestX-ray in ResNet50; Figure S14: PR and ROC curves for each disease label in test data for concatenating transfer learning of ImageNet + CheXpert in ResNet50; Figure S15: PR and ROC curves for each disease label in test data for concatenating transfer learning of ImageNet + ChestX-ray in DenseNet121; Figure S16: PR and ROC curves for each disease label in test data for concatenating transfer learning of ImageNet + CheXpert in DenseNet121; Figure S17: PR and ROC curves for each disease label in test data for co-training transfer learning of ChestX-ray + CheXpert in ResNet50; Figure S18: PR and ROC curves for each disease label in test data for co-training transfer learning of ChestX-ray + CheXpert in DenseNet121.

Author Contributions: Conceptualization, G.-H.H. and T.-B.C.; Data curation, M.-Z.G., N.-H.L., K.-Y.L. and T.-B.C.; Formal analysis, G.-H.H., Q.-J.F. and M.-Z.G.; Funding acquisition, G.-H.H.; Investigation, N.-H.L., K.-Y.L. and T.-B.C.; Methodology, G.-H.H., Q.-J.F. and M.-Z.G.; Project administration, G.-H.H.; Resources, T.-B.C.; Software, Q.-J.F. and M.-Z.G.; Supervision, G.-H.H.; Writing—original draft, G.-H.H., Q.-J.F. and M.-Z.G.; Writing—review and editing, G.-H.H., N.-H.L., K.-Y.L. and T.-B.C. 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 107-2118-M-009-005-MY2, and MOST 109-2118-M-009-004-MY2).

Institutional Review Board Statement: The study was conducted in accordance with the guidelines of the Declaration of Helsinki. All experimental procedures were approved by the Institutional Review Board of the E-Da Hospital, Kaohsiung, Taiwan (approval number EMRP-108-115).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the target data set of this study.

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

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

References

1. Simonyan, K.; Zisserman, A. Very deep convolutional networks for largescale image recognition. *arXiv* **2014**, arXiv:1409.1556.
2. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. *arXiv* **2014**, arXiv:1409.4842.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385.
4. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. *arXiv* **2018**, arXiv:1608.06993.
5. 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]
6. Bressem, K.K.; Adams, L.C.; Erxleben, C.; Hamm, B.; Niehues, S.M.; Vahldiek, J.L. Comparing different deep learning architectures for classification of chest radiographs. *Sci. Rep.* **2020**, *10*, 12590.
7. Tang, Y.-X.; Tang, Y.-B.; Peng, Y.; Yan, K.; Bagheri, M.; Redd, B.A.; Brandon, C.J.; Lu, Z.; Han, M.; Xiao, J.; et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digit. Med.* **2020**, *3*, 70.
8. Baltruschat, I.M.; Nickisch, H.; Grass, M.; Knopp, T.; Saalbach, A. Comparison of deep learning approaches for multi-label chest X-ray classification. *Sci. Rep.* **2019**, *9*, 6381. [CrossRef]
9. Hu, T.; Khishe, M.; Mohammadi, M.; Parvizi, G.R.; Taher Karim, S.H.; Rashid, T.A. Real-time COVID-19 diagnosis from X-ray images using deep CNN and extreme learning machines stabilized by chimp optimization algorithm. *Biomed. Signal Process. Control* **2021**, *68*, 102764. [CrossRef]
10. Wu, C.; Khishe, M.; Mohammadi, M.; Taher Karim, S.H.; Rashid, T.A. Evolving deep convolutional neutral network by hybrid sine–cosine and extreme learning machine for real-time COVID19 diagnosis from X-ray images. *Soft Comput.* **2021**. [CrossRef]
11. Wang, X.; Gong, C.; Khishe, M.; Mohammadi, M.; Rashid, T.A. Pulmonary diffuse airspace opacities diagnosis from chest X-ray images using deep convolutional neural networks fine-tuned by whale optimizer. *Wirel. Pers. Commun.* **2021**, *124*, 1355–1374. [CrossRef] [PubMed]
12. Khishe, M.; Caraffini, F.; Kuhn, S. Evolving deep learning convolutional neural networks for early COVID-19 detection in chest X-ray images. *Mathematics* **2021**, *9*, 1002. [CrossRef]
13. Ibrahim, D.M.; Elshennawy, N.M.; Sarhan, A.M. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Comput. Biol. Med.* **2021**, *132*, 104348. [CrossRef] [PubMed]
14. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks. *arXiv* **2014**, arXiv:1411.1792.
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
16. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv* **2017**, arXiv:1705.02315.
17. 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.
18. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, 8–12 October 2019; Volume 33, pp. 590–597.
19. Azizpour, H.; Razavian, A.S.; Sullivan, J.; Maki, A.; Carlsson, S. Factors of transferability for a generic convnet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1790–1802. [CrossRef]
20. Cui, Y.; Song, Y.; Sun, C.; Howard, A.; Belongie, S. Large scale fine-grained categorization and domain-specific transfer learning. *arXiv* **2018**, arXiv:1806.06193.
21. Tajbakhsh, N.; Shin, J.Y.; Gurudu, S.R.; Hurst, R.T.; Kendall, C.B.; Gotway, M.B.; Liang, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. Med. Imaging* **2016**, *35*, 1299–1312. [CrossRef]
22. Gozes, O.; Greenspan, H. Deep feature learning from a hospital-scale chest X-ray dataset with application to TB detection on a small-scale dataset. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23–27 July 2019; pp. 4076–4079.

23. Python. Available online: https://www.python.org/ (accessed on 1 May 2020).
24. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
25. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
27. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*; ECML PKDD 2011; Lecture Notes in Computer Science; Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6913, pp. 145–158.
28. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
29. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. CoAtNet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.

diagnostics

MDPI

Article

A Deep Modality-Specific Ensemble for Improving Pneumonia Detection in Chest X-rays

Sivaramakrishnan Rajaraman *,†, Peng Guo †, Zhiyun Xue and Sameer K. Antani

Computational Health Research Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; peng.guo@nih.gov (P.G.); zhiyun.xue@nih.gov (Z.X.); santani@mail.nih.gov (S.K.A.)
* Correspondence: sivaramakrishnan.rajaraman@nih.gov
† These authors contributed equally to this work.

Abstract: Pneumonia is an acute respiratory infectious disease caused by bacteria, fungi, or viruses. Fluid-filled lungs due to the disease result in painful breathing difficulties and reduced oxygen intake. Effective diagnosis is critical for appropriate and timely treatment and improving survival. Chest X-rays (CXRs) are routinely used to screen for the infection. Computer-aided detection methods using conventional deep learning (DL) models for identifying pneumonia-consistent manifestations in CXRs have demonstrated superiority over traditional machine learning approaches. However, their performance is still inadequate to aid in clinical decision-making. This study improves upon the state of the art as follows. Specifically, we train a DL classifier on large collections of CXR images to develop a CXR modality-specific model. Next, we use this model as the classifier backbone in the RetinaNet object detection network. We also initialize this backbone using random weights and ImageNet-pretrained weights. Finally, we construct an ensemble of the best-performing models resulting in improved detection of pneumonia-consistent findings. Experimental results demonstrate that an ensemble of the top-3 performing RetinaNet models outperformed individual models in terms of the mean average precision (mAP) metric (0.3272, 95% CI: (0.3006,0.3538)) toward this task, which is markedly higher than the state of the art (mAP: 0.2547). This performance improvement is attributed to the key modifications in initializing the weights of classifier backbones and constructing model ensembles to reduce prediction variance compared to individual constituent models.

Keywords: chest X-ray; deep learning; modality-specific knowledge; object detection; RetinaNet; ensemble learning; pneumonia; mean average precision

Citation: Rajaraman, S.; Guo, P.; Xue, Z.; Antani, S.K. A Deep Modality-Specific Ensemble for Improving Pneumonia Detection in Chest X-rays. *Diagnostics* **2022**, *12*, 1442. https://doi.org/10.3390/diagnostics12061442

Academic Editor: Henk A. Marquering

Received: 17 May 2022
Accepted: 8 June 2022
Published: 11 June 2022

1. Introduction

Pneumonia is an acute respiratory infectious disease that can be caused by various pathogens such as bacteria, fungi, or viruses [1]. The infection affects the alveoli in the lungs by filling them up with fluid or pus, thereby resulting in reduced intake of oxygen and causing difficulties in breathing. The potency of the disease depends on several factors including age, health, and the source of infection. According to the World Health Organization (WHO) report (https://www.who.int/news-room/fact-sheets/detail/pneumonia, accessed on 11 December 2021), pneumonia is reported to be an infectious disease that results in a higher mortality rate, particularly in children. About 22% of all deaths in pediatrics from 1 to 5 years of age are reported to result from this infection. Effective diagnosis and treatment of pneumonia are therefore critical to improving patient care and survival rate.

Chest X-rays (CXRs) are commonly used to screen for pneumonia infection [2,3]. Analysis of CXR images can be particularly challenging in low and middle-income countries due to a lack of expert resources, socio-economic factors, etc. [4]. Computer-aided detection systems using conventional deep learning (DL) methods, a sub-class of machine learning (ML) algorithms can alleviate this burden and have demonstrated superiority over traditional machine learning methods in detecting disease regions of interest (ROIs) [5,6]. Such

algorithms (i) automatically detect pneumonia-consistent manifestations on CXRs; and (ii) can support clinical-decision making by facilitating swift referrals for critical cases to improve patient care.

1.1. Related Works

A study of the literature reveals several studies that propose automated methods using DL models for detecting pneumonia-consistent manifestations on CXRs. However, DL models vary in their architecture and learn discriminative features from different regions in the feature space. They are observed to be highly sensitive to data fluctuations resulting in poor generalizability due to varying degrees of biases and variances. An approach to achieving a low bias and variance and ensuring reliable outcomes is using ensemble learning which is an established ML paradigm that combines predictions from multiple diverse DL models and improves performance compared to individual constituent models [7]. The authors of [8] proposed an ensemble of Faster-RCNN [9], Yolov5 [8], and EfficientDet [8] models to localize and predict bounding boxes containing pneumonia-consistent findings in the publicly available VinDr-CXR [8] dataset and reported a mean Average Precision (mAP) of 0.292. The following methods used ensembled object detection models to detect pneumonia-consistent findings using the CXR collection hosted for the RSNA Kaggle pneumonia detection challenge (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge accessed on 3 March 2022). The current state-of-the-art method according to the challenge leaderboard (https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge/leaderboard accessed on 3 March 2022) has a mAP of 0.2547. In [10], an ensemble of RetinaNet [11] and Mask RCNN models with ResNet-50 and ResNet-101 classifier backbones delivered a performance with a mAP of 0.2283 using the RSNA Kaggle pneumonia detection challenge CXR dataset. Another study [12] proposed a weighted-voting ensemble of the predictions from Mask R-CNN and RetinaNet models to achieve an mAP of 0.2174 in detecting pneumonia-consistent manifestations. These studies used the randomized test set split from the challenge-provided training data. This is a serious concern since the organizers have not made the blinded test set used during the challenge available for further use. This cripples follow-on research, such as ours, from making fair comparisons.

1.2. Rationale for the Study

All above studies used off-the-shelf DL object detection models with ImageNet [13] pretrained classifier backbones. However, ImageNet is a collection of stock photographic images whose visual characteristics, including shape and texture among others, are distinct from CXRs. As well, the disease-specific ROIs in CXRs are relatively small and many go unnoticed which may result in suboptimal predictions [14]. Our prior works and other literature have demonstrated that the knowledge transferred from DL models that are retrained on a large collection of CXR images is shown to improve performance on relevant target medical visual recognition tasks [15–17]. To the best of our knowledge, we observed that no literature discussed the use of CXR modality-specific backbones in object detection models, particularly applied to detecting pneumonia-consistent findings in CXRs.

1.3. Contributions of the Study

Our study improves upon the state-of-the-art as follows:

(i). To the best of our knowledge, this is the first study that studies the impact of using CXR modality-specific classifier backbones in a RetinaNet-based object detection model, particularly applied to detecting pneumonia-consistent findings in CXRs.

(ii). We train state-of-the-art DL classifiers on large collections of CXR images to develop CXR modality-specific models. Next, we use these models as the classifier backbone in the RetinaNet object detection network. We also initialize this backbone using random weights and ImageNet-pretrained weights to compare detection performance.

Finally, we construct an ensemble of the aforementioned models resulting in improved detection of pneumonia-consistent findings.

(iii). Through this approach, we aim to study the combined benefits of various weight initializations for classifier backbones and construct an ensemble of the best-performing models to improve detection performance. The models' performance is evaluated in terms of mAP and statistical significance is reported in terms of confidence intervals (CIs) and *p*-values.

Section 2 discusses the datasets, model architecture, training strategies, loss functions, evaluation metrics, statistical methods, and computational resources, Section 3 elaborates on the results and Section 4 concludes this study.

2. Materials and Methods

2.1. Data Collection and Preprocessing

The following data collections are used for this study:

(i). CheXpert CXR [18]: The dataset includes 223,648 frontal and lateral CXR images that are collected from 65,240 patients at Stanford Hospital, California, USA. The CXRs are labeled for 14 cardiopulmonary disease manifestations, the details are extracted from the associated radiology reports using an automated labeling algorithm.

(ii). TBX11K CXR [19]: This collection includes 11,200 CXRs collected from normal patients and those with other cardiopulmonary abnormalities. The abnormal CXRs are collected from patients tested with the microbiological gold standard. There are 5000 CXRs showing no abnormalities and 6200 CXRs showing other abnormal findings including those collected from sick patients (n = 5000), active Tuberculosis (TB) (n = 924), latent Tuberculosis (n = 212), active and latent TB (n = 54), and other uncertain (n = 10) cases. The regions showing TB-consistent manifestations are labeled for the abnormal regions using coarse rectangular bounding boxes.

(iii). RSNA CXR [20]: This CXR collection is released by RSNA for the RSNA Kaggle Pneumonia detection challenge. The collection consists of 26,684 CXRs that include 6012 CXR images showing pneumonia-consistent manifestations, 8851 CXRs showing no abnormal findings, and 11,821 CXRs showing other cardiopulmonary abnormalities. The CXRs showing pneumonia-consistent findings are labeled for abnormal regions using rectangular bounding boxes and are made available for the detection challenge.

We used the frontal CXRs from the CheXpert and TBX11K data collection during CXR image modality-specific retraining and those from the RSNA CXR collection to train the RetinaNet-based object detection models. All images are resized to 512×512 spatial dimensions to reduce computation complexity. The contrast of the CXRs is further increased by saturating the top 1% and bottom 1% of all the image pixel values. For CXR modality-specific retraining, the frontal CXR projections from the CheXpert and TBX11K datasets are divided at the patient level into 70% for training, 10% for validation, and 20% for testing. This patient-level split prevents the leakage of data and subsequent bias during model training. For object detection, the frontal CXRs from the RSNA CXR dataset that shows pneumonia-consistent manifestations are divided at the patient level into 70% for training, 10% for validation, and 20% for testing. Table 1 shows the number of CXR images across the training, validation, and test sets used for CXR modality-specific retraining and object detection, respectively.

Table 1. Patient-level dataset splits show the number of images for CXR modality-specific retraining and object detection. Note: TBX11K and RSNA datasets have one image per patient.

Dataset	Train		Validation		Test	
	Abnormal	**Normal**	**Abnormal**	**Normal**	**Abnormal**	**Normal**
CXR Modality-specific retraining						
CheXpert	13,600	13,600	1700	1700	1700	1700
TBX11k	3040	3040	380	380	380	380
RetinaNet-based object detection						
Dataset	Train		Validation		Test	
RSNA	4212		600		1200	

2.2. Model Architecture

2.2.1. CXR Modality-Specific Retraining

The ImageNet-pretrained DL models, viz., VGG-16, VGG-19, DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNet have demonstrated promising performance in several medical visual recognition tasks [14,19,21–23]. These models are further retrained on a large collection of CXR images to classify them as showing cardiopulmonary abnormal manifestations or no abnormalities. Such retraining helps the models to learn CXR image modality-specific features that can be transferred and fine-tuned to improve performance in a relevant task using CXR images. The best-performing model with the learned CXR image modality-specific weights is used as the classifier backbone to train the RetinaNet-based object detection model toward detecting pneumonia-consistent manifestations. Figure 1 shows the block diagram illustrating the steps involved in CXR image modality-specific retraining.

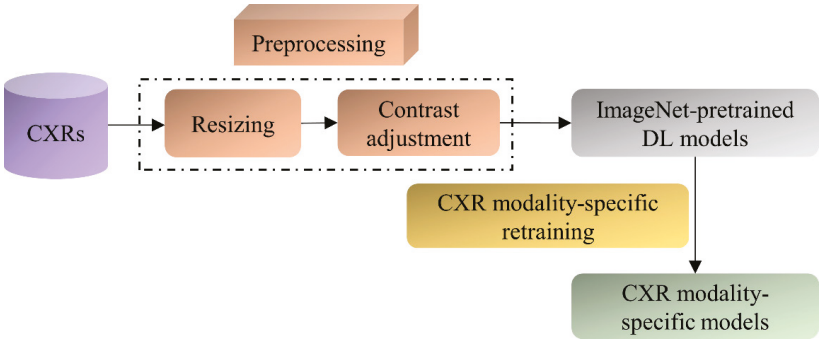


Figure 1. Steps illustrating CXR image modality-specific retraining of the ImageNet-pretrained models.

2.2.2. RetinaNet Architecture

We used RetinaNet as the base object detection architecture in our experiments. The architecture of the RetinaNet model is shown in Figure 2. As a single-stage object detection structure, RetinaNet shares a similar concept of "anchor proposal" with [24]. It used a feature pyramid network (FPN) [25] where features on each of the image scales are computed separately in the lateral connections and then summed up through convolutional operations via the top-down pathways. The FPN network combines low-resolution features with strong semantic information, and high-resolution features with weak semantics through top-down paths and horizontal connections. Thus, feature maps with rich semantic information are obtained that would prove beneficial for detecting relatively smaller ROIs consistent with pneumonia compared to the other parts of the CXR image. Furthermore, when trained to minimize the focal loss [5], the RetinaNet was reported to deliver significant performance focusing on hard, misclassified examples.

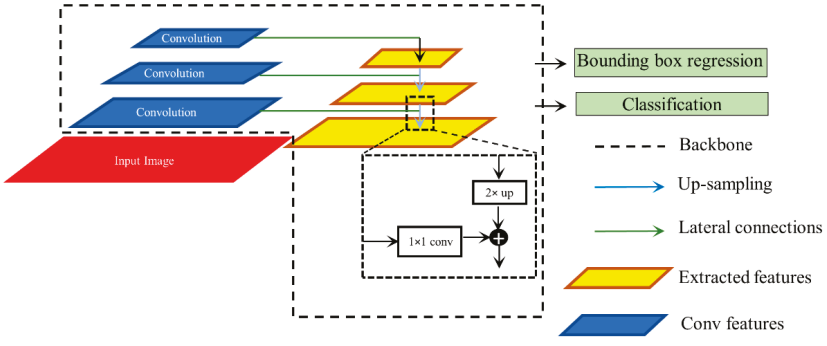


Figure 2. Method flowchart for the RetinaNet network.

2.2.3. Ensemble of RetinaNet Models with Various Backbones

We initialized the weights of the VGG-16 and ResNet-50 classifier backbones used in the RetinaNet model using three strategies: (i) Random weights; (ii) ImageNet-pretrained weights, and (iii) CXR image modality-specific retrained weights as discussed in Section 2.2.1. Each model is trained for 80 epochs and the model weights (snapshots) are stored at the end of each epoch. Varying modifications of the RetinaNet model classifier backbones and loss functions are mentioned in Table 2.

Table 2. RetinaNet model classifier backbones with varying weight initializations and loss functions. The loss functions mentioned are used for classification. For bounding box regression, only the smooth-L1 loss function [26] is used in all cases.

ResNet-50 Backbone and Classification Loss Functions	VGG-16 Backbone and Classification Loss Functions
ResNet-50 with random weights + focal loss	VGG-16 with random weights + focal loss
ResNet-50 with random weights + focal Tversky loss	VGG-16 with random weights + focal Tversky loss
ResNet-50 with ImageNet pretrained weights + focal loss	VGG-16 with ImageNet pretrained weights + focal loss
ResNet-50 with ImageNet pretrained weights + focal Tversky loss	VGG-16 with ImageNet pretrained weights + focal Tversky loss
ResNet-50 with CXR image modality-specific weights + focal loss	VGG-16 with CXR image modality-specific weights + focal loss
ResNet-50 with CXR image modality-specific weights + focal Tversky loss	VGG-16 with CXR image modality-specific weights + focal Tversky loss

We adopted the non-maximum suppression (NMS) in the RetinaNet training with an IoU threshold of 0.5 and evaluated the models using all the predictions with a confidence score over 0.9. A weighted averaging ensemble is constructed using (i) the top-3 performing models from the 12 RetinaNet models mentioned in Table 2, and (ii) the top-3 performing snapshots (model weights) using each classifier backbone. We empirically assigned the weights as 1, 0.9, and 0.8 for the predictions of the 1st, 2nd, and 3rd best performing models. A schematic of the ensemble procedure is shown in Figure 3. An ensembled bounding box is generated if the IOU of the weighted average of the predicted bounding boxes and the ground truth (GT) boxes is greater than 0.5. The ensembled model is evaluated based on the mean average precision (mAP) metric.

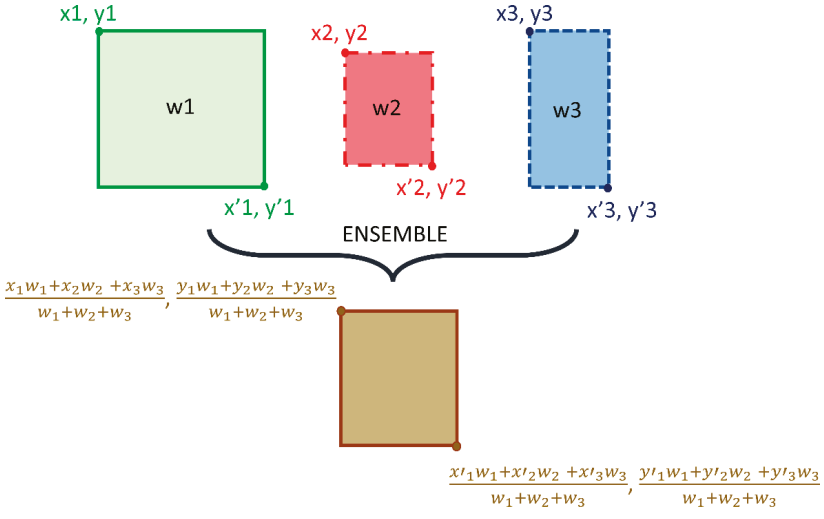


Figure 3. Method Schematic of the ensemble approach.

2.2.4. Loss Functions and Evaluation Metrics

CXR Image Modality-Specific Retraining

During CXR image modality-specific retraining, the DL models are retrained on a combined selection of the frontal CXR projections from the CheXpert and TBX11K datasets (details in Table 1). The training is performed for 128 epochs to minimize the categorical cross-entropy (CCE) loss. The CCE loss is the most commonly used loss function in classification tasks, and it helps to measure the distinguishability between two discrete probability distributions. It is expressed as shown in Equation (1).

$$CCE_{loss} = -\sum_{k=1}^{output\ size} y_k log\hat{y_k} \quad (1)$$

Here, $\hat{y_k}$ denotes the kth scalar value in the model output, y_k denotes the corresponding target, and the *output size* denotes the number of scalar values in the model output. The term y_k denotes the probability that event k occurs and the sum of all $y_k = 1$. The minus sign in the CCE loss equation ensures the loss is minimized when the distributions become less distinguishable. We used a stochastic gradient descent optimizer with an initial learning rate of 1×10^{-4} and momentum of 0.9 to reduce the CCE loss and improve performance. Callbacks are used to store the model checkpoints and the learning rate is reduced after a patience parameter of 10 epochs when the validation performance ceased to improve. The weights of the model that delivered a superior performance with the validation set are used to predict the test set. The models are evaluated in terms of accuracy, the area under the receiver-operating characteristic curve (AUROC), the area under the precision-recall (PR) curve (AUPRC), sensitivity, precision, F-score, Matthews correlation coefficient (MCC), and Kappa statistic.

RetinaNet-Based Detection of Pneumonia-Consistent Findings

Considering medical images, the disease ROIs span a relatively smaller portion of the whole image. This results in a considerably high degree of imbalance in the foreground ROI and the background pixels. These issues are particularly prominent in applications such as detecting cardiopulmonary manifestations like pneumonia where the number of pixels showing pneumonia-consistent manifestations is markedly lower compared to the total number of image pixels. Generalized loss functions such as balanced cross-entropy loss do

not take this data imbalance into account. This may lead to a learning bias and subsequent adversity in learning the minority ROI pixels. Appropriate selection of the loss function is therefore critical for improving detection performance. In this regard, the authors of [11] proposed the focal loss for object detection, an extension of the cross-entropy loss, which alleviates this learning bias by giving importance to the minority ROI pixels while down-weighting the majority background pixels. Minimizing the focal loss thereby reduces the loss contribution from majority background examples and increases the importance of correctly detecting the minority disease-positive ROI pixels. The focal loss is expressed as shown in Equation (2).

$$Focal\ loss(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{2}$$

Here, p_t denotes the probability the object detection model predicts for the GT. The parameter γ decides the rate of down-weighting the majority (background non-ROI) samples. The equation converges to the conventional cross-entropy loss when $\gamma = 0$. We empirically selected the value of $\gamma = 2$ which delivered superior detection performance.

Another loss function called the Focal Tversky loss function [27], a generalization of the focal loss function, is proposed to tackle the data imbalance problem and is given in Equation (3). The Focal Tversky loss function generalizes the Tversky loss which is based on the Tversky index that helps achieve a superior tradeoff between recall and precision when trained on class-imbalanced datasets. The Focal Tversky loss function uses a smoothing parameter γ that controls the non-linearity of the loss at different values of the Tversky index to balance between the minority pneumonia-consistent ROI and majority background classes. In Equation (3), TI denotes the Tversky index, expressed as shown in Equation (4).

$$FT_{loss_c} = \sum_{c} 1 - TI_c^{\gamma} \tag{3}$$

$$TI_c = \frac{\sum_{i=1}^{M} t_{ic}g_{ic} + \in}{\sum_{i=1}^{M} t_{ic}g_{ic} + \alpha \sum_{i=1}^{M} t_{i\hat{c}}g_{ic} + \beta \sum_{i=1}^{M} t_{ic}g_{i\hat{c}} + \in} \tag{4}$$

Here, g_{ic} and t_{ic} denote the ground truth and predicted labels for the pneumonia class c, where g_{ic} and $t_{ic} \in \{0,1\}$. That is, t_{ic} denotes the probability that the pixel i belongs to the pneumonia class c and $t_{i\hat{c}}$ denotes the probability that the pixel i belongs to the background class $\hat{c}$. The same holds for g_{ic} and $g_{i\hat{c}}$. The term M denotes the total number of image pixels. The term $\in$ provides numerical stability to avoid divide-by-zero errors. The hyperparameters α and β are tuned to emphasize recall under class-imbalanced training conditions. The Tversky index is adapted to a loss function by minimizing $\sum_c 1 - TI_c$. After empirical evaluations, we fixed the value of $\gamma = 4/3$, $\alpha = 0.7$ and $\beta = 0.75$.

As is known, the loss function within RetinaNet is a summation of a couple of loss functions, one for classification and the other for bounding box regression. We left the Smooth-L1 loss that is used for bounding box regression unchanged. For classification, we explored the performance with focal loss and focal Tversky loss functions individually for training the RetinaNet models with varying weight initializations. We used the bounding box annotations [20] associated with the RSNA CXRs showing pneumonia-consistent manifestations as the GT bounding boxes and measured its agreement with that generated by the models initialized with random weights, ImageNet-pretrained, and CXR image modality-specific retrained classifier backbones. Let TP, FP, and FN denote the true positives, false positives, and false negatives, respectively. Given a pre-defined IOU threshold, a predicted bounding box is considered to be TP if it overlaps with the GT bounding box by a value equal to or exceeding this threshold. FP denotes that the predicted bounding box has no associated GT bounding box. FN denotes the GT bounding box has no associated predicted bounding box. The mAP is measured as the area under the precision-recall curve (AUPRC) as shown in Equation (5). Here, P denotes precision which measures the accuracy of predictions, and R denotes recall which measures how well the model identifies all the

TPs. They are computed as shown in Equations (6) and (7). The value of mAP lies in the range [0, 1].

$$mean\ average\ precision\ (mAP) = \int_0^1 P(R)dR \tag{5}$$

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{7}$$

We used a Linux system with 1080Ti GPU, the Tensorflow backend (v. 2.6.2) with Keras, and CUDA/CUDNN libraries for accelerating the graphical processing unit (GPU) toward training the object detection models that are configured in the Python environment.

2.3. Statistical Analysis

We evaluated statistical significance using the mAP metric achieved by the models trained with various weight initializations and loss functions. The 95% confidence intervals (CIs) are measured as the binomial interval using the Clopper-Pearson method.

3. Results and Discussion

We organized the results from our experiments into the following sections: Evaluating the performance of (i) CXR image modality-specific retrained models and (ii) RetinaNet object detection models using classifier backbones with varying weight initializations and loss functions.

3.1. Classification Performance during CXR Image Modality-Specific Retraining

Recall that the ImageNet-pretrained DL models are retrained on the combined selection of CXRs from the CheXpert and TBX11K collection. Such retraining is performed to convert the weight layers specific to the CXR image modality and let the models learn CXR modality-specific features to improve performance when the learned knowledge is transferred and fine-tuned for a related medical image visual recognition task. The performance achieved by the CXR image modality-specific retrained models using the hold-out test set is listed in Table 3 and the performance curves are shown in Figure 4. The *no-skill* line in Figure 4 denotes the performance when a classifier would fail to discriminate between the normal and abnormal CXRs and therefore would predict a random outcome or a specific category under all circumstances.

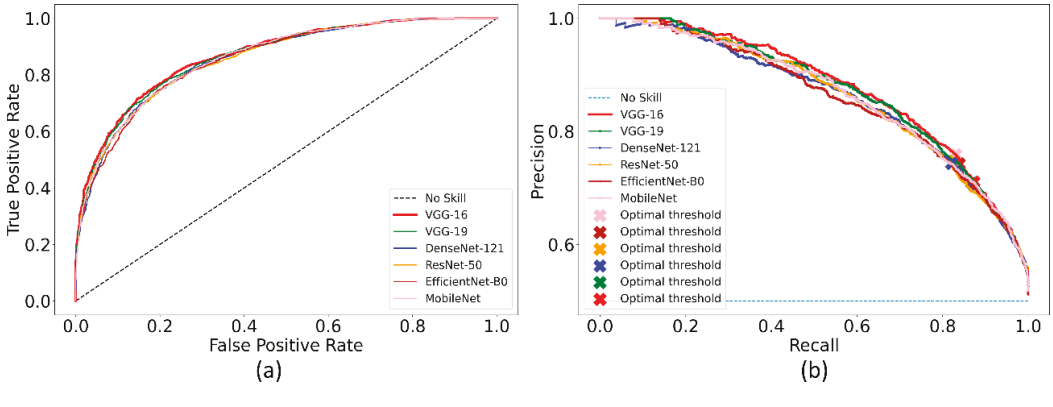


Figure 4. The collection of performance curves for the CXR image modality-specific retrained models. The performance is recorded at the optimal classification threshold measured with the validation data. (**a**) ROC and (**b**) PR curves.

Table 3. Performance of the CXR image modality-specific retrained models with the hold-out test set. Bold numerical values denote superior performance. The values in parenthesis denote the 95% CI for the MCC metric.

Models	Accuracy	AUROC	AUPRC	Sensitivity	Precision	F-Score	MCC	Kappa
VGG-16	**0.7834**	**0.8701**	**0.8777**	0.8303	**0.7591**	**0.7931**	**0.5693 (0.5542, 0.5844)**	**0.5668**
VGG-19	0.7743	0.8660	0.8727	0.8389	0.7429	0.7880	0.5532 (0.5380, 0.5684)	0.5486
DenseNet-121	0.7738	0.8582	0.8618	0.8264	0.7477	0.7851	0.5507 (0.5355, 0.5659)	0.5476
ResNet-50	0.7685	0.8586	0.8646	0.8207	0.7431	0.7800	0.5400 (0.5248, 0.5552)	0.5370
EfficientNet-B0	0.7553	0.8568	0.8612	0.8678	0.7084	0.7800	0.5240 (0.5088, 0.5392)	0.5106
MobileNet	0.7584	0.8609	0.8655	**0.8726**	0.7104	0.7832	0.5309 (0.5157, 0.5461)	0.5168

We could observe from Table 3 that the CXR image modality-specific retrained VGG-16 model demonstrates the best performance compared to other models in terms of all metrics except sensitivity. Of these, the MCC metric is a good measure to use because unlike F-score because it considers a balanced ratio of TPs TNs, FPs, and FNs. We noticed that the differences in the MCC values achieved by the various CXR image modality-specific retrained models are not significantly different ($p > 0.05$). Based on its performance, we used VGG-16 as the backbone for the RetinaNet detector. However, to enable fair comparison with other conventional RetinaNet-based results, we included the ResNet-50 backbone for detecting pneumonia-consistent manifestations. The VGG-16 and ResNet-50 classifier backbones are also initialized with random and ImageNet-pretrained weights for further comparison.

3.2. Detection Performance Using RetinaNet Models and Their Ensembles

Recall that the RetinaNet models are trained with different initializations of the classifier backbones. The performance achieved by these models using the hold-out test set is listed in Table 4. Figure 5 shows the PR curves obtained with the RetinaNet model using varying weight initializations for the selected classifier backbones. These curves show the precision and recall value of the model's bounding box predictions on every sample in the test set. We observe from Table 4 that the RetinaNet model with the CXR image modality-specific retrained ResNet-50 classifier backbone and trained using the focal loss function demonstrates superior performance in terms of mAP. Figure 6 shows the bounding box predictions of the top-3 performing RetinaNet models for a sample CXR from the hold-out test set.

We used two approaches to combine the bounding box predictions. They are (i) using the bounding box predictions from the top-3 performing RetinaNet models, viz., ResNet-50 with CXR image modality-specific weights + focal loss, ResNet-50 with CXR image modality-specific weights + focal Tversky loss, and ResNet-50 with random weights + focal loss; and, (ii) using the bounding box predictions from the top-3 performing snapshots (weights) within each model. The results are presented in Table 5 and Figure 7. A weighted averaging ensemble of the bounding boxes is generated when the IoU of the predicted bounding boxes is greater than the threshold value which is set at 0.5. Recall that the models are trained for 80 epochs and a snapshot (i.e., the model weights) is stored at the end of each epoch. We observed that the ensemble of the top-3 performing RetinaNet models delivered superior performance in terms of mAP metric compared to other models and ensembles. Figure 8 shows a sample CXR image with GT and predicted bounding

boxes using the weighted averaging ensemble of the top-3 individual models and the top-3 snapshots of the best-performing model.

Table 4. Performance of RetinaNet with the varying weight initializations for the classifier backbones and training losses. The values in parenthesis denote the 95% CI for the mAP metric. Bold numerical values denote superior performance.

Models	AUPRC (mAP)
ResNet-50 with random weights + focal loss	0.2763 (0.2509, 0.3017)
ResNet-50 with random weights + focal Tversky loss	0.2627 (0.2377, 0.2877)
ResNet-50 with ImageNet pretrained weights + focal loss	0.2719 (0.2467, 0.2971)
ResNet-50 with ImageNet pretrained weights + focal Tversky loss	0.2737 (0.2484, 0.2990)
ResNet-50 with CXR image modality-specific weights + focal loss	**0.2865 (0.2609, 0.3121)**
ResNet-50 with CXR image modality-specific weights + focal Tversky loss	0.2859 (0.2603, 0.3115)
VGG-16 with random weights + focal loss	0.2549 (0.2302, 0.2796)
VGG-16 with random weights + focal Tversky loss	0.2496 (0.2251, 0.2741)
VGG-16 with ImageNet pretrained weights + focal loss	0.2734 (0.2481, 0.2987)
VGG-16 with ImageNet pretrained weights + focal Tversky loss	0.2666 (0.2415, 0.2917)
VGG-16 with CXR image modality-specific weights + focal loss	0.2686 (0.2435, 0.2937)
VGG-16 with CXR image modality-specific weights + focal Tversky loss	0.2648 (0.2398, 0.2898)

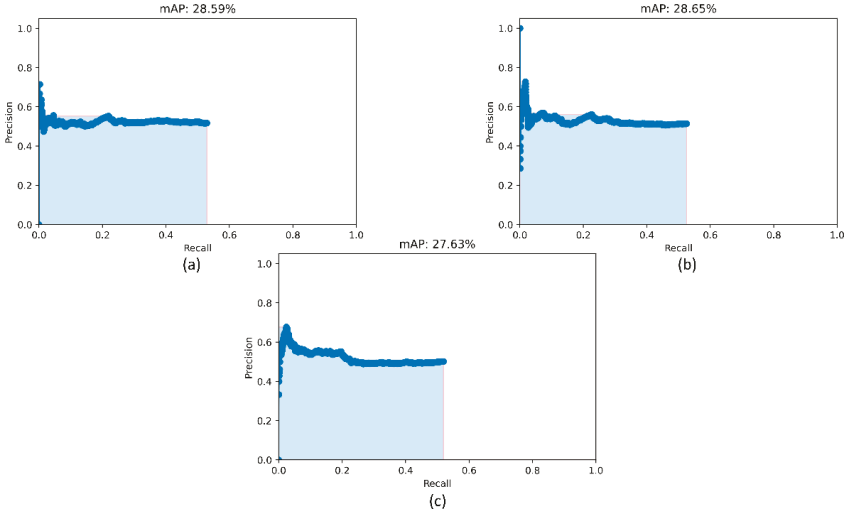


Figure 5. PR curves of the RetinaNet models initialized with varying weights for the classifier backbones. (**a**) ResNet-50 with CXR image modality-specific weights + focal Tversky loss; (**b**) ResNet-50 with CXR image modality-specific weights + focal loss, and (**c**) ResNet-50 with random weights + focal loss.

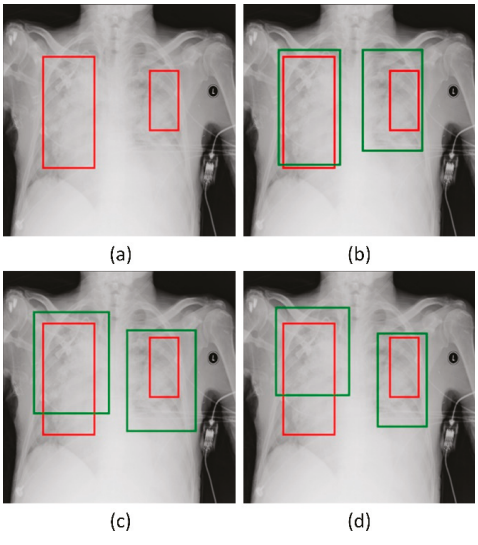


Figure 6. Bounding box predictions of the RetinaNet models initialized with varying weights for the classifier backbones. Green boxes denote the model predictions and red boxes denote the ground truth. (**a**) A sample CXR with ground truth bounding boxes. (**b**) ResNet-50 with CXR image modality-specific weights + focal Tversky loss; (**c**) ResNet-50 with CXR image modality-specific weights + focal loss, and (**d**) ResNet-50 with random weights + focal loss.

Table 5. Ensemble performance with the top-3 performing models (from Table 4) and the top-3 snapshots for each of the models trained with various classifier backbones and weight initializations. Values in parenthesis denote the 95% CI for the mAP metric. Bold numerical values denote superior performance.

Ensemble Method	mAP
Top-3 model ensemble (ResNet-50 with CXR image modality-specific weights + focal loss, ResNet-50 with CXR image modality-specific weights + focal Tversky loss, and ResNet-50 with random weights + focal loss	**0.3272 (0.3006, 0.3538)**
Ensemble of the top-3 snapshots for each model	
ResNet-50 with random weights + focal loss	0.2777 (0.2523, 0.3031)
ResNet-50 with random weights + focal Tversky loss	0.2630 (0.2380, 0.2880)
ResNet-50 with ImageNet pretrained weights + focal loss	0.2788 (0.2534, 0.3042)
ResNet-50 with ImageNet pretrained weights + focal Tversky loss	0.2812 (0.2557, 0.3067)
ResNet-50 with CXR image modality-specific weights + focal loss	0.2973 (0.2714, 0.3232)
ResNet-50 with CXR image modality-specific weights + focal Tversky loss	0.2901 (0.2644, 0.3158)
VGG-16 with random weights + focal loss	0.2633 (0.2383, 0.2883)
VGG-16 with random weights + focal Tversky loss	0.2556 (0.2309, 0.2803)
VGG-16 with ImageNet pretrained weights + focal loss	0.2823 (0.2568, 0.3078)
VGG-16 with ImageNet pretrained weights + focal Tversky loss	0.2715 (0.2463, 0.2967)
VGG-16 with CXR image modality-specific weights + focal loss	0.2813 (0.2558, 0.3068)
VGG-16 with CXR image modality-specific weights + focal Tversky loss	0.2698 (0.2446, 0.2950)

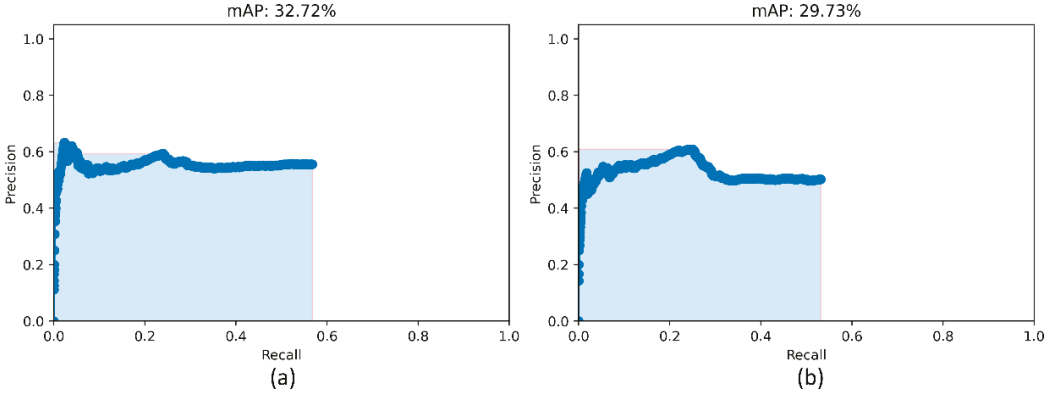


Figure 7. PR curves of the model ensembles. (**a**) PR curve obtained with the weighted-averaging ensemble of top-3 performing models (ResNet-50 with CXR modality-specific weights + focal loss, ResNet-50 with CXR modality-specific weights + focal Tversky loss, and ResNet-50 with random weights + focal loss and (**b**) PR curve obtained with the ensemble of top-3 performing snapshots while training the ResNet-50 with CXR modality-specific weights + focal loss model.

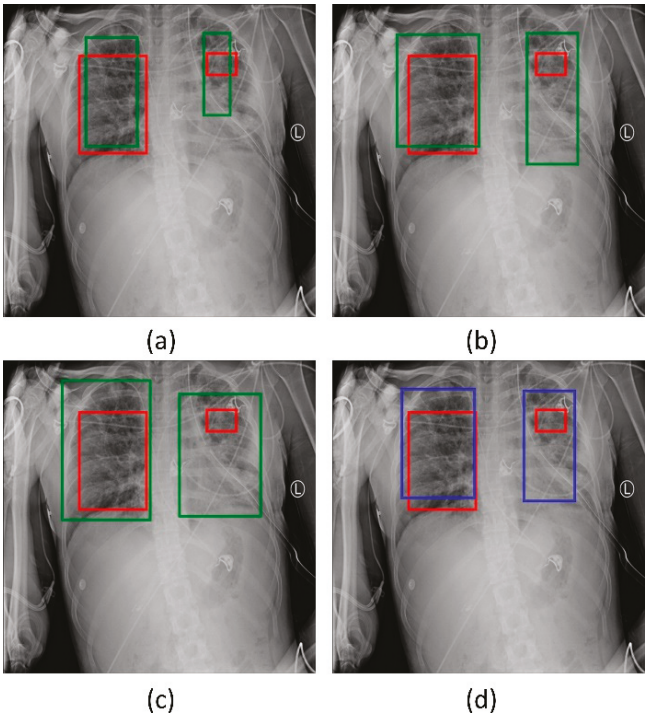


Figure 8. Bounding box predictions using the ensemble of RetinaNet models initialized with varying weights for the classifier backbones. Green boxes denote the individual model predictions, blue boxes denote the ensemble predictions and red boxes denote the ground truth. (**a**) ResNet-50 with CXR image modality-specific weights + focal Tversky loss; (**b**) ResNet-50 with CXR image modality-specific weights + focal loss; (**c**) ResNet-50 with random weights + focal loss, and (**d**) the ensembled bounding box prediction.

4. Conclusions and Future Work

In this study, we demonstrated the combined benefits of training CXR image modality-specific models, using them as backbones in an object detection model, evaluating them in different loss settings, and constructing ensembles of the best-performing models to improve performance in a pneumonia detection task. We observed that both CXR image modality-specific classifier backbones and ensemble learning improved detection performance compared to the individual constituent models. This study, however, suffers from the limitation that we have only investigated the effect of using CXR modality-specific classifier backbones in a RetinaNet-based object detection model to improve detecting pneumonia-consistent findings. The efficacy of this approach in detecting other cardiopulmonary disease manifestations is a potential avenue for future research. Additional diversity in the training process could be introduced by using CXR images and their disease-specific annotations collected from multiple institutions. With the advent of high-performance computing and current advancements in DL-based object detection, future studies could explore the use of mask x-RCNN, transformer-based models, and other advanced detection methods [28–31] and their ensembles in improving detection performance. Novel model optimization methods and loss functions can be proposed to further improve detection performance. However, the objective of this study is not to propose a new objection detection model but to validate the use of CXR modality-specific classifier backbones in existing models to improve performance. As the organizers of the RSNA Kaggle pneumonia detection challenge have not made the blinded GT annotations of the test set publicly available, we are unable to compare our results with the challenge leaderboard. However, the performance of our method on a random split from the challenge-provided training set, where we sequester 10% of the images for testing, using 70% for training and 20% for validation, respectively, is markedly superior to the best performing method on the leaderboard.

Author Contributions: Conceptualization, S.R., P.G. and S.K.A.; Data curation, S.R. and P.G.; Formal analysis, S.R., P.G. and Z.X.; Funding acquisition, S.K.A.; Investigation, Z.X. and S.K.A.; Methodology, S.R. and P.G.; Project administration, S.K.A.; Resources, S.K.A.; Software, S.R. and P.G.; Supervision, Z.X. and S.K.A.; Validation, S.R., P.G. and S.K.A.; Visualization, S.R. and P.G.; Writing—original draft, S.R. and P.G.; Writing—review & editing, S.R., P.G., Z.X. and S.K.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Institutional Review Board Statement: Ethical review and approval were waived for this study because of the retrospective nature of the study and the use of anonymized patient data.

Informed Consent Statement: Patient consent was waived by the IRBs because of the retrospective nature of this investigation and the use of anonymized patient data.

Data Availability Statement: The data required to reproduce this study is publicly available and cited in the manuscript.

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

References

1. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122–1131.e9. [CrossRef] [PubMed]
2. Rajaraman, S.; Candemir, S.; Thoma, G.; Antani, S. Visualizing and Explaining Deep Learning Predictions for Pneumonia Detection in Pediatric Chest Radiographs. *Proc. SPIE* **2019**, *10950*, 109500S.
3. Nishio, M.; Noguchi, S.; Matsuo, H.; Murakami, T. Automatic Classification between COVID-19 Pneumonia, Non-COVID-19 Pneumonia, and the Healthy on Chest X-Ray Image: Combination of Data Augmentation Methods. *Sci. Rep.* **2020**, *10*, 17532. [CrossRef]

4. Balabanova, Y.; Coker, R.; Fedorin, I.; Zakharova, S.; Plavinskij, S.; Krukov, N.; Atun, R.; Drobniewski, F. Variability in Interpretation of Chest Radiographs among Russian Clinicians and Implications for Screening Programmes: Observational Study. *BMJ* **2005**, *331*, 379–382. [CrossRef] [PubMed]
5. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
6. Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S.; Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S. Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs. *Appl. Sci.* **2018**, *8*, 1715. [CrossRef]
7. Mouhafid, M.; Salah, M.; Yue, C.; Xia, K. Deep Ensemble Learning-Based Models for Diagnosis of COVID-19 from Chest CT Images. *Healthcare* **2022**, *10*, 166. [CrossRef] [PubMed]
8. Pham, V.T.; Tran, C.M.; Zheng, S.; Vu, T.M.; Nath, S. Chest X-Ray Abnormalities Localization via Ensemble of Deep Convolutional Neural Networks. *Int. Conf. Adv. Technol. Commun.* **2021**, *2021*, 125–130. [CrossRef]
9. Xie, X.; Liao, Q.; Ma, L.; Jin, X. Gated Feature Pyramid Network for Object Detection. *Lect. Notes Comput. Sci.* **2018**, *11259*, 199–208. [CrossRef]
10. Mao, L.; Yumeng, T.; Lina, C. Pneumonia detection in chest X-rays: A deep learning approach based on ensemble RetinaNet and mask R-CNN. In Proceedings of the 8th International Conference on Advanced Cloud and Big Data, Taiyuan, China, 19–20 September 2020; pp. 213–218. [CrossRef]
11. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
12. Ko, H.; Ha, H.; Cho, H.; Seo, K.; Lee, J. Pneumonia detection with weighted voting ensemble of CNN models. In Proceedings of the 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 26–29 May 2019; pp. 306–310. [CrossRef]
13. Fei-Fei, L.; Deng, J.; Li, K. ImageNet: Constructing a Large-Scale Image Database. *J. Vis.* **2010**, *9*, 1037. [CrossRef]
14. Suzuki, K. Overview of Deep Learning in Medical Imaging. *Radiol. Phys. Technol.* **2017**, *10*, 257–273. [CrossRef] [PubMed]
15. Rajaraman, S.; Sornapudi, S.; Kohli, M.; Antani, S. Assessment of an ensemble of machine learning models toward abnormality detection in chest radiographs. In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019.
16. Rajaraman, S.; Folio, L.R.; Dimperio, J.; Alderson, P.O.; Antani, S.K. Improved Semantic Segmentation of Tuberculosis—Consistent Findings in Chest x-Rays Using Augmented Training of Modality-Specific u-Net Models with Weak Localizations. *Diagnostics* **2021**, *11*, 616. [CrossRef] [PubMed]
17. Yadav, O.; Passi, K.; Jain, C.K. Using deep learning to classify x-ray images of potential tuberculosis patients. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018.
18. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proc. Conf. AAAI Artif. Intell.* **2019**, *33*, 590–597. [CrossRef]
19. Liu, Y.; Wu, Y.H.; Ban, Y.; Wang, H.; Cheng, M.M. Rethinking computer-aided tuberculosis diagnosis. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
20. Shih, G.; Wu, C.C.; Halabi, S.S.; Kohli, M.D.; Prevedello, L.M.; Cook, T.S.; Sharma, A.; Amorosa, J.K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiol. Artif. Intell.* **2019**, *1*, e180041. [CrossRef] [PubMed]
21. Rajaraman, S.; Jaeger, S.; Thoma, G.R.; Antani, S.K.; Silamut, K.; Maude, R.J.; Hossain, M.A. Understanding the Learned Behavior of Customized Convolutional Neural Networks toward Malaria Parasite Detection in Thin Blood Smear Images. *J. Med. Imaging* **2018**, *5*, 034501. [CrossRef] [PubMed]
22. Ganesan, P.; Rajaraman, S.; Long, R.; Ghoraani, B.; Antani, S. Assessment of data augmentation strategies toward performance improvement of abnormality classification in chest radiographs. In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019.
23. Rajaraman, S.; Antani, S.K.; Poostchi, M.; Silamut, K.; Hossain, M.A.; Maude, R.J.; Jaeger, S.; Thoma, G.R.; Hossain, A.; Maude, R.J.; et al. Pre-Trained Convolutional Neural Networks as Feature Extractors toward Improved Malaria Parasite Detection in Thin Blood Smear Images. *PeerJ* **2018**, *6*, e4568. [CrossRef] [PubMed]
24. 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]
25. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
26. Liu, C.; Yu, S.; Yu, M.; Wei, B.; Li, B.; Li, G.; Huang, W. Adaptive smooth L1 loss: A better way to regress scene texts with extreme aspect ratios. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC), Athens, Greece, 5–8 September 2021. [CrossRef]
27. Abraham, N.; Khan, N.M. A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation. In Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI), Venice, Italy, 8–11 April 2019.

28. Qi, W.; Su, H. A Cybertwin Based Multimodal Network for ECG Patterns Monitoring Using Deep Learning. *IEEE Trans. Ind. Inform.* **2022**, *3203*, 1–9. [CrossRef]
29. Su, H.; Hu, Y.; Karimi, H.R.; Knoll, A.; Ferrigno, G.; De Momi, E. Improved Recurrent Neural Network-Based Manipulator Control with Remote Center of Motion Constraints: Experimental Results. *Neural Netw.* **2020**, *131*, 291–299. [CrossRef]
30. Qi, W.; Aliverti, A. A Multimodal Wearable System for Continuous and Real-Time Breathing Pattern Monitoring during Daily Activity. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2199–2207. [CrossRef]
31. Su, H.; Mariani, A.; Ovur, S.E.; Menciassi, A.; Ferrigno, G.; De Momi, E. Toward Teaching by Demonstration for Robot-Assisted Minimally Invasive Surgery. *IEEE Trans. Autom. Sci. Eng.* **2021**, *18*, 484–494. [CrossRef]

diagnostics

MDPI

Article

Automated 3D Segmentation of the Aorta and Pulmonary Artery on Non-Contrast-Enhanced Chest Computed Tomography Images in Lung Cancer Patients

Hao-Jen Wang [1,†], Li-Wei Chen [1,†], Hsin-Ying Lee [2], Yu-Jung Chung [1], Yan-Ting Lin [1], Yi-Chieh Lee [2], Yi-Chang Chen [1], Chung-Ming Chen [1,*] and Mong-Wei Lin [3,*]

[1] Department of Biomedical Engineering, College of Medicine and College of Engineering, National Taiwan University, Taipei 106, Taiwan; d04548013@ntu.edu.tw (H.-J.W.); f04548034@ntu.edu.tw (L.-W.C.); r08548005@ntu.edu.tw (Y.-J.C.); r08548047@ntu.edu.tw (Y.-T.L.); scsnake@gmail.com (Y.-C.C.)
[2] Department of Medicine, National Taiwan University, Taipei 100, Taiwan; b04401128@ntu.edu.tw (H.-Y.L.); b06401027@ntu.edu.tw (Y.-C.L.)
[3] Department of Surgery, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei 100, Taiwan
* Correspondence: chung@ntu.edu.tw (C.-M.C.); mwlin@ntu.edu.tw (M.-W.L.)
† These authors contributed equally to this work.

Abstract: Pulmonary hypertension should be preoperatively evaluated for optimal surgical planning to reduce surgical risk in lung cancer patients. Preoperative measurement of vascular diameter in computed tomography (CT) images is a noninvasive prediction method for pulmonary hypertension. However, the current estimation method, 2D manual arterial diameter measurement, may yield inaccurate results owing to low tissue contrast in non-contrast-enhanced CT (NECT). Furthermore, it provides an incomplete evaluation by measuring only the diameter of the arteries rather than the volume. To provide a more complete and accurate estimation, this study proposed a novel two-stage deep learning (DL) model for 3D aortic and pulmonary artery segmentation in NECT. In the first stage, a DL model was constructed to enhance the contrast of NECT; in the second stage, two DL models then applied the enhanced images for aorta and pulmonary artery segmentation. Overall, 179 patients were divided into contrast enhancement model (n = 59), segmentation model (n = 120), and testing (n = 20) groups. The performance of the proposed model was evaluated using Dice similarity coefficient (DSC). The proposed model could achieve 0.97 ± 0.007 and 0.93 ± 0.002 DSC for aortic and pulmonary artery segmentation, respectively. The proposed model may provide 3D diameter information of the arteries before surgery, facilitating the estimation of pulmonary hypertension and supporting preoperative surgical method selection based on the predicted surgical risks.

Keywords: aorta; computed tomography; deep learning; lung cancer; pulmonary artery; pulmonary hypertension

Citation: Wang, H.-J.; Chen, L.-W.; Lee, H.-Y.; Chung, Y.-J.; Lin, Y.-T.; Lee, Y.-C.; Chen, Y.-C.; Chen, C.-M.; Lin, M.-W. Automated 3D Segmentation of the Aorta and Pulmonary Artery on Non-Contrast-Enhanced Chest Computed Tomography Images in Lung Cancer Patients. *Diagnostics* **2022**, *12*, 967. https://doi.org/10.3390/diagnostics12040967

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 28 February 2022
Accepted: 9 April 2022
Published: 12 April 2022
Corrected: 2 August 2022

1. Introduction

Low-dose computed tomography (CT) screening has recently increased the detection rate of early-stage lung cancer [1–3]. Thoracic surgical resection is the major treatment approach for patients with early-stage lung cancer [4–7]. Surgical planning may vary from patient to patient owing to different surgical risks across patients. Extensive resection (lobectomy) is the treatment of choice for patients with a low surgical risk and high tumor invasiveness. However, limited resection (wedge resection or segmentectomy) is indicated for patients with high surgical risks and low tumor invasiveness [8–11]. Preoperative pulmonary hypertension associated with postoperative heart failure has been indicated to be exacerbated by surgery, leading to an increase in mortality risk (four to five times higher than that in patients without pulmonary hypertension) [12,13]. Furthermore, Wei et al.

showed that the failure rate of the right ventricle was significantly higher in patients with pulmonary hypertension before surgery (10.5%) than in patients without pulmonary hypertension (2.2%) [14]. Therefore, pulmonary hypertension is a surgical risk factor that may result in malignant behaviors; thus, it is important to preoperatively evaluate the presence of pulmonary hypertension, supporting surgical management. The gold standard approach for the diagnosis of pulmonary hypertension is the direct measurement of pulmonary artery pressure by cardiac catheterization [12,14]. However, this invasive measurement method may not be commonly used for the preoperative evaluation of patients with lung cancer. Cardiac ultrasound examination before lung cancer surgery would be an alternative approach to confirm the presence of pulmonary hypertension before surgery [12,14]. However, these results are unreliable and lack accuracy.

Several previous studies have indicated that the diameter of the pulmonary artery, or the ratio of the pulmonary artery to the aorta, is an effective tool for assessing pulmonary hypertension [4]. Chung et al. published imaging studies that measured these two parameters found that the diameter of the pulmonary artery increased significantly after lobectomy (23.9–25.6 mm, $p < 0.0001$) [15]. However, the method used in that study was a 2D measurement of contrast-enhanced chest CT images. As the majority of patients with lung cancer are diagnosed during low-dose CT (LDCT) screening, these patients may not have undergone contrast-enhanced CT before surgery. In addition, post-surgery tracking is usually by non-contrast-enhanced CT, and the technology of image measurement based on non-contrast-enhanced CT still needs to be developed.

The present study aimed to develop an automatic 3D segmentation method for the aorta and pulmonary artery on non-contrast-enhanced CT images to accurately calculate the 3D diameter information of the two arteries before surgery, facilitate the estimation of pulmonary hypertension, and support preoperative surgical management.

2. Materials and Methods

2.1. Data Information

Preoperative chest CT images of 179 patients with lung cancer were collected from the National Taiwan University Hospital between January 2011 and December 2019, and all patients had a set of CT images without a contrast agent and with a contrast agent. The inclusion criteria of the study were as follows: (1) pathologically confirmed lung cancer, and (2) available thin-cut chest CT image data. The Research Ethics Committee of the National Taiwan University Hospital approved this study (project approval number 201712087RIND) and waived the need for informed consent because of the retrospective study design.

The overall flowchart of the pulmonary hypertension assessment method is shown in Figure 1. In this study, two types of models, namely, the contrast-enhanced and segmentation models, were developed to achieve segmentation of the aorta and pulmonary artery in two stages. CT images obtained with a contrast agent and without a contrast agent were used in the contrast-enhanced model. The training data were allocated to 49 patients, and the validation data were allocated to 10 patients, with a total of 59 patients. On the contrary, because the clinical application of this model presupposes that patients are not evaluated with a contrast agent, the data used for the segmentation model were taken from non-contrast-enhanced CT images. In the segmentation model, 120 non-contrast-enhanced CT images from the dataset were used, divided into 80 patients (training data), 20 patients (validation data), and 20 patients (testing data), and this configuration was used for segmentation model training.

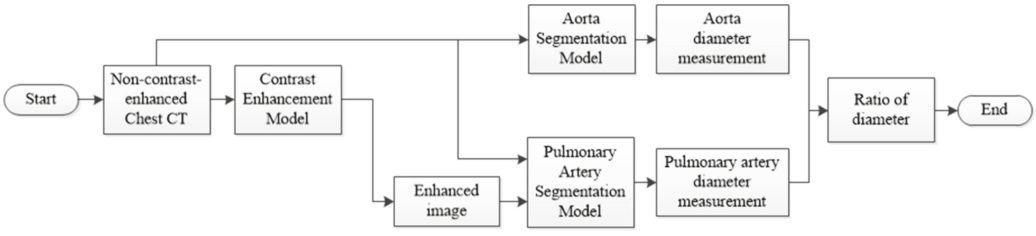


Figure 1. Flowchart of the pulmonary hypertension assessment method.

2.2. CT Image Acquisition

Chest CT scans used in this study were acquired from the following manufacturers using a multidetector (16-, 32-, or 64-detector row) CT scanner: GE (LightSpeed VCT, LightSpeed 16, and HiSpeed CT/I, Chicago, IL, USA), Siemens (Definition AS+, Emotion 16, and Sensation 64, Erlangen, Germany), and Philips (iCT 256 and Ingenuity CT, Amsterdam, Netherlands) Healthcare systems. The CT image parameters were as follows: detector collimation, 0.6–1.25 mm; field of view, 20–38 cm; beam pitch, 0.800–1.396; beam width, 10–40 mm; gantry speed, 0.5 or 0.8 s per rotation; 100–130 kVp; 47–351 mA; reconstruction interval, 0.39–6 mm; matrix, 512 × 512 mm^2.

2.3. Pre-Processing of CT Images

The original image was a set of chest CT Dicom images, and each data were resampled to 0.1 mm in data preprocessing, and the value was between −160 and 240 HU. To ensure that the deep learning (DL) model learns the location of blood vessels accurately, the aorta and pulmonary artery are considered as the center to cut out two sizes of volume of interest (VOI) (96 × 96 × 32 and 128 × 128 × 64). The aorta requires a larger VOI because of its anatomical shape that covers more slices. As there is currently no commercial software that can accurately annotate the aorta and pulmonary artery for CT images without contrast agents, this study invited two professional thoracologists to assist in the annotation of the ground truth (GT) of the two targets in this study. The preprocessing flow of the data is shown in Figure 2.

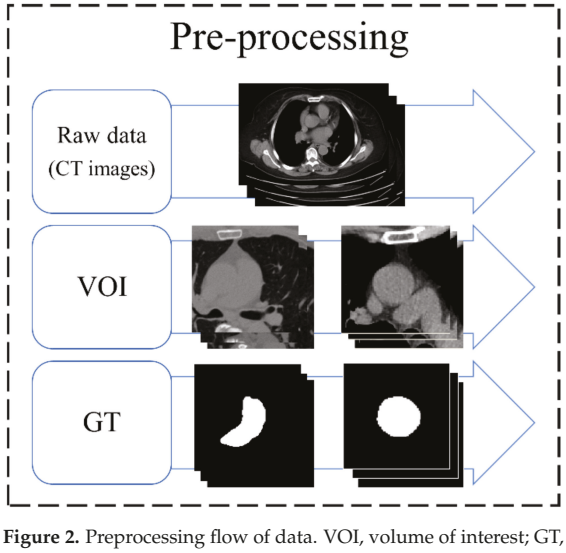


Figure 2. Preprocessing flow of data. VOI, volume of interest; GT, ground truth.

2.4. Architecture

To evaluate the complications, measurements of the diameter of the aorta and pulmonary artery and calculation of the ratio are required. This study proposes a two-stage DL architecture for the 3D segmentation of blood vessels. Because this study used non-contrast-enhanced CT images, a contrast enhancement model was used in the first stage to enhance the non-contrast images of the aorta and pulmonary artery; thus, the contrast between the blood vessels and the surrounding tissues in the non-contrast-enhanced CT images was improved. This model increases the sharpness of the blood vessel edge to facilitate the effective learning of the backward segmentation model. In the second stage, two 3D vessel segmentation models were developed for the aorta and pulmonary artery, as shown in the flowchart (Figure 3). After successfully segmenting the two vessels, the vessel sections were extracted to obtain the average diameter of the 3D vessel. The overall architecture is shown in Figure 3.

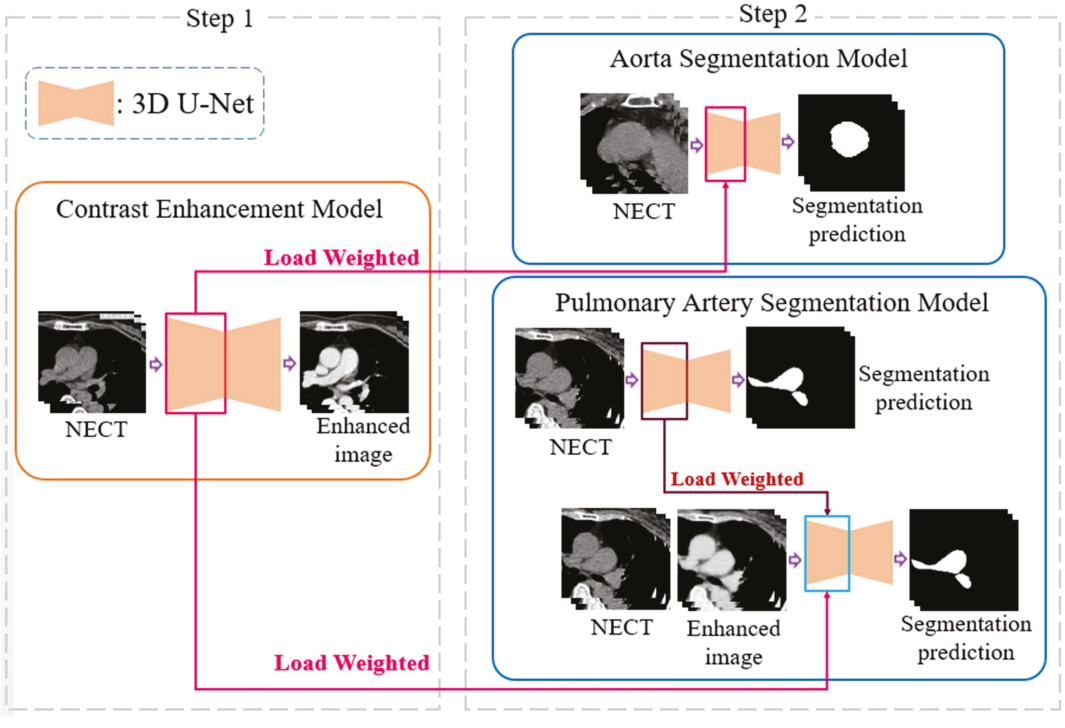


Figure 3. A two-stage deep learning architecture for 3D segmentation. NECT, non-contrast-enhanced Chest CT.

2.5. 3D U-Net

The contrast enhancement model and segmentation model proposed in this study are improved models based on U-Net [16,17]. U-Net is a convolution-based model that can be modeled by point-by-point convolution and superimposed on each convolution layer. It is a type of fully convolutional network (FCN) model [18] and is composed of a downsampling (contraction) path to aggregate high-level information using context modules and an upsampling (expansion) pathway to combine feature and spatial information for localization. In the downsampling path, each layer consists of two 3×3 convolutional layers, and then downsampling with a stride of 2 is used to extract information and capture the contour features of the input image with missing spatial information. This gradually restores the image size through upsampling with a step size of 2, extracting information

on important features from the original image information and integrating contextual information. Therefore, this model can perform feature extraction and multi-information transmission through two paths to achieve semantic segmentation. The network used in this study, called 3D U-Net, was changed from its original 2D architecture to a 3D architecture by using 3D volumes as input and processing them with corresponding 3D operations, such as 3D convolutions and 3D max pooling, as shown in Figure 4 [19].

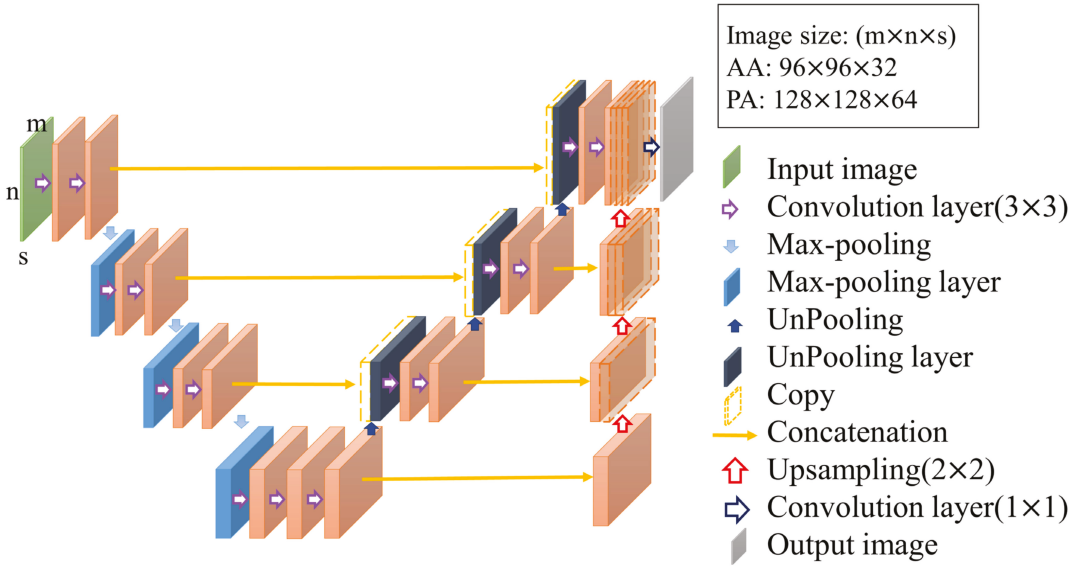


Figure 4. Structure of 3D U-Net.

2.6. Contrast Enhancement Model

In non-contrast-enhanced images, because the image contrast is not significant and the blood vessel boundaries are relatively blurred, it is not easy to segment the blood vessels directly. Therefore, this study proposes a contrast-enhancement model that uses the corresponding non-contrast- and contrast-enhanced images as the input and GT of the model, respectively, as shown in Figure 5. This model learns how to generate contrast-enhanced images from non-contrast-enhanced images so that the second-stage model in the architecture can more easily achieve segmentation. In this model, a combination of mean absolute error (MAE) and structural dissimilarity (DSSIM) loss functions is used as the loss function [20,21]. The MAE is the sum of the absolute values of the difference between the target value and the predicted value. It measures only the average error of the predicted value, regardless of the direction, and ranges from 0 to positive infinity; therefore, it can be used to judge the overall contrast enhancement performance of this model. DSSIM was derived from a formula based on the structural similarity index (SSIM) [21,22]. SSIM combines luminance, contrast, and structure to reflect the structural information heavily relied on by anthropology. In the chest CT images used in this study, there is a strong correlation between adjacent pixels in the same anatomical structure; therefore, it is suitable for use as the loss function of this model [22]. Therefore, this study adopts the advantages of the two loss functions and sets the trade-off parameter α to the optimal value of 0.7 after many experiments in this study, as shown in Formula (1), where I_gt and I_op are the GT and model output, respectively. In this study, this loss function is used in a U-

Net model with the ability to integrate contextual information for contrast enhancement model training.

$$Loss(I\_op,\ I\_gt) = \alpha MAE(I\_op,\ I\_gt) + (1-\alpha)DSSIM(I\_op,\ I\_gt) \quad (1)$$

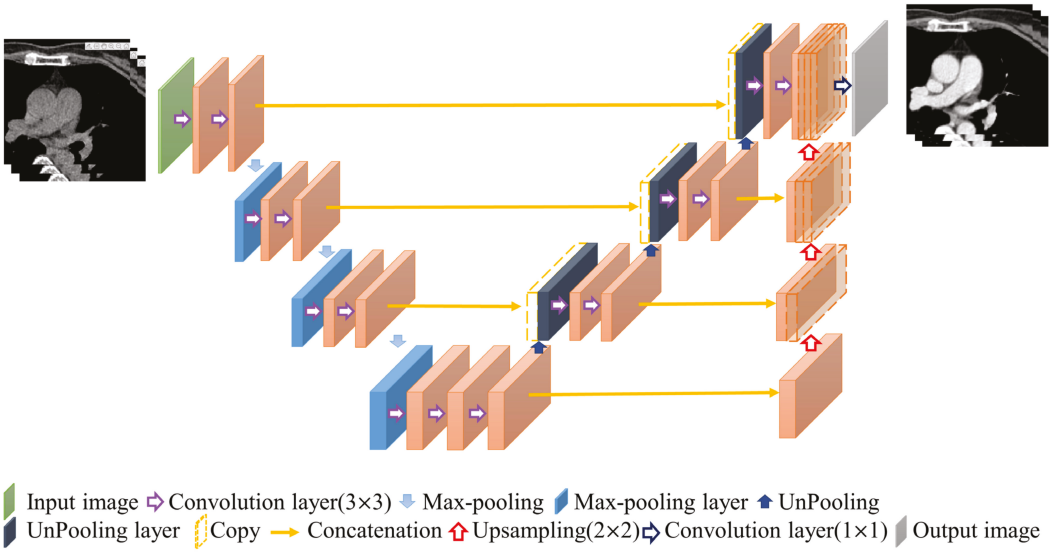


Figure 5. Contrast enhancement model.

2.7. Segmentation Model

In the second stage of the architecture, this study developed relative 3D vessel segmentation models for two vessels, namely the aorta and pulmonary artery. The ratio of the diameter of the aorta to that of the pulmonary artery is an important indicator of the presence or absence of pulmonary hypertension. To obtain the diameters of the two blood vessels, this study developed a 3D segmentation model to overcome the disadvantage of non-contrast-enhanced CT images to segment the anatomical structures of the two blood vessels and then extract the blood vessel sections to calculate the average blood vessel diameter.

2.7.1. Aorta Segmentation Model

The training process of the contrast enhancement model also learns the difference between the voxels of the blood vessels and those of other thoracic anatomical structures, which is similar to the purpose of finding the position of the blood vessel boundary in the segmentation model. Therefore, the training weights obtained in the first stage of the architecture are suitable for transfer learning to improve the learning efficiency of the segmentation model [23]. The location and method of the weights used in the transfer learning are shown in Figure 6. The aorta is relatively simple in the thoracic structure; therefore, this study directly uses unenhanced CT images for training and combines the loss function of the Dice loss function commonly used in segmentation models to achieve the training and development of the aorta segmentation model, in which both Pi and Gi are a single voxel of GT and model output, respectively, and N is the total number of voxels of the data, as shown in Formula (2).

$$\mathrm{DSC} = \frac{2\sum_i^N P_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \quad (2)$$

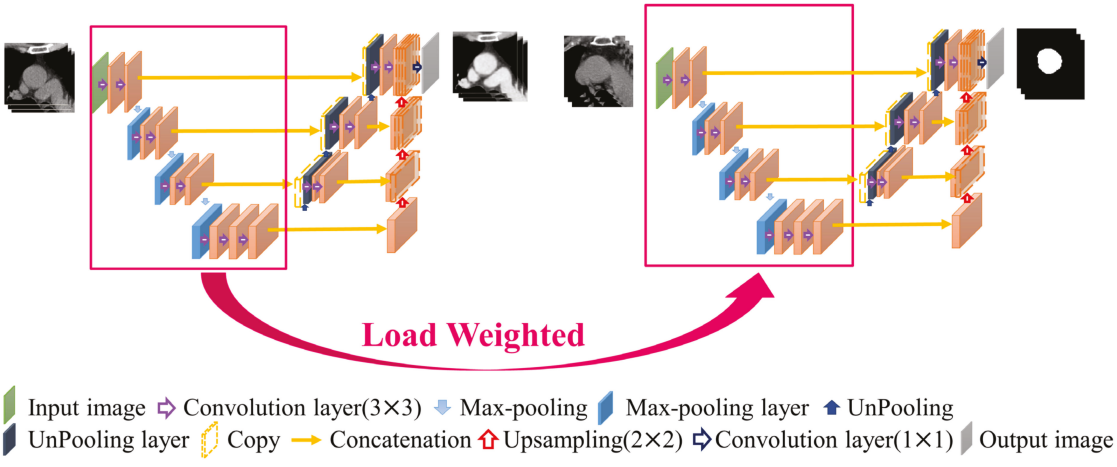


Figure 6. Aorta segmentation model.

2.7.2. Pulmonary Artery Segmentation Model

The pathological structure of the pulmonary artery is relatively variable in the direction and shape of the blood vessels. From the cross-section of the CT image, it can be observed that the shape of each slice is very different (Figure 7). To enable the model to learn the target more accurately, this study designed this model as a two-channel model. In addition to the original non-contrast-enhanced CT image as the input, the contrast-enhanced image learned by the contrast enhancement model is used as the second channel to input the pulmonary artery segmentation model. In addition, consistent with the aorta segmentation model, this model also uses transfer learning for the augmentation training of the model. The difference is that for the model to learn more accurate pulmonary artery voxel information, the model will pre-train the segmentation model and then concatenate the weights obtained in this pre-training with the training weights in the contrast enhancement model. Based on this, transfer learning was performed on the pulmonary artery segmentation model, as shown in Figure 8, to achieve a more complete pulmonary artery segmentation result.

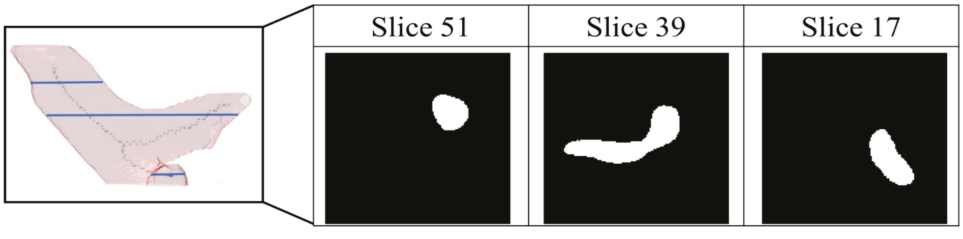


Figure 7. Pathological shape images of pulmonary aorta in different slice of the computed tomography images.

In this segmentation model, the combination of the weights obtained by segmentation pre-training and the training weights in the contrast enhancement model is shown in Figure 8. In this study, each layer of downsampling was combined individually to ensure that, during the feature extraction stage, both channels maintained the training impact. The model hyperparameters used in this study are presented in Table 1.

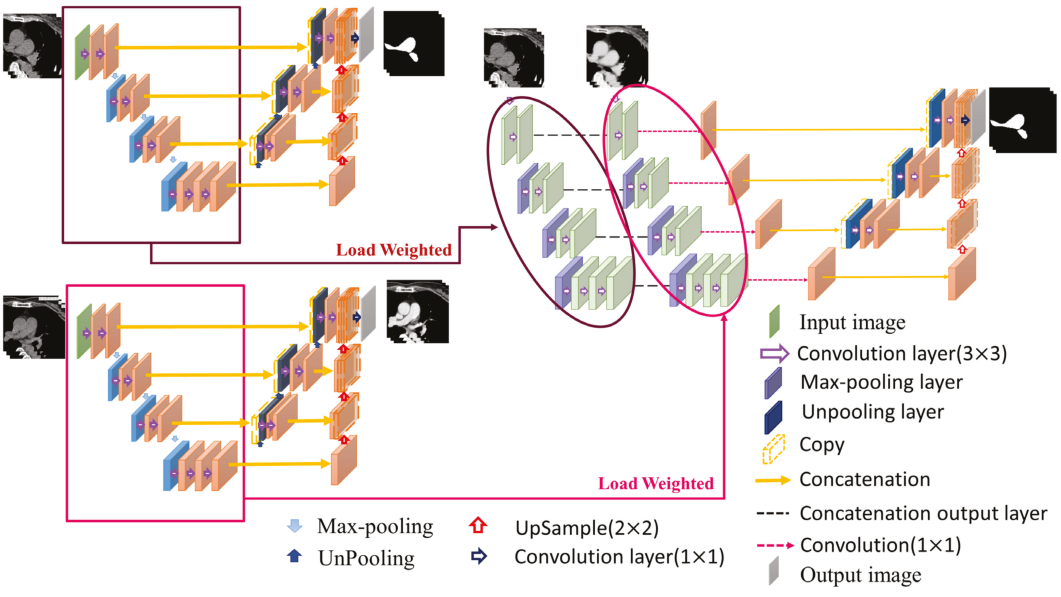


Figure 8. Pulmonary artery segmentation model.

Table 1. Hyperparameters of the two types of models used in the proposed architecture.

AA and PA	Learning Rate	Decay	Epochs	Loss Function	Spatial Dropout 3D	Convolution Kernel Size	Activation Function	Output Layer Activation Function
Contrast enhancement model Segmentation model	10^{-2}	10^{-6}	500	Combination of MAE and DSSIM Dice loss function	0.25	$3 \times 3 \times 3$	ReLU	Sigmoid

2.8. Vessel Diameter Measurement

After obtaining the aorta and pulmonary artery from the two-stage segmentation architecture, this study developed a mean diameter measurement method for both vessels for the vessel diameter measurement segment required for assessing pulmonary hypertension. To measure the vessel diameter, this study determined the centerline from the segmented 3D vessel images. Find the corresponding blood vessel section by the point on the centerline and calculate its diameter. After summation and averaging, the average diameters of the two blood vessels were obtained. However, the vascular shape of the pulmonary artery is more tortuous than that of the aorta; therefore, two different methods were used in the anterior segment of the measurement process (Figure 9).

According to the characteristics of the aorta blood vessel itself, the blood vessel was measured from 0.5 cm after exiting the heart to the position of 2.5 cm, as the range for calculating the average diameter of the entire aorta. The blood vessel diameter was measured (Figure 10). The original three-dimensional blood vessel is eroded to obtain a region as a limiting range to find the center line; second, the direction vector between points and the blood vessel surface are used. The normal vector is used as the inner product, and the minimum inner product value between each candidate point is compared to select the next point and so on until the entire blood vessel is searched; finally, the diameter of the

blood vessel section perpendicular to the centerline of each segment is used to calculate the average diameter.

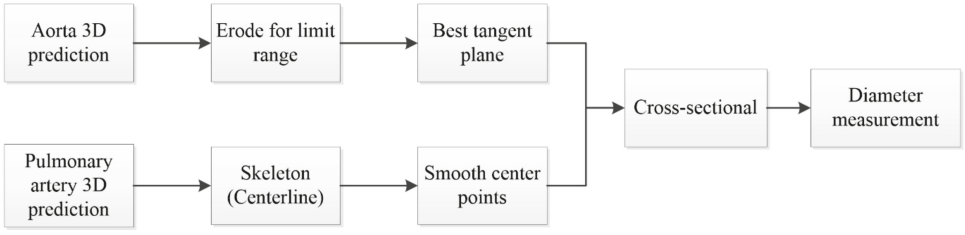


Figure 9. Flowchart of diameter measurement.

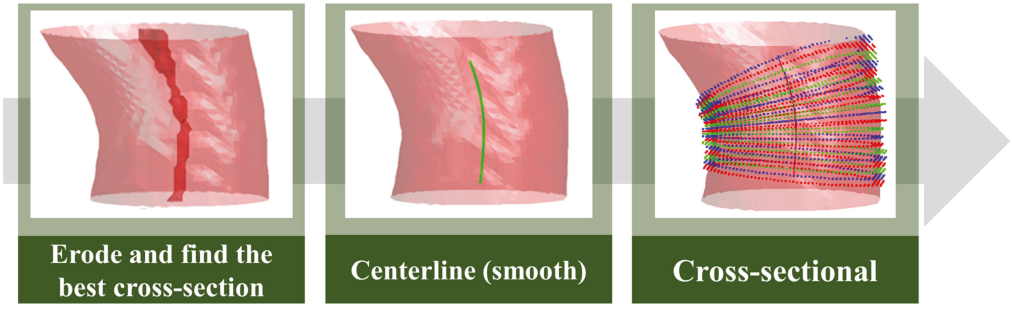


Figure 10. Aorta diameter measurement method.

To obtain the average diameter of the pulmonary artery, this study first used skeletonization [24] to determine the rough centerline (Figure 11) and started to measure along the main vessel of the pulmonary artery 0.5–1.5 cm from the position of the branch points. Furthermore, the points on the centerline were discretized by interpolation. This step smoothens the centerline. Finally, the diameter of the blood vessel section perpendicular to each segment of the centerline was calculated at every 0.04-cm interval to obtain the average diameter of the pulmonary artery.

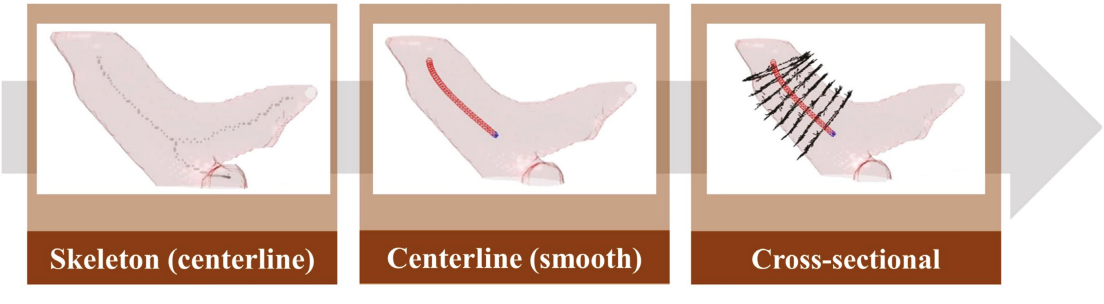


Figure 11. Pulmonary artery diameter measurement method.

3. Results

3.1. Patient Clinicopathological Features and Perioperative Results

This study cohort comprised 179 patients diagnosed with lung cancer who underwent lobectomies between 2011 and 2019. The mean age of all the 258 patients was 78.6 ± 3.3 years (range: 75−90). The majority of patients were females (64.2%) and nonsmokers (78.8%). The mean postoperative intensive care unit stay and hospital stay were

0.3 and 5.3 days, respectively. There was no 30-day mortality in the study cohort. Patient clinicopathological features and perioperative results are presented in Supplementary Table S1.

3.2. Contrast Enhancement Model

In the first stage of the architecture, there was only a slight difference between the contrast enhancement generated by the non-contrast agent CT image and the real contrast agent CT image. The method developed in this study can significantly enhance the vascular contrast of the non-contrast agent CT image, as shown in Figure 12.

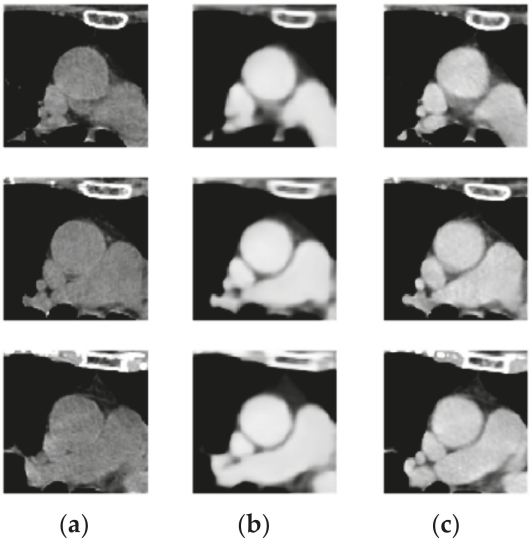


Figure 12. (**a**) Non-contrast-enhanced chest CT, (**b**) enhanced image, and (**c**) contrast CT image.

3.3. Segmentation Model

The training curves of the segmentation model of the aorta and pulmonary artery in this study are shown in Figures 13 and 14. It can be seen that regardless of whether it is the segmentation of the aorta or the segmentation model of the dual-channel pulmonary artery, in the later stage of model training, reliable and stable results can be achieved for both the training dataset and the validation dataset.

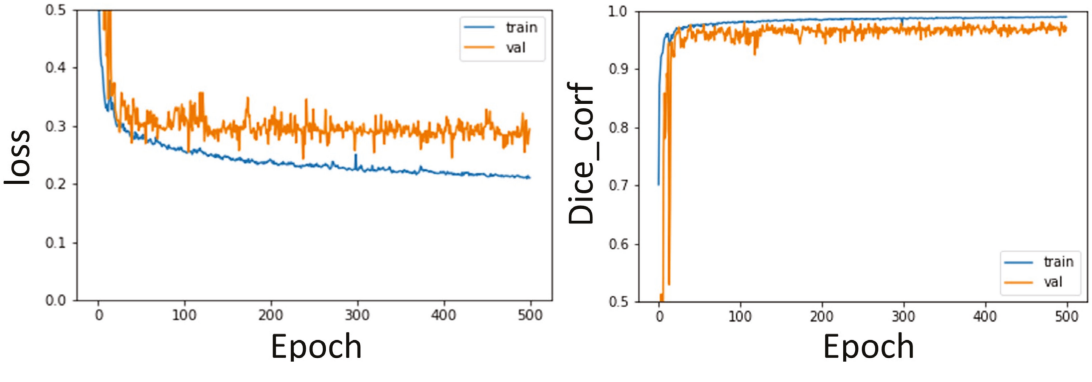


Figure 13. Aorta segmentation model training curve: Loss, Dice curve.

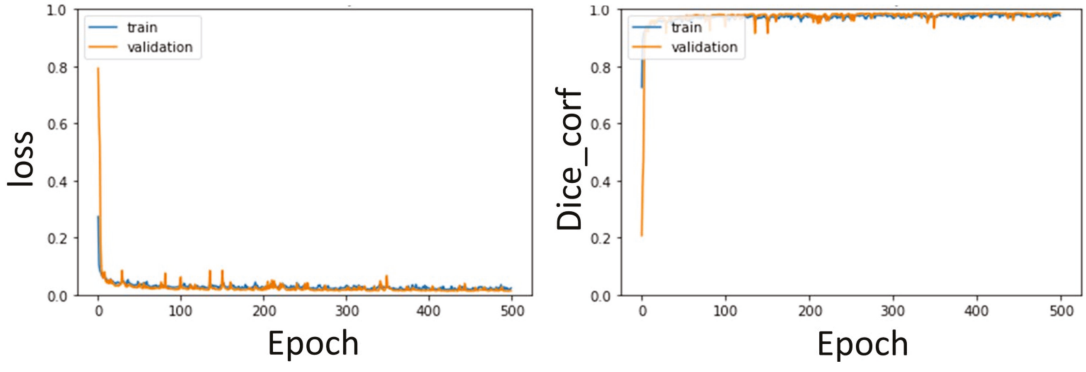


Figure 14. Pulmonary artery segmentation model training curve: Loss, Dice curve.

In the second stage of the architecture, the weights of the first stage are utilized by transfer learning, which can improve the learning performance of the model. Therefore, the segmentation results of the aorta and pulmonary artery by the method proposed in this study can achieve Dice coefficients of 0.97 ± 0.007 and 0.93 ± 0.002, respectively, after fivefold cross-validation.

In addition, this study compared the results of the proposed model with the unimproved 3D U-Net mentioned in Section 2.5, as shown in Table 2.

Table 2. Segmentation performance of the two-stage segmentation architecture.

Aorta		Pulmonary Artery	
Model	**DSC**	**Model**	**DSC**
1-AA	0.97 ± 0.007	1-PA	0.91 ± 0.002
		2-PA	0.93 ± 0.002
3D U-Net	0.87 ± 0.025	3D U-Net	0.87 ± 0.0004

1-AA, aorta segmentation model; 1-PA, one-channel pulmonary artery segmentation model by inputting non-contrast-enhanced image; 2-PA, two-channel model by inputting non-contrast-enhanced image and enhanced image; DSC, Dice similarity coefficient stage.

From the results, the two-stage DL segmentation model proposed in this study can efficiently complete the three-dimensional segmentation of the two major blood vessels, and for the difficult pulmonary artery, additional input imaging enhanced images can effectively improve segmentation performance. Among them, the pulmonary artery segmentation model adds contrast-enhanced images as the second channel, and it can be seen from the segmentation results that its performance is better than that of the single-channel input of only non-contrast-enhanced CT images. As shown in Table 2, the method proposed in this study is far superior to the 3D U-Net in either aortic segmentation or pulmonary artery segmentation.

4. Discussion

This study used fivefold cross-validation and DSC as the evaluation metrics, and the results are shown in Table 2. For aorta segmentation, the performance of this segmentation model was 0.97 ± 0.007, and it was only required to input non-contrast-enhanced CT images, which was in line with clinical use. Pulmonary artery segmentation is more difficult than aorta segmentation because of its complex vessel orientation. As shown in Table 2, the result of this model is 0.91 ± 0.002 when inputting only non-contrast-enhanced CT images, which is relatively poor. Therefore, this study adds the contrast-enhanced images obtained in the first stage of the architecture to improve segmentation performance. The result of this two-channel pulmonary artery segmentation model is 0.93 ± 0.002, which is approximately 0.02 higher than the input of only non-contrast-enhanced CT images.

To verify the effectiveness of the two-stage method in this study, we compared the segmentation performance of several previous studies, as shown in Table 3 [25–39]. In terms of aorta segmentation, the method proposed in this study achieved the highest segmentation performance, whereas in terms of pulmonary artery segmentation, it was only slightly inferior to the method developed by Gamechi et al. The method used in this study still has a high-precision segmentation performance.

Table 3. Comparison of segmentation performance between the method in this research method and those in previous research.

	Method	DSC
Aorta	2016 Jang et al. [25]	0.95 ± 0.02
	2009 Išgum et al. [26]	0.87 ± 0.03
	2012 Kurugol et al. [27]	0.93 ± 0.01
	2013 Avila-Montes et al. [28]	0.88 ± 0.05
	2017 Dasgupta et al. [29]	0.88 ± 0.06
	2014 Xie et al. [30]	0.93 ± 0.01
	2015 Kurugol et al. [31].	0.92 ± 0.01
	2019 Gamechi et al. [32]	0.95 ± 0.01
	2018 Noothout et al. [33]	0.91 ± 0.04
	2021 Lartaud et al. [34]	0.92 ± 0.02
	2020 Haq et al. [35]	$0.75 \leq DSC \leq 0.94$
	2020 Morris et al. [36]	0.85 ± 0.03
	2021 Sedghi Gamechi et al. [37]	0.96 ± 0.01
	Proposed method	0.97 ± 0.007
Pulmonary artery	2015 Xie et al. [38]	0.88
	2018 López-Linares et al. [39]	0.89 ± 0.07
	2020 Haq et al. [35]	$0.80 \leq DSC \leq 0.91$
	2020 Morris et al. [36]	0.85 ± 0.03
	2021 Sedghi Gamechi et al. [37]	0.94 ± 0.02
	Proposed method	0.93 ± 0.002

In the past, research on segmentation of the aorta and pulmonary artery on CT images has been conducted for many years. Therefore, this study also compared the performance of a previous study with that of the method used in this study. Previous studies on aorta segmentation mostly used images taken with a contrast agent for algorithm development. Compared with images taken with a non-contrast agent, those taken with a contrast agent has better vascular contrast and the vascular lumen presents a more obvious grayscale contrast with the surrounding tissue. Therefore, blood vessel segmentation is easier to perform. Jang et al. used CT images of contrast agents and proposed a method of automatic segmentation of the ascending aorta using geodesic distance transformation combined with Hough circles, which was applied to the diagnosis of cardiovascular diseases [25]. The proposed method outperforms this method in terms of segmentation performance based on non-contrast images; therefore, it is more competitive.

CT screening for early detection of thoracic cavity disease was performed without contrast agents. In addition, contrast injections should not be used in patients with allergy to these contrast media. Therefore, in recent years, most studies have been conducted on CT images without the use of contrast agents. In studies on aorta segmentation on non-contrast images, the following methods have been proposed based on prior knowledge of the vessel shape [26,28–30,32]. Isgum et al. proposed a multiple atlas-based segmentation method that registers multiple manually segmented atlas to the target image and uses decision fusion to obtain segmentation results [26]. Avila-Montes et al. proposed the extraction of the aorta centerline by Hough transform and dynamic programming and used the entropy-based cost function for boundary detection [28]. Dasgupta et al. used the circular Hough transform method to locate the vessel region to obtain aorta segmentation results [29]. Xie et al. proposed an algorithm that uses anatomy label maps and cylinder tracking to

segment the aorta [30]. Gamechi et al. combined multi-atlas registration to obtain seed points, aorta centerline extraction, and optimal surface segmentation to extract the aorta in non-contrast-enhanced CT images [32]. However, the extraction of the aorta centerline or boundary based on such shape priors is prone to errors in the locations where some vessels are narrowed, dilated, or where plaques appear.

In a study based on the active contour [40] method to segment the aorta, Kurugol et al. used the Frenet framework and 3D level set method to develop a fully automated and unsupervised segmentation of the aorta. The Dice coefficient was 0.93 ± 0.01. The aorta segmentation results can be used to quantify the degree of aorta calcification [27]. Kurugol et al. exploited the cross-sectional circularity of the aorta in axial slices and the aortic arch in reformatted oblique slices to detect initial aorta boundaries and used the 3D level-set method to modify the final results. The efficacy yields a Dice coefficient of 0.92 ± 0.01 [31]. Shown in Table 3, such active contour-based methods have a slightly higher average performance than those of studies that rely on shape priors.

In a study applying DL to aorta segmentation, Noothout et al. used a dilated convolutional neural network for segmentation. To obtain the final segmentation results, the probabilities obtained from the three planes were averaged per class. The Dice coefficients were 0.83 ± 0.07, 0.86 ± 0.06, and 0.88 ± 0.05 for the aorta arch and descending aorta, respectively, and 0.91 ± 0.04 for the aorta [33]. Lartaud et al. segmented multiple cardiac anatomical structures on spectral dual-energy CT images by using a multi-label U-Net, where a Dice coefficient of 0.92 ± 0.02 was obtained for the aorta [34]. However, the DL method relies on effective feature learning of the model or huge training data. Therefore, the aforementioned method has no obvious advantage over the traditional algorithm in terms of performance results. Based on a DL network, this research uses transfer learning through the first stage of the architecture. Thus, the effective learning of the model can be enhanced, and better segmentation performance can be obtained.

Related studies on pulmonary artery segmentation include the following: Moses et al. obtained a high correlation with the manually determined parameters for both mid-cross-sectional area (R = 0.96) and length (R = 0.93) [41]. Xie et al. used the shape before using the cylindrical registration method to segment the pulmonary artery and obtained the mean diameter according to the triangular mesh model and the anatomy label map [38]. Román et al. proposed a 3D convolutional neural network architecture, using realistic deformations to augment data, and obtained a Dice coefficient performance of 0.89 ± 0.07 for pulmonary artery segmentation in CT angiography images [39]. This shows that segmentation of the pulmonary artery remains a challenge even with contrast imaging.

In the aforementioned studies, only one of the segmentation targets required in this study was discussed, which could not meet the needs of this study to be applied to the study of pulmonary hypertension. The following studies have discussed the segmentation of both the aorta and pulmonary artery. Haq et al. established and validated a multi-label DL segmentation model for 2D segmentation for automatic segmentation of 12 cardiopulmonary substructures, including the aorta and pulmonary artery, with segmentation efficiencies of $0.80 \leq \text{DSC} \leq 0.91$ and 0.75, respectively, $0.75 \leq \text{DSC} \leq 0.94$ [35]. Morris et al. used a 3D U-Net to segment multiple structures of the heart and post-processed them using 3D-CRF. The result of the segmentation Dice coefficient of the aorta and pulmonary aorta, collectively called Great Vessels was 0.85 ± 0.03. However, this study requires simultaneous input of CT and magnetic resonance imaging images, which are difficult to obtain simultaneously under normal conditions [36]. Sedghi Gamechi et al. proposed to cut the centerline based on the optimal surface map, and the Dice coefficient of the segmentation result can be obtained as 0.94 ± 0.02 for the pulmonary artery and 0.96 ± 0.01 for the aorta [37]. However, this study is still based on shape priors; therefore, it is easy to encounter the aforementioned problems.

In Table 3, the performance comparison results also show that the segmentation models proposed in this study are superior to other methods in aortic segmentation and only slightly inferior to those of the segmentation algorithms developed based on traditional

methods in pulmonary artery segmentation. The method proposed in this study is a DL model; therefore, it is more generalized and robust than traditional methods with good architectural design and training. In the comparison of previous related studies that also used DL models, it can also be seen from Table 3 that the method proposed in this study achieved the best segmentation performance of the aorta and pulmonary artery among the related DL methods.

Several studies have shown a correlation between preoperative pulmonary hypertension and postoperative complications [12,14]. The gold standard approach for the diagnosis of pulmonary hypertension is the direct measurement of pulmonary artery pressure by cardiac catheterization [12,14]. However, this invasive measurement method may not be commonly used for preoperative evaluation of patients with lung cancer. Consequently, owing to the relationship between elevated pulmonary artery pressure and vessel diameter, recent studies have shown the correlation of enlarged pulmonary artery to postoperative complications [42]. However, the method used in the previous studies was 2D measurement of single-cut axial view contrast-enhanced computed tomography image. Automatic 3D segmentation method for both the aorta and pulmonary artery on CT images to accurately calculate the mean 3D diameter has not been reported before. Our proposed model may automatically provide 3D diameter information of the aorta and pulmonary artery before surgery, facilitating the estimation of pulmonary hypertension and supporting preoperative surgical method selection based on the predicted surgical risks.

This study has the following limitations. In the pulmonary artery segmentation model, the contrast enhancement model developed in the first stage of this architecture still needs to be used to provide contrast-enhanced images as inputs for clinical applications. Therefore, this model is more time-consuming and energy consuming than the aorta segmentation model, and the input and model construction methods of this model can be further improved in the future. Second, the types and quantities of data used in this study need to be expanded and increased so that the model in this study can achieve more effective generalization capabilities and improve the applicability of this study model on various non-contrast-enhanced CT images, such as low-dose CT. Third, there are still many artificial parameter settings in the calculation of the diameters of the two blood vessels, which can be further improved into a more automated extraction method.

5. Conclusions

To overcome the difficulty of segmenting non-imaging CT images of the aorta and pulmonary artery, this study proposes a two-stage DL segmentation architecture consisting of a contrast enhancement model and segmentation model. This method uses transfer learning to enhance the performance of the segmentation model. The DL method proposed in this study can efficiently complete the segmentation of the aorta and pulmonary artery. Compared with previous research methods for aorta and pulmonary artery segmentation, this study can achieve a high level of segmentation performance. In conclusion, the proposed model may provide the 3D diameter information of two arteries before surgery, facilitating the estimation of pulmonary hypertension and supporting the preoperative surgical method selection based on the predicted surgical risks.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics12040967/s1, Table S1: Patient clinic pathological features and preoperative results.

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

Funding: This research was funded by the Ministry of Science and Technology, Taiwan (MOST 107-2221-E-002-074-MY3, 107-2221-E-002-080-MY3) and National Taiwan University Hospital, Taipei, Taiwan (NTUH 111-S0199).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Research Ethics Committee of the National Taiwan University Hospital (protocol code: 201712087RIND; date of approval on 23 January 2018).

Informed Consent Statement: The Research Ethics Committee of the National Taiwan University Hospital approved this study and waived the need for informed consent because of the retrospective study design.

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

References

1. de Koning, H.J.; van der Aalst, C.M.; de Jong, P.A.; Scholten, E.T.; Nackaerts, K.; Heuvelmans, M.A.; Lammers, J.W.J.; Weenink, C.; Yousaf-Khan, U.; Horeweg, N.; et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N. Engl. J. Med.* **2020**, *382*, 503–513. [CrossRef] [PubMed]
2. Lin, M.-W.; Tseng, Y.-H.; Lee, Y.-F.; Hsieh, M.-S.; Ko, W.-C.; Chen, J.-Y.; Hsu, H.-H.; Chang, Y.-C.; Chen, J.-S. Computed tomography-guided patent blue vital dye localization of pulmonary nodules in uniportal thoracoscopy. *J. Thorac. Cardiovasc. Surg.* **2016**, *152*, 535–544.e2. [CrossRef] [PubMed]
3. Chen, P.-H.; Hsu, H.-H.; Yang, S.-M.; Tsai, T.-M.; Tsou, K.-C.; Liao, H.-C.; Lin, M.-W.; Chen, J.-S. Preoperative dye localization for thoracoscopic lung surgery: Hybrid versus computed tomography room. *Ann. Thorac. Surg.* **2018**, *106*, 1661–1667. [CrossRef]
4. Ginsberg, R.J.; Rubinstein, L.V.; Group, L.C.S. Randomized trial of lobectomy versus limited resection for T1 N0 non-small cell lung cancer. *Ann. Thorac. Surg.* **1995**, *60*, 615–623. [CrossRef]
5. Chiang, X.-H.; Hsu, H.-H.; Hsieh, M.-S.; Chang, C.-H.; Tsai, T.-M.; Liao, H.-C.; Tsou, K.-C.; Lin, M.-W.; Chen, J.-S. Propensity-matched analysis comparing survival after sublobar resection and lobectomy for cT1N0 lung adenocarcinoma. *Ann. Surg. Oncol.* **2020**, *27*, 703–715. [CrossRef] [PubMed]
6. Lin, M.-W.; Kuo, S.-W.; Yang, S.-M.; Lee, J.-M. Robotic-assisted thoracoscopic sleeve lobectomy for locally advanced lung cancer. *J. Thorac. Dis.* **2016**, *8*, 1747–1752. [CrossRef] [PubMed]
7. Lin, Y.-J.; Chiang, X.-H.; Lu, T.-P.; Hsieh, M.-S.; Lin, M.-W.; Hsu, H.-H.; Chen, J.-S. Thoracoscopic Lobectomy Versus Sublobar Resection for pStage I Geriatric Non-Small Cell Lung Cancer. *Front. Oncol.* **2021**, *11*, 11. [CrossRef]
8. Kagimoto, A.; Tsutani, Y.; Kushitani, K.; Kai, Y.; Kambara, T.; Miyata, Y.; Takeshima, Y.; Okada, M. Segmentectomy vs Lobectomy for Clinical Stage IA Lung Adenocarcinoma with Spread Through Air Spaces. *Ann. Thorac. Surg.* **2021**, *112*, 935–943. [CrossRef]
9. Hu, S.-Y.; Hsieh, M.-S.; Hsu, H.-H.; Tsai, T.-M.; Chiang, X.-H.; Tsou, K.-C.; Liao, H.-C.; Lin, M.-W.; Chen, J.-S. Correlation of tumor spread through air spaces and clinicopathological characteristics in surgically resected lung adenocarcinomas. *Lung Cancer* **2018**, *126*, 189–193. [CrossRef]
10. Lin, M.-W.; Su, K.-Y.; Su, T.-J.; Chang, C.-C.; Lin, J.-W.; Lee, Y.-H.; Yu, S.-L.; Chen, J.-S.; Hsieh, M.-S. Clinicopathological and genomic comparisons between different histologic components in combined small cell lung cancer and non-small cell lung cancer. *Lung Cancer* **2018**, *125*, 282–290. [CrossRef]
11. Li, C.; Kuo, S.-W.; Hsu, H.-H.; Lin, M.-W.; Chen, J.-S. Lung adenocarcinoma with intraoperatively diagnosed pleural seeding: Is main tumor resection beneficial for prognosis? *J. Thorac. Cardiovasc. Surg.* **2018**, *155*, 1238–1249.e1. [CrossRef] [PubMed]
12. Ramakrishna, G.; Sprung, J.; Ravi, B.S.; Chandrasekaran, K.; McGoon, M.D. Impact of pulmonary hypertension on the outcomes of noncardiac surgery: Predictors of perioperative morbidity and mortality. *J. Am. Coll. Cardiol.* **2005**, *45*, 1691–1699. [CrossRef] [PubMed]
13. Crabtree, T.; Puri, V.; Timmerman, R.; Fernando, H.; Bradley, J.; Decker, P.A.; Paulus, R.; Putnum Jr, J.B.; Dupuy, D.E.; Meyers, B. Treatment of stage I lung cancer in high-risk and inoperable patients: Comparison of prospective clinical trials using stereotactic body radiotherapy (RTOG 0236), sublobar resection (ACOSOG Z4032), and radiofrequency ablation (ACOSOG Z4033). *J. Thorac. Cardiovasc. Surg.* **2013**, *145*, 692–699. [CrossRef]
14. Wei, B.; D'Amico, T.; Samad, Z.; Hasan, R.; Berry, M.F. The impact of pulmonary hypertension on morbidity and mortality following major lung resection. *Eur. J. Cardio-Thorac. Surg.* **2014**, *45*, 1028–1033. [CrossRef] [PubMed]
15. Chung, M.; Lewis, E.; Yip, R.; Jirapatnakul, A.; Reeves, A.; Yankelevitz, D.; Henschke, C.; Bhora, F. P2. 16-023 Changes of the Pulmonary Artery After Resection of Stage I Lung Cancer. *J. Thorac. Oncol.* **2017**, *12*, S2197. [CrossRef]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin, Germany, 2015; pp. 234–241.
17. Siddique, N.; Paheding, S.; Elkin, C.P.; Devabhaktuni, V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* **2021**, *9*, 82031–82057. [CrossRef]
18. Bi, L.; Feng, D.; Kim, J. Dual-path adversarial learning for fully convolutional network (FCN)-based medical image segmentation. *Vis. Comput.* **2018**, *34*, 1043–1052. [CrossRef]

19. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; Springer: Berlin, Germany, 2016; pp. 424–432.
20. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
21. Chen, L.; Liang, X.; Shen, C.; Jiang, S.; Wang, J. Synthetic CT generation from CBCT images via deep learning. *Med. Phys.* **2020**, *47*, 1115–1125. [CrossRef]
22. Brunet, D.; Vrscay, E.R.; Wang, Z. On the mathematical properties of the structural similarity index. *IEEE Trans. Image Processing* **2011**, *21*, 1488–1499. [CrossRef]
23. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. In Proceedings of the International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; Springer: Berlin, Germany, 2018; pp. 270–279.
24. Saha, P.K.; Borgefors, G.; di Baja, G.S. A survey on skeletonization algorithms and their applications. *Pattern Recognit. Lett.* **2016**, *76*, 3–12. [CrossRef]
25. Jang, Y.; Jung, H.Y.; Hong, Y.; Cho, I.; Shim, H.; Chang, H.-J. Geodesic distance algorithm for extracting the ascending aorta from 3D CT images. *Comput. Math. Methods Med.* **2016**, *2016*, 4561979. [CrossRef] [PubMed]
26. Isgum, I.; Staring, M.; Rutten, A.; Prokop, M.; Viergever, M.A.; Van Ginneken, B. Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans. *IEEE Trans. Med. Imaging* **2009**, *28*, 1000–1010. [CrossRef]
27. Kurugol, S.; Estepar, R.S.J.; Ross, J.; Washko, G.R. Aorta segmentation with a 3D level set approach and quantification of aortic calcifications in non-contrast chest CT. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; IEEE: New York, NY, USA, 2012; pp. 2343–2346.
28. Avila-Montes, O.C.; Kurkure, U.; Nakazato, R.; Berman, D.S.; Dey, D.; Kakadiaris, I.A. Segmentation of the thoracic aorta in noncontrast cardiac CT images. *IEEE J. Biomed. Health Inform.* **2013**, *17*, 936–949. [CrossRef] [PubMed]
29. Dasgupta, A.; Mukhopadhyay, S.; Mehre, S.A.; Bhattacharyya, P. Morphological geodesic active contour based automatic aorta segmentation in thoracic CT images. Proceedings of International Conference on Computer Vision and Image Processing, Roorkee, India, 9–12 September 2017; Springer: Berlin, Germany, 2017; pp. 187–195.
30. Xie, Y.; Padgett, J.; Biancardi, A.M.; Reeves, A.P. Automated aorta segmentation in low-dose chest CT images. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 211–219. [CrossRef] [PubMed]
31. Kurugol, S.; Come, C.E.; Diaz, A.A.; Ross, J.C.; Kinney, G.L.; Black-Shinn, J.L.; Hokanson, J.E.; Budoff, M.J.; Washko, G.R.; San Jose Estepar, R. Automated quantitative 3D analysis of aorta size, morphology, and mural calcification distributions. *Med. Phys.* **2015**, *42*, 5467–5478. [CrossRef]
32. Sedghi Gamechi, Z.; Bons, L.R.; Giordano, M.; Bos, D.; Budde, R.P.; Kofoed, K.F.; Pedersen, J.H.; Roos-Hesselink, J.W.; de Bruijne, M. Automated 3D segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced CT. *Eur. Radiol.* **2019**, *29*, 4613–4623. [CrossRef] [PubMed]
33. Noothout, J.M.; De Vos, B.D.; Wolterink, J.M.; Išgum, I. Automatic segmentation of thoracic aorta segments in low-dose chest CT. In *Medical Imaging 2018: Image Processing*; SPIE: Washington, DC, USA, 2018; p. 105741S.
34. Lartaud, P.-J.; Hallé, D.; Schleef, A.; Dessouky, R.; Vlachomitrou, A.S.; Douek, P.; Rouet, J.-M.; Nempont, O.; Boussel, L. Spectral augmentation for heart chambers segmentation on conventional contrasted and unenhanced CT scans: An in-depth study. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 1699–1709. [CrossRef]
35. Morris, E.D.; Ghanem, A.I.; Dong, M.; Pantelic, M.V.; Walker, E.M.; Glide-Hurst, C.K. Cardiac substructure segmentation with deep learning for improved cardiac sparing. *Med. Phys.* **2020**, *47*, 576–586. [CrossRef]
36. Sedghi Gamechi, Z.; Arias-Lorza, A.M.; Saghir, Z.; Bos, D.; de Bruijne, M. Assessment of fully automatic segmentation of pulmonary artery and aorta on noncontrast CT with optimal surface graph cuts. *Med. Phys.* **2021**, *48*, 7837–7849. [CrossRef]
37. Moses, D.; Sammut, C.; Zrimec, T. Automatic segmentation and analysis of the main pulmonary artery on standard post-contrast CT studies using iterative erosion and dilation. *Int. J. Comput. Assist. Radiol. Surg.* **2016**, *11*, 381–395. [CrossRef] [PubMed]
38. López-Linares Román, K.; Bruere, I.D.L.; Onieva, J.; Andresen, L.; Qvortrup Holsting, J.; Rahaghi, F.N.; Macía, I.; González Ballester, M.A.; San José Estepar, R. 3D pulmonary artery segmentation from CTA scans using deep learning with realistic data augmentation. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*; Springer: Berlin, Germany, 2018; pp. 225–237.
39. Haq, R.; Hotca, A.; Apte, A.; Rimner, A.; Deasy, J.O.; Thor, M. Cardio-pulmonary substructure segmentation of radiotherapy computed tomography images using convolutional neural networks for clinical outcomes analysis. *Phys. Imaging Radiat. Oncol.* **2020**, *14*, 61–66. [CrossRef] [PubMed]
40. Chan, T.; Vese, L. An active contour model without edges. In Proceedings of the International Conference on Scale-Space Theories in Computer Vision, Corfu, Greece, 26–27 September 1999; Springer: Berlin, Germany, 1999; pp. 141–151.
41. Xie, Y.; Liang, M.; Yankelevitz, D.F.; Henschke, C.I.; Reeves, A.P. Automated measurement of pulmonary artery in low-dose non-contrast chest CT images. In *Medical Imaging 2015: Computer-Aided Diagnosis*; SPIE: Washington, DC, USA, 2015; pp. 375–383.
42. Asakura, K.; Mitsuboshi, S.; Tsuji, M.; Sakamaki, H.; Otake, S.; Matsuda, S.; Kaseda, K.; Watanabe, K. Pulmonary arterial enlargement predicts cardiopulmonary complications after pulmonary resection for lung cancer: A retrospective cohort study. *J. Cardiothorac. Surg.* **2015**, *10*, 113. [CrossRef] [PubMed]

diagnostics

MDPI

Article

A Rotational Invariant Neural Network for Electrical Impedance Tomography Imaging without Reference Voltage: RF-REIM-NET

Jöran Rixen [1,*], Benedikt Eliasson [1], Benjamin Hentze [1,2], Thomas Muders [2], Christian Putensen [2], Steffen Leonhardt [1] and Chuong Ngo [1]

[1] Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, 52074 Aachen, Germany; benedikt.eliasson@rwth-aachen.de (B.E.); hentze@hia.rwth-aachen.de (B.H.); leonhardt@hia.rwth-aachen.de (S.L.); ngo@hia.rwth-aachen.de (C.N.)

[2] Department of Anaesthesiology and Intensive Care Medicine, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany; thomas.muders@ukbonn.de (T.M.); christian.putensen@ukbonn.de (C.P.)

* Correspondence: rixen@hia.rwth-aachen.de

Abstract: *Background*: Electrical Impedance Tomography (EIT) is a radiation-free technique for image reconstruction. However, as the inverse problem of EIT is non-linear and ill-posed, the reconstruction of sharp conductivity images poses a major problem. With the emergence of artificial neural networks (ANN), their application in EIT has recently gained interest. *Methodology*: We propose an ANN that can solve the inverse problem without the presence of a reference voltage. At the end of the ANN, we reused the dense layers multiple times, considering that the EIT exhibits rotational symmetries in a circular domain. To avoid bias in training data, the conductivity range used in the simulations was greater than expected in measurements. We also propose a new method that creates new data samples from existing training data. *Results*: We show that our ANN is more robust with respect to noise compared with the analytical Gauss–Newton approach. The reconstruction results for EIT phantom tank measurements are also clearer, as ringing artefacts are less pronounced. To evaluate the performance of the ANN under real-world conditions, we perform reconstructions on an experimental pig study with computed tomography for comparison. *Conclusions*: Our proposed ANN can reconstruct EIT images without the need of a reference voltage.

Keywords: artificial intelligence; deep learning; Electrical Impedance Tomography; lung imaging; cardiopulmonary monitoring

Citation: Rixen, J.; Eliasson, B.; Hentze, B.; Muders, T.; Putensen, C.; Leonhardt, S.; Ngo, C. A Rotational Invariant Neural Network for Electrical Impedance Tomography Imaging without Reference Voltage: RF-REIM-NET. *Diagnostics* **2022**, *12*, 777. https://doi.org/10.3390/diagnostics12040777

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 11 February 2022
Accepted: 19 March 2022
Published: 22 March 2022

1. Introduction

Electrical Impedance Tomography (EIT) enables the non-invasive visualization of the dielectric properties of a medium of interest. EIT has a wide range of applications, including the status monitoring of concrete [1], the monitoring of semiconductor manufacturing [2], and observing cell cultures [3]. In the medical domain, the applications are broader, and include the monitoring of lung recruitment and collapse [4], lung ventilation [5] and perfusion, the monitoring of 3D brain activity [6], size and volume estimation of the bladder [7], breast cancer imaging [8], and cardiopulmonary monitoring [9]. Here, EIT can be used to assess metrics such as regional ventilation, end-expiratory lung volume, compliance, regional respiratory system compliance, and regional pressure–volume curves [9].

The versatility of EIT stems from the fact that the measurements can be made non-invasively and inexpensively. For an image to be reconstructed, electrodes need to be placed around the domain. Small, low-frequency currents in the range of 100 kHz are fed through these electrodes. Then, the voltage across the electrodes is measured, and an image is reconstructed. Despite the advantages of EIT, it has one major drawback: it suffers from a relatively low spatial resolution.

This issue is due to the fact that EIT image reconstruction belongs to the class of inverse problems [10]. Large changes in the conductivity of the medium may lead to only small changes in the voltage measurements. To still be able to solve the problem, different types of algorithms have been proposed in the literature. From a mathematical perspective, three different types of algorithms can be distinguished.

The first set of algorithms is variational regularization methods. Their goal is to minimize a cost function that contains two parts. First, the physical behavior of the medium of interest is modeled. Given a set of voltage measurements, the algorithm helps to find the best fit for the conductivities that could produce these voltages. Second, the regularization strategy is applied, which plays a crucial role in finding a valid solution. Two common examples of regularization strategies are total variation [11] and the Tikhonov regularization [12].

The second type of algorithms is statistical inversion methods. Here, image reconstruction is modeled as a problem of statistical inference. The measurements and conductivities are modeled as random variables from which an a posteriori distribution can be estimated, through, e.g., Markov Chain Monte Carlo iterations. From this, the conductivity can be derived [13]. This can be accomplished, by, for example, first obtaining a starting distribution through the one-step Gauss–Newton method. Thereafter, Markov Chain Monte Carlo methods can be used to refine the starting distribution [14].

The final type is direct inversion algorithms. In these methods, the problem is analyzed through the partial differential equations governing the system behavior. From this, a solution strategy is developed. An example of these kinds of methods is the D-Bar algorithm [15].

Artificial neural networks belong to the variational regularization methods, as they solve the optimization problem once during training and then act like a complex look-up table. The regularization performed by artificial neural networks is not straightforward: first, the neural network architecture provides a part of the regularization. A very deep architecture may provide sophisticated results for the training data set, but may lead to profound over fitting, such that the results for slightly different data bring far worse results. The second part of the regularization comes from the training data. There is no reference technique to capture the conductivity distribution of body tissue. Thus, in EIT the training data are simulated with the help of, for example, finite element method (FEM) software such as EIDORS [16]. However, for simulations a multitude of assumptions have to be made: What does the model shape look like? What are the electrode positions? Do they change? What shape do the conductivity enclosures have? What is the range of conductivity? All of these assumptions act as some kind of regularization.

Artificial neural networks are beginning to gain more relevance in the field of EIT. In 2017, Kłosowski and Rymarczyk [17] presented an ANN with fully connected layers and convolutional layers. However, the proposed ANN can only reconstruct single targets. Their outputs are the coordinates and the radius of the conductivity enclosure. Other approaches used ANNs to enhance the reconstructions of traditional EIT reconstructions [18]. In 2019, Hu et al. [19] used the spatial invariant properties of the EIT to improve upon these results. However, to aid in the reconstruction, their approach is based on calibration. Thus, their artificial neural network is not usable when the background data are missing. By contrast, Chan et al. [20] proposed a network which does not need this preprocessing. However, the structure of the artificial neural network does not account for the symmetry of EIT measurements. We settled for artificial neural networks, as they have been used in the past within the domain of EIT and show the greatest potential due to their ability to recreate non-linear functions.

In the following, we propose an artificial neural network structure which can reconstruct images without dependence on a reference voltage, while still using the rotational symmetry of EIT adjacent measurements in adjacent drive. We call this structure the **R**eference **F**ree **R**otational **E**lectrical **I**mpedance **M**ap **Net**work (RF-REIM-NET).

The novelties of this research are:

1. We use real-world animal trial data with CT references to confirm that RF-REIM-NET gives meaningful results in such a setting.
2. Our training data are unbiased, as we used a conductivity range bigger than what is expected in the thorax region and did not try to model the conductivity distributions typically encountered in the thorax region.
3. We present a method for time-effective data augmentation using the existing training data.
4. Even though RF-REIM-NET uses fully connected layers, it still preserves the rotational invariance of adjacent measurements.

2. Materials and Methods

2.1. Fundamentals

In EIT, the goal is to find an optimal conductivity distribution given a set of voltage measurements. When using variational reconstruction methods, this is expressed as

$$\boldsymbol{\sigma}_{rec} = \arg\min \frac{1}{2}\|F(\boldsymbol{\sigma}) - \boldsymbol{V}_{meas}\|^2 + \lambda\|\boldsymbol{L}\boldsymbol{\sigma}\|^2, \tag{1}$$

where $\boldsymbol{\sigma}$ is the conductivity, $F(\boldsymbol{\sigma})$ is the forward model, $\boldsymbol{V}_{meas}$ is the measured voltage, λ is the weighting of the regularization term and $\boldsymbol{L}$ is the regularization matrix. When using artificial neural networks (ANNs), the general scheme of this minimization is still true; however, it is achieved differently. While variational reconstruction methods minimize each measurement according to Equation (1), ANNs will perform the minimization on a given dataset. This can be formulated as

$$\arg\min \frac{1}{2}\left\|\boldsymbol{Y}'(\boldsymbol{V}_{meas}) - \boldsymbol{Y}\right\|^2, \tag{2}$$

where $\boldsymbol{Y}$ denotes the ground truth value and $\boldsymbol{Y}'(\boldsymbol{V}_{meas})$ denotes the ANNs output depending on the input of the network. During runtime, the ANN behaves deterministically like a look-up table. When assuming that the dataset represents real measurements, the ANN still minimizes the actual measurements.

2.2. Electrical Impedance Maps

Hu et al., pointed out the advantages of packing EIT measurements into the electrical impedance map (EIM). EIMs can be used to represent EIT data in adjacent–adjacent measurement mode. For 16 electrodes, the data are represented in a 16×16 matrix; see Figure 1. Along the matrix column are the measurement electrodes, while the excitation electrodes are arranged along the rows. $\boldsymbol{EIM}[j,k]$ contains the measurement of the jth electrode pair, while the kth electrode pair drives the current. Since four probe measurements are used, voltages from injecting electrodes cannot be used. On those spots, the $\boldsymbol{EIM}$ matrix is filled with zeros, causing the superdiagonal, diagonal and subdiagonal elements to become 0.

When a conductivity distribution is rotated by an angle of $\frac{2\pi k}{16}$, where k is an integer number, the features of the $\boldsymbol{EIM}$ map do not change. The features are moved diagonally across the image. Thus, a convolutional ANN can extract features from the EIM independent of rotation.

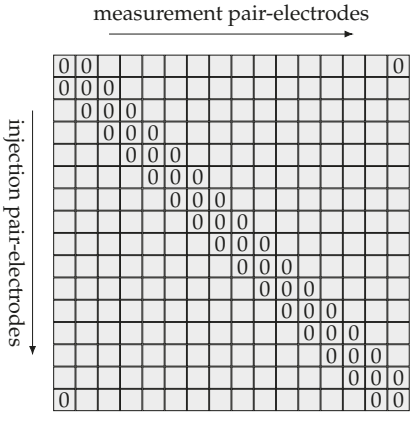


Figure 1. A 16 × 16 electrical impedance map (EIM) arrangement from an adjacent injection pattern. The zeros represent values which are not gathered from adjacent–adjacent measurement mode, as at least one electrode is used for current injection.

2.3. Training Data Set

In machine learning, the data set used for training is an important part of the algorithm's performance [21]. For EIT, there is no general high-resolution ground truth dataset. Instead, the data have to be carefully designed. It is easy and tempting to craft a data set that gives meaningful results on the available test data. If there is a relatively narrow band of possible conductivities in the test data, using this conductivity band in the simulated training data would bring a bias to the network—it might look better than it actually is.

To avoid this fallacy, we designed our data with as few assumptions as possible. We used FEM simulations to create the training data. These simulations were executed using EIDORS [16]. The first practical constraint faced was simulation time, and in general, higher mesh density is better for the quality of the simulations. However, the time taken for meshing and actual computation increases non-linearly. Thus, we used a mesh density of 0.075, while the model radius was chosen to be 28, as this is a feasible trade-off between simulation quality and computation time. Our domain shape was cylindrical. We used 16 electrodes, each with a height of 40 mm and a width of 20 mm. The electrodes were placed equidistantly around the domain. This setup was chosen as it imitates the typical measurement configuration of the clinically available device for thoracic images from Draeger (*Draeger Pulmo Vista 500*, Draeger Medical GmbH, Lübeck, Germany).

2.3.1. Basic Object Shapes

To create conductivity enclosures, we used three different basic shapes: an ellipsoid, a cube and an octahedron. The basic size of these objects is 1 in all directions, and their center of gravity is in the origin of the coordinate system. To save computation time, we did not re-mesh each impedance enclosure from scratch. Instead, we created a mask for each conductivity enclosure and then changed the conductivities of mesh elements inside the mask m. The formulas for the three basic shapes are given as:

$$mask_{\text{sphere}} = (x^2 + y^2 + z^2) \leq 1 \tag{3}$$

$$mask_{\text{cube}} = \max(|x|, |y|, |z|) \leq 1 \tag{4}$$

$$mask_{\text{octahedron}} = |x| + |y| + |z| \leq 1 \tag{5}$$

2.3.2. Transformation of the Basic Objects

Only inserting the same shape at the same place in the FEM model would be of no use for real-world reconstructions. Thus, the basic shapes have to be transformed. Our transformation involves the translation, rotation and scaling of the enclosures. This can be mathematically described as:

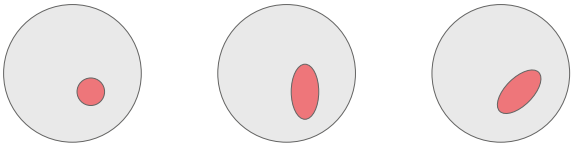
$$\boldsymbol{v}' = (\boldsymbol{R}(\boldsymbol{v} - \boldsymbol{t})) \oslash \boldsymbol{s}, \tag{6}$$

where $\boldsymbol{v} = \{x, y, z\}$ is the coordinate vector, $\boldsymbol{v}'$ is the transformed coordinate vector, $\boldsymbol{t} \in \mathbb{R}^3$ is the translation vector, $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ is the rotation matrix, $\boldsymbol{s} \in \mathbb{R}^3$ is the scaling vector and $\oslash$ denotes the element-wise division.

The positioning of the enclosures is important. As the ANN should be able to detect any conductivity enclosures in the domain with the same quality, the distribution of the object's center of gravity should be uniform across the domain. Thus, we sampled the values of $\boldsymbol{t}$ from a uniform distribution, such that every component of $\boldsymbol{t} = \{x, y, z\}$ is well inside the domain boundaries. Figure 2 gives a visual example of the transformations applied to the data.

Each entry of the scaling vector $\boldsymbol{s}$ is uniformly sampled between 10% and 80% of the model radius.

The angle of the rotation matrix $\boldsymbol{R}$ is uniformly sampled from $[0, 2\pi)$, and thus the basic shape can be rotated in any direction. This enables the ANN to learn features that are valid for a variety of positions, as only the position of the feature changes. In Figure 2, the transformation is visualized.



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

Figure 2. An example of the random transformation used in dataset generation, applied to a circular shape. (**a**) Circle with offset. (**b**) Circle with offset and scaling. (**c**) Circle with offset, scaling and rotation.

2.3.3. Conductivity Range

Another important degree of freedom is the conductivity range used. When using EIT tanks for testing, the conductivity range of the test data is typically known. Thus, it would be very tempting to just use this conductivity range for the training of the ANN. However, in practice, the conductivity is typically not known to this level of detail. Through a slice of the chest, conductivity values can range from 3.5×10^{-3} S/m (cortical bone) up to 4.64×10^{-1} S/m (deflated lung) [22]. We used a range of 1×10^{-5} S/m to 1 S/m for the background conductivity, as this covers the conductivity values typically encountered in chest measurements, while at the same time providing a margin well outside to improve generalization. The values were sampled uniformly from a logarithmic arrangement of the mentioned conductivity range. This is also known as a reciprocal or log-uniform distribution. This was chosen because the ANN should be able to differentiate objects that are one order of magnitude bigger than the background, regardless of the actual background conductivity.

The next step is an appropriate choice of the conductivity enclosures. As mentioned, it is important that the ANN is able to distinguish conductivity contrasts. At the same time, the ANN shall also be able to distinguish those contrasts symmetrically in the lower and upper bound of the conductivity range. To achieve this, the enclosure's conductivity is chosen with respect to the background conductivity. This ensures that the ANN has

no bias towards a conductivity contrast higher or lower than the background. Thus, the enclosure conductivity is chosen by multiplying the background conductivity with values from a range of 1×10^{-2} to 1×10^{2}. Again, this is sampled uniformly from the logarithmic arrangement of those values.

In a real-world setting, conductivities are rarely perfectly homogeneous across a tissue type. Because of this, the enclosures, as well as the background, are perturbed. We again scale each node of the FEM model by different values. This is achieved by using a Gaussian distribution with a *mean* value of 1. For each training sample and chosen conductivity value, a different standard deviation (*std*) from 1×10^{-8} to 1×10^{-2} was chosen. When the *std* is chosen, the values of one conductivity are perturbed by multiplication with the sampled values.

2.3.4. Electrode Contact Impedance

Another effect to consider is the electrode contact impedance. Although the adjacent drive pattern used here relies solely on four probe measurements and, thus, will reduce the effect of electrode contact impedance, we included the effect into our training data. We multiplied EIDORS default contact impedance by a value randomly sampled from a Gaussian distribution with a *mean* value of 1. To simulate high and low differences, we sampled the values from three different distributions with an *std* of 1×10^{-5}, 1×10^{-3} and 1×10^{-1}.

2.3.5. Measurement Noise

ANNs typically struggle with generalizing learned samples to cases that the ANN has not yet seen [23]. To tackle this problem, further data augmentation strategies need to be used. While the previously mentioned steps required new simulations for each training sample, the following steps rely on already simulated data. This saves computation time.

EIT measurements can be affected by several sources of noise. Paired with the ill-conditioned nature of the EIT problem, this can cause artifacts in the reconstruction. Often, reconstruction algorithms have a hyperparameter, which in essence balances the robustness to noise and the quality of the reconstruction. As for ANNs, the sensitivity to noise can be adjusted through the noise in the training data.

A major component in EIT systems is the analog digital converter; the noise consists primarily of thermal, jitter, and quantization noise [24]. The first two depend on the magnitude of the signals. The greater the signal, the bigger the noise. We can model this by multiplying the noise-free signal with a constant drawn from Gaussian noise:

$$U^{\mathrm{i,j}}_{\mathrm{ther,jit}} = U^{\mathrm{i,j}} \cdot n_{\mathrm{mult}}, \quad n_{\mathrm{mult}} \sim \mathcal{N}(1, \sigma^2), \forall \mathrm{i,j} \in \{1, \ldots, 16\} \tag{7}$$

where $U^{\mathrm{i,j}}_{\mathrm{ther,jit}}$ is the thermal noise-affected measurement, $U^{\mathrm{i,j}}$ is the noise-free measurement and n_{mult} is the noise sampled from a normal distribution. The quantization noise does not depend on the signal level, and can be modeled by adding a noise term to the voltage signals:

$$U^{\mathrm{i,j}}_{\mathrm{quant}} = U^{\mathrm{i,j}} + n_{\mathrm{add}}, \quad n_{\mathrm{add}} \sim \mathcal{N}(0, \sigma^2), \forall \mathrm{i,j} \in \{1, \ldots, 16\} \tag{8}$$

where $U^{\mathrm{i,j}}_{\mathrm{quant}}$ is the quantization noise-affected measurement and n_{add} is the noise sampled from a normal distribution. However, there is still another source of noise. Different measurement channels of a given EIT system can have different gains. This is due to different gains in the multiplexers [25]. The noise can be described through

$$U^{\mathrm{i,j}}_{\mathrm{gain}} = U^{\mathrm{i,j}} \cdot n_{\mathrm{mult}}, \quad n_{\mathrm{mult}} \sim \mathcal{N}(0, \sigma^2), \forall \mathrm{j} \in \{1, \ldots, 16\} \tag{9}$$

Note that this noise only affects one row of the EIM, compared with Equation (7), where every entry is affected individually. For the additive noise, n_{add} a *std* of 1×10^{-8} was chosen. For the multiplicative noise, n_{mult}, an *std* of 1×10^{-6} was chosen.

2.3.6. Rotation of the Data

To specifically incorporate the rotational invariance into the ANN, the voltage data were prepared with minimal computational costs, as follows. A shift of n columns along the EIM results in a rotation of the reconstructed image by $\frac{2\pi n}{16}$. Thus, the target image must be shifted according to that angle. To get rid of the rotational variance, we produced 15 additional shifted voltages for each training sample, described previously.

2.3.7. Alpha-Blending

With the given data set, there is still potential for obtaining entirely new training samples. In the field of image classification, there is a technique called α-blending [26–28]. It produces a new image from a linear combination of two other images, and an $\alpha \in [0, 1]$ factor weights these images. An α of 0.3 would mean that the resulting image is a combination of 30% of the first image and 70% of the second image. For EIT images, we can describe the technique as

$$\sigma_{comb} = \alpha\sigma_1 + (1 - \alpha)\sigma_2 \tag{10}$$

From Ohm's law with conductivities and the constant injection current follows the procedure to combine the voltages accordingly

$$\boldsymbol{Y}_{comb} = \alpha\boldsymbol{Y}_1 + (1 - \alpha)\boldsymbol{Y}_2 \tag{11}$$

$$\Leftrightarrow \frac{\boldsymbol{I}}{\boldsymbol{U}_{comb}} = \alpha\frac{\boldsymbol{I}}{\boldsymbol{U}_1} + (1 - \alpha)\frac{\boldsymbol{I}}{\boldsymbol{U}_2} \tag{12}$$

$$\Leftrightarrow \frac{1}{\boldsymbol{U}_{comb}} = \alpha\frac{1}{\boldsymbol{U}_1} + (1 - \alpha)\frac{1}{\boldsymbol{U}_2} \tag{13}$$

$$\Leftrightarrow \boldsymbol{U}_{comb} = \left(\frac{\alpha}{\boldsymbol{U}_1} + \frac{1 - \alpha}{\boldsymbol{U}_2}\right)^{-1} \tag{14}$$

where $\boldsymbol{Y}$ denotes the admittance between the voltage measurement electrodes.

2.3.8. Conclusion on Trainign Dataset

All in all, the choice of the simulated training data was made such that it was as realistic as possible, but at the same time no major assumptions were made on the structure and content of the simulated data nor on the bias in the dataset (e.g., restricting conductivity values to the range expected in the testing data). Furthermore, the described augmentation techniques impose no bias on the simulated data.

2.4. On the ANN Structure

In the domain of classification, ANNs can often be separated into two parts. The first part, consisting of convolutional layers, is used for the extraction of features, while the second part is used for processing these features into educated guesses about the class label. Two very famous examples are AlexNet and VGG19 [29,30]. The second part is realized through fully connected layers. In this work, we modified this basic approach and tailored it specifically for use in EIT. The structure can be seen in Figure 3. When calculating the receptive field of the convolutional layers, it can be seen that the receptive field is of the shape 21×21, although the EIM only has a shape of 16×16. However, Luo et al. showed that the receptive field exhibits a Gaussian distribution [31], which means that features in the center are strongly recognized by the network, while closer to the boundary the features are less recognized. To dampen this problem, we increased the receptive field of RF-REIM-NET.

From Figure 3, it can also be seen that the shape of the input after the convolutions does not change. To achieve this, we used circular padding rather than the standard zero

padding. This choice can be understood by the nature of adjacent–adjacent measurements. On the boundaries of the EIM, values from the other side are inserted, as the neighbor of the 16th electrode pair would also be the 1st electrode pair.

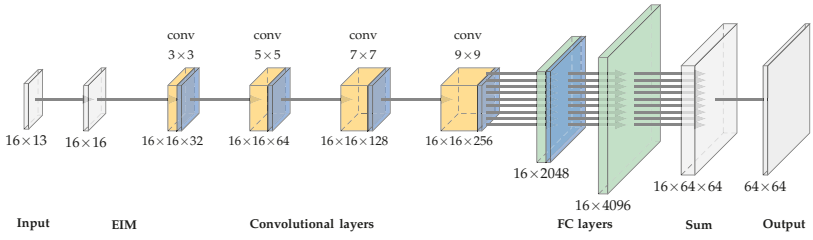


Figure 3. Illustration of the RF-REIM-NET structure. At the beginning, the $16 \times 13 = 208$ voltages are transformed into an EIM. From these, features are extracted with the help of convolutional layers. At the end, these features are processed by fully connected layers, which reconstruct the image for each injection electrode.

Our proposed RF-REIM-NET structure also comes without any form of pooling layers. In general, pooling layers tend to increase the efficiency of ANNs; however, this comes at the cost of broken location invariant properties of the convolutional layers [32]. Another problem is that down sampling, when carried out by pooling, causes aliasing [33]. Thus, we did not use pooling and increased the receptive field of the RF-REIM-NET.

Instead of batch normalization, we used layer normalization. Instead of normalization along the batch, layer normalization computes the normalization along the features of the layer's output. We found that this works best for training.

The second part of RF-REIM-NET consists of a fully connected layer, adapted such that the rotational invariance is considered. The input to the first fully connected layer has the shape 16×2048. This was purposeful, as the 16 represents the 16 different current injection pairs. Thus, instead of passing a vector of $16 \times 2048 = 32{,}768$, 16 passes of a 2048 vector are used. This saves time during training and considerably decreases the size of the RF-REIM-NET. The second fully connected layer works the same way, but at the same time the output will be doubled to obtain a 64×64 reconstruction image.

2.5. Training of the Neural Network

Our training procedure was as follows. α-blending was used during training, and two batches were randomly combined as described in the methods section. As a regularization strategy, L2 weight regularization, dropout with a dropout rate of 0.1 and total variation (TV) regularization were used. L2 weight regularization was added, scaled to the loss of RF-REIM-NET, and can be described as

$$L_{w^2} = \sum_{i=1}^{N} w_i, \tag{15}$$

where N is the total number of weights and w_i is the i^{th} weight. For details about dropout, see [34]. TV regularization is commonly used for image de-noising and de-blurring [35]. The TV regularization loss is computed with

$$L_{TV} = \sum_{i,j} |Y'_{i+1,j} - Y'_{i,j}| + |Y'_{i,j+1} - Y'_{i,j}|, \tag{16}$$

where $\boldsymbol{Y}'$ is the output of RF-REIM-NET.

For the loss function we used the mean squared logarithmic error (MSLE). This error was chosen as our training data vary in orders of magnitude, and RF-REIM-NET should be

able to predict the conductivities in the same way across the whole range. MSLE can be described as

$$L_{MSLE} = \frac{1}{N}\sum(\log(\boldsymbol{Y}+1) - \log(\boldsymbol{Y}'+1))^2, \tag{17}$$

$\boldsymbol{Y}$ is the ground truth conductivity distribution.

The total loss of RF-REIM-NET is described by

$$L_{total} = L_{w^2} + \lambda_1 \cdot L_{TV} + \lambda_2 \cdot L_{MSLE} \tag{18}$$

with $\lambda_1 = 0.1$ and $\lambda_2 = 1 \times 10^{-6}$.

RF-REIM-NET was then trained with the help of the TensorFlow [36] and the Adam optimizer [37]. Additionally, we used a learning rate decay: whenever the loss reached a plateau, the learning rate was reduced by 70%.

2.6. Evaluating RF-REIM-NET

We compared RF-REIM-NET to the standard Gauss–Newton (GN) reconstruction for absolute EIT. To compare the two algorithms quantitatively, we used a modified version of the GREIT figures of merit [38]. Some modifications were necessary, as the original GREIT figures of merit only allow the evaluation of difference images.

First, the calculation of the evaluation mask needs to be modified. The median value of the reconstructed image is subtracted from the original image,

$$\sigma_{eval} = \sigma - \tilde{\sigma}, \tag{19}$$

where $\tilde{\sigma}$ denotes the median value of σ. The median was chosen, as it is more robust to outliers in the data. When inserting only one target for evaluation, an ideal reconstruction would have just two values, the background and the target. If we then subtracted the *mean*, the background would be slightly negative. This is prevented with the help of the subtraction of the median. The mask is then composed of all values, which are 50% less or more than the minimum/maximum value. The choice is dependent on the value of the target with respect to the background. In our evaluation case, the target is less conductive than the background. Thus, our mask is defined as

$$\boldsymbol{m} = \begin{cases} 1 & if\ \sigma_{eval} < \frac{1}{2} \cdot \min(\sigma_{eval}) \\ 0 & else \end{cases}, \tag{20}$$

where $\boldsymbol{m}$ is the evaluation mask. We denote all pixels inside the mask as $\hat{\sigma}_{eval}$, while all pixels outside the mask are denoted as $\sim\hat{\sigma}_{eval}$.

2.6.1. Amplitude Response (AR)

The AR is now defined as

$$AR = \sum \hat{\sigma}_{eval}. \tag{21}$$

The *std* of the AR should be low.

2.6.2. Position Error (PE)

The PE is defined as

$$PE = \sqrt{(\textcircled{\scriptsize x}(\hat{\sigma}_{eval}) - t_x)^2 + (\textcircled{\scriptsize y}(\hat{\sigma}_{eval}) - t_y)^2}, \tag{22}$$

where $\textcircled{\scriptsize x}$ denotes the x-component of the center of gravity, $\textcircled{\scriptsize y}$ denotes the y-component accordingly, t_x is the ground truth x-position and t_y the ground truth y-position. The *mean* and the *std* of the PE should be low.

2.6.3. Ringing (RNG)

The RNG was defined as the *std* of all pixels outside the mask $\boldsymbol{m}$. Formally, this is written as

$$RNG = std(\sim\hat{\sigma}_{eval}). \tag{23}$$

The *mean* and standard RNG deviation should be low.

2.7. Evaluation Data

To validate our RF-REIM-NET, we used three different types of input and analyzed the output according to the three introduced figures of merit (AR, PE and RNG).

2.7.1. FEM Data

First, we used FEM data that the network had not yet seen. Multiple enclosures were simulated, and the enclosure was positioned such that it move from the domain center to the outside. RF-REIM-NET is compared with GN. For the hyperparameter selection of GN, we at first used the L-curve criterion, but did not find usable results. We are convinced that this is due to the reference-free reconstruction, which destabilizes the EIT problem compared with differential EIT. Thus, we made multiple sweeps of the hyperparameter to narrow down the optimal hyperparameter iteratively.

2.7.2. Noise Performance on FEM Data

Second, we compared the noise performance on a FEM data sample. For that, an enclosure near the boundary was simulated and the noise level was increased from 200 to 5 db. For evaluation, GN reconstructions are given.

2.7.3. Tank Data

Third, we used data from a circular EIT tank. The tank had a diameter of 28 cm and had 16 electrodes attached equidistantly around the surface. The tank was filled with 0.9% saline solution and the target was a pickle with a circumference of 4.5 cm. The pickle was moved from the center in the direction of one electrode in nine steps, where the last position was 9 cm in front of an electrode. The measurements were performed with the EIT evaluation kit 2 (*Draeger EEK2*, Draeger Medical GmbH).

2.7.4. Experimental Data

Finally, we give an impression of the performance of RF-REIM-NET on real-world data. The data were taken from an experimental pig trial using a clinical EIT device (*Draeger Pulmo Vista 500*, Draeger Medical GmbH). For the trial, eight pigs were anesthetized and tracheotomized in supine position [39]. During the trial, CT measurements from the pigs were taken. In our data sample, we used two measurement points from a single pig, which was healthy in the time span we chose. The length of the data sample was around 30 s.

As there is no ground truth regarding the conductivity, we show two pictures. The first picture shows the *mean* conductivity over an entire breathing cycle, while the second shows the *std* over an entire breathing cycle. As the background, a CT image is given, as it will give a sense of quality. This is given for transparency, as tank data have more ideal conditions, which are closer to the training data. The absolute Gauss–Newton algorithm did not yield any meaningful results after a thorough hyperparameter sweep and was thus not given as a reference.

3. Results

At first, we give the results of simulated FEM data, outside the distribution of the training data used to train RF-REIM-NET. The results for the figures of merit are given in Table 1. The *std* of the AR is bigger for the GN algorithm; however, the *std* is one order of magnitude lower compared with the *mean*. Visually, this is confirmed by the reconstructions

in Figure 4. The AR for both reconstructions stays roughly the same. For the PE, GN has a lower *mean* and *std*. The PE *mean* is half that of the RF-REIM-NET, while the *std* is a quarter. The RNG for the GN is also lower compared with RF-REIM-NET. This, again, can be seen in the images. The RNG for RF-REIM-NET is larger, due to the higher differences in magnitude outside the mask $\boldsymbol{m}$.

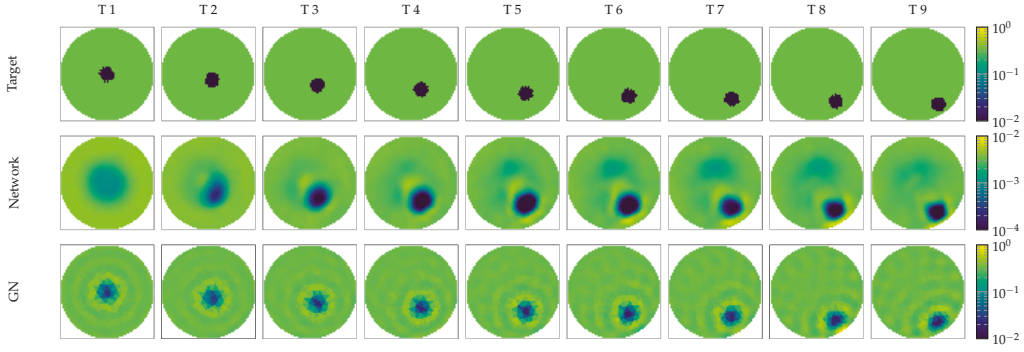


Figure 4. Illustration of RF-REIM-NET (**middle**) and GN (**bottom**) reconstructions of FEM that mimic the position of the tank pickle data. The ground truth target positions are given at the (**top**).

Table 1. Figures of merit for the simulated FEM data. Given is the *mean* $\pm$ *std*.

Algorithm/Metric	AR	PE	RNG
GN	0.069 ± 0.0069	2.7 ± 1.1	0.11 ± 0.0094
RF-REIM-NET	0.066 ± 0.0046	5.6 ± 4.0	0.14 ± 0.019

To better see the difference between the original ground truth image and the reconstruction, we present in Figure 5 the ground truth, the reconstruction of RF-REIM-NET and the MSLE error. It can be seen that the error is for the most part on the edges of the enclosures. In the second column, we can see that the middle target is barely visible in the reconstruction.

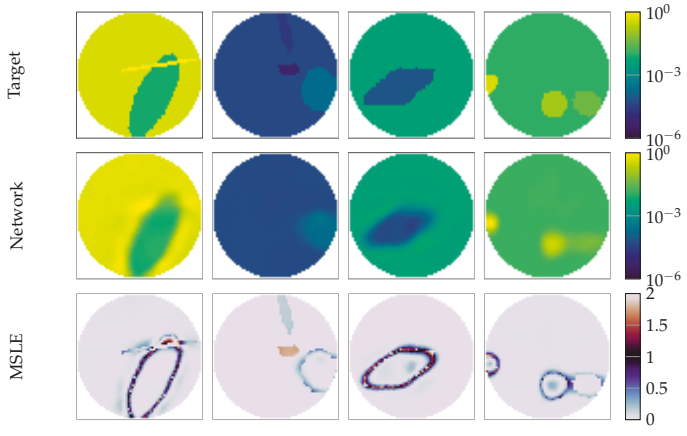


Figure 5. Illustration of RF-REIM-NET reconstructions on the validation dataset. At the top, the original targets are given. In the middle, the reconstructions of RF-REIM-NET are presented. At the bottom, the MSLE error between the original image and the reconstruction are presented.

3.1. Noise Comparison on Simulated Data

Here, we compare the noise performance of the RF-REIM-NET compared with GN. We positioned a target near the boundary of the FEM domain and simulated the voltages. In Figure 6, the results can be seen. The reconstructions of RF-REIM-NET are more robust to noise, compared with GN. At 100 db the reconstruction of GN is barely visible, while RF-REIM-NET is still clearly visible. At 15 db the reconstruction of RF-REIM-NET begins to degrade and also becomes less visible.

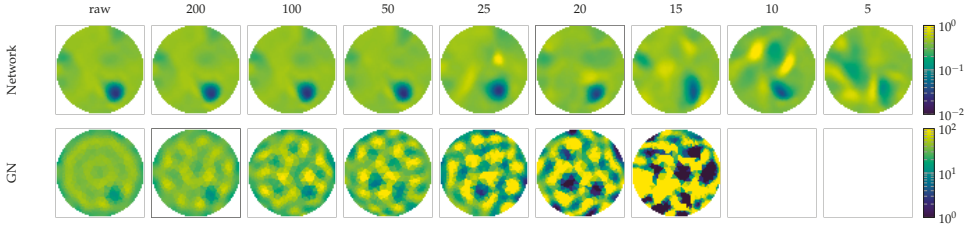


Figure 6. Evaluation of the noise performance of RF-REIM-NET (**top**) compared with GN (**bottom**). The columns represent different noise levels added with the EIDORS function `add_noise`. The target position is equal to T6 in Figure 4.

3.2. Tank Results

Next, we provide the results from the tank measurement. As shown in Figure 7, the pickle was moved from the center to an electrode. The reconstruction from the Gauss–Newton algorithm shows a more diffuse boundary, while RF-REIM-NET has a more clear boundary. The background from the Gauss–Newton algorithm shows many, but small background disturbances, while the background of RF-REIM-NET has fewer disturbances, where one is at the top and the other surrounds the conductivity enclosure.

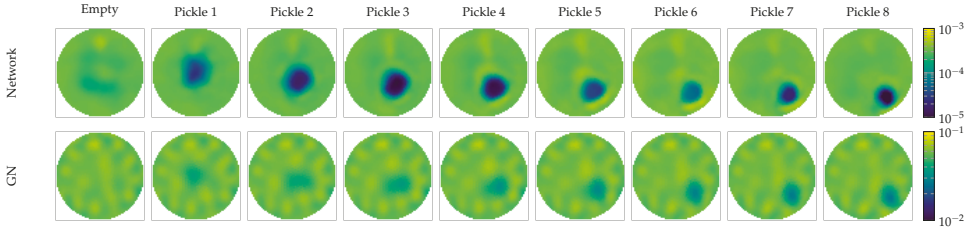


Figure 7. Illustration of RF-REIM-NET and GN reconstructions on the pickle measurements of the tank dataset. The empty measurement was not used in both reconstructions, and is given only for a better view of the reconstruction artifacts. Pickle 1 is the center pickle, while Pickle 7 is the outer pickle.

These observations are also reflected in the figures of merit given in Table 2. The *mean* and *std* of the AR from RF-REIM-NET are bigger than the ones of Gauss–Newton. This can also be visually confirmed in Figure 7. The PE and its *std*, however, are lower in RF-REIM-NET. The *mean* of the PE from RF-REIM-NET is ∼35% lower than that of Gauss–Newton, while its *std* of the PE is ∼50% lower. The *mean* RNG of RF-REIM-NET is 20% lower compared with Gauss–Newton. However, the inverse is true for the *std*: the RNG *std* is 20% higher compared with Gauss–Newton. However, the *std* is $\frac{1}{14}$th of the *mean*.

Table 2. Figures of merit for the tank experiment. The left number in each cell is the *mean* of the metric, while the right number is its *std*.

Algorithm/Metric	AR	PE	RNG
GN	0.1 ± 0.0056	11 ± 5.7	0.1 ± 0.0045
RF-REIM-NET	0.14 ± 0.0086	7.1 ± 2.9	0.08 ± 0.0054

3.3. Experimental Data

For the experimental data, only qualitative analysis is given. The results are shown in Figure 8. On the left, a CT measurement is given to better judge the results. In the middle, the mean reconstruction over 20 s of mechanical ventilation is given. At the top, there is an artifact in the reconstruction. The two lungs are visible, but they are smaller compared with the original size. The heart on the other hand, between the lungs and the artifact, matches the position given in the CT image. The standard deviation picture on the right confirms the findings. The artifact stays in the position, as the standard deviation is near zero at this position. At the position of the heart a high standard deviation is visible, and the same holds true for the lungs. The shape of the standard deviation picture for the lungs better resembles the general shape of the lung.

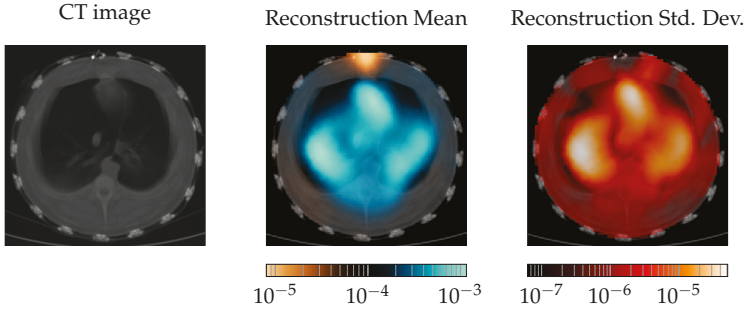


Figure 8. Comparison of the RF-REIM-NET reconstruction with CT scans. On the very left, the CT scan of the pig thorax is given. In the middle, the mean over 20 s of mechanical ventilation is given. On the very right the *std* is given.

3.4. Discussion

In the FEM setting, GN outperformed RF-REIM-NET in the metrics of PE and RNG, both with the mean and the *std*. The mean and *std* of the AR are a little higher using GN. However, the values only differ slightly. Thus, we would argue that GN outperformed RF-REIM-NET in the FEM setting. This is probably due to the fact that the setting has not much disturbance by factors such as hardware or the imperfect conductivity of the target.

In a tank setting, RF-REIM-NET has a lower mean PE and mean RNG. The PE also has a *std* that is roughly half that of the GN PE *std*. This can be visually observed in Figure 7. However, at the same time, the mean AR and its *std* is higher. This, again, can be seen in Figure 7. In the samples "pickle 2", "pickle 3" and "pickle 4", the reconstruction is clearly larger than in the other samples. In contrast, GN has a less clear object boundary. Thus, we argue that on the tank dataset, RF-REIM-NET has a better performance. We showed that RF-REIM-NET is able to give reconstructions from experimental data, even though the ANN does not need any reference voltage, as can be seen in Figure 8. To the best of our knowledge, this is the first work to evaluate the performance of ANNs for EIT reconstructions on real-world experimental data from an ANN solely trained on simulated data.

While the heart was reconstructed accurately, the lungs were too small, which is at that point not fully useful for clinical diagnostics. Another shortcoming is the artifact at the top, which constantly stays in that position. We assume that the artifact is due to electrode

position errors. At the top of the picture, the EIT belt is closed. Thus, the electrodes have a larger distance from each other at that place.

4. Conclusions and Outlook

We present an ANN (RF-REIM-NET) that is able to reconstruct conductivity enclosures without using a reference voltage. RF-REIM-NET is inspired by ANNs that are commonly used for classification: the first part of these ANNs extracts features, while the second part is responsible for the evaluation. Compared with GN on FEM and tank data, our approach tends to give clearer reconstructions. However, the images tend to be a little bigger than in real life. We also showed the performance on real-world subject data. Compared with GN, which did not obtain any meaningful reconstructions from the experimental data set, RF-REIM-NET was able to give reconstructions. For future work, the network needs to be made more robust against electrode position errors and domain shape influences, which may be the biggest impact factors on the experimental data performance. Thus, in the future, altering the electrode positions in the training data might improve the overall reconstructions. Second, the boundary shape has to be altered more drastically, as this might further increase the performance of RF-REIM-NET.

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

Funding: The authors gratefully acknowledge financial support provided by the project InDiThera (13GW0361E) through the Federal Ministry of Education and Research. This research was supported by a grant (PU 219/2-1) from the German Research Foundation (DFG).

Institutional Review Board Statement: The study from which the real world data was used, was approved by the regional ethical committee (No. 5.8.18-15771/2017, Uppsala, Sweden).

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

EIT	Electrical Impedance Tomography
FEM	Finite Element Method
ANN	Artifical Neural Network
EIM	Electrical Impedance Map
GN	Gauss–Newton
AR	Amplitude Response
PE	Position Error
RNG	Ringing

References

1. Hallaji, M.; Seppänen, A.; Pour-Ghaz, M. Electrical impedance tomography-based sensing skin for quantitative imaging of damage in concrete. *Smart Mater. Struct.* **2014**, *23*, 085001. [CrossRef]
2. Kruger, M.; Poolla, K.; Spanos, C.J. A class of impedance tomography based sensors for semiconductor manufacturing. In Proceedings of the 2004 American Control Conference, Boston, MA, USA, 30 June–2 July 2004; Volume 3, pp. 2178–2183.
3. Yang, Y.; Wu, H.; Jia, J.; Bagnaninchi, P.O. Scaffold-based 3-D cell culture imaging using a miniature electrical impedance tomography sensor. *IEEE Sens. J.* **2019**, *19*, 9071–9080. [CrossRef]

4. Meier, T.; Luepschen, H.; Karsten, J.; Leibecke, T.; Großherr, M.; Gehring, H.; Leonhardt, S. Assessment of regional lung recruitment and derecruitment during a PEEP trial based on electrical impedance tomography. *Intensive Care Med.* **2008**, *34*, 543–550. [CrossRef] [PubMed]
5. Hentze, B.; Muders, T.; Luepschen, H.; Maripuu, E.; Hedenstierna, G.; Putensen, C.; Walter, M.; Leonhardt, S. Regional lung ventilation and perfusion by electrical impedance tomography compared to single-photon emission computed tomography. *Physiol. Meas.* **2018**, *39*, 065004. [CrossRef]
6. Abascal, J.F.P.; Arridge, S.R.; Atkinson, D.; Horesh, R.; Fabrizi, L.; De Lucia, M.; Horesh, L.; Bayford, R.H.; Holder, D.S. Use of anisotropic modelling in electrical impedance tomography; description of method and preliminary assessment of utility in imaging brain function in the adult human head. *Neuroimage* **2008**, *43*, 258–268. [CrossRef]
7. Leonhardt, S.; Cordes, A.; Plewa, H.; Pikkemaat, R.; Soljanik, I.; Moehring, K.; Gerner, H.J.; Rupp, R. Electric impedance tomography for monitoring volume and size of the urinary bladder. *Biomed. Eng./Biomed. Tech.* **2011**, *56*, 301–307. [CrossRef]
8. Hong, S.; Lee, K.; Ha, U.; Kim, H.; Lee, Y.; Kim, Y.; Yoo, H.J. A 4.9 mΩ-sensitivity mobile electrical impedance tomography IC for early breast-cancer detection system. *IEEE J. Solid-State Circuits* **2014**, *50*, 245–257. [CrossRef]
9. Putensen, C.; Hentze, B.; Muenster, S.; Muders, T. Electrical impedance tomography for cardio-pulmonary monitoring. *J. Clin. Med.* **2019**, *8*, 1176. [CrossRef]
10. Costa, E.L.; Lima, R.G.; Amato, M.B. Electrical impedance tomography. In *Yearbook of Intensive Care and Emergency Medicine*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009; pp. 394–404.
11. Borsic, A.; Graham, B.M.; Adler, A.; Lionheart, W.R. Total Variation Regularization in Electrical Impedance Tomography. 2007. Available online: http://eprints.maths.manchester.ac.uk/id/eprint/813 (accessed on 10 February 2022).
12. Vauhkonen, M.; Vadasz, D.; Karjalainen, P.A.; Somersalo, E.; Kaipio, J.P. Tikhonov regularization and prior information in electrical impedance tomography. *IEEE Trans. Med. Imaging* **1998**, *17*, 285–293. [CrossRef]
13. Kaipio, J.P.; Kolehmainen, V.; Somersalo, E.; Vauhkonen, M. Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse Probl.* **2000**, *16*, 1487. [CrossRef]
14. Kolehmainen, V.; Somersalo, E.; Vauhkonen, P.; Vauhkonen, M.; Kaipio, J. A Bayesian approach and total variation priors in 3D electrical impedance tomography. In Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Volume 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286), Hong Kong, China, 1 November 1998; IEEE: Washington, DC, USA, 1998; Volume 2, pp. 1028–1031.
15. Isaacson, D.; Mueller, J.L.; Newell, J.C.; Siltanen, S. Reconstructions of chest phantoms by the D-bar method for electrical impedance tomography. *IEEE Trans. Med. Imaging* **2004**, *23*, 821–828. [CrossRef]
16. Adler, A.; Lionheart, W.R. Uses and abuses of EIDORS: An extensible software base for EIT. *Physiol. Meas.* **2006**, *27*, S25. [CrossRef]
17. Kłosowski, G.; Rymarczyk, T. Using neural networks and deep learning algorithms in electrical impedance tomography. *Informatyka Automatyka Pomiary w Gospodarce i Ochronie Środowiska* **2017**, *7*. [CrossRef]
18. Hamilton, S.J.; Hauptmann, A. Deep D-bar: Real-time electrical impedance tomography imaging with deep neural networks. *IEEE Trans. Med. Imaging* **2018**, *37*, 2367–2377. [CrossRef]
19. Hu, D.; Lu, K.; Yang, Y. Image reconstruction for electrical impedance tomography based on spatial invariant feature maps and convolutional neural network. In Proceedings of the 2019 IEEE International Conference on Imaging Systems and Techniques (IST), Abu Dhabi, United Arab Emirates, 9–10 December 2019; pp. 1–6.
20. Tan, C.; Lv, S.; Dong, F.; Takei, M. Image reconstruction based on convolutional neural network for electrical resistance tomography. *IEEE Sens. J.* **2018**, *19*, 196–204. [CrossRef]
21. Paullada, A.; Raji, I.D.; Bender, E.M.; Denton, E.; Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv* **2020**, arXiv:2012.05345.
22. Hasgall, P.; Di Gennaro, F.; Baumgartner, C.; Neufeld, E.; Lloyd, B.; Gosselin, M.; Payne, D.; Klingenböck, A.; Kuster, N. *IT'IS Database for Thermal and Electromagnetic Parameters of Biological Tissues*; Version 4.0, 15 May 2018; Technical Report, VIP21000-04-0. itis. swiss/database; ScienceOpen, Inc.: Burlington, MA, USA, 2018.
23. Talman, A.; Chatzikyriakidis, S. Testing the generalization power of neural network models across NLI benchmarks. *arXiv* **2018**, arXiv:1810.09774.
24. De Teyou, G.K.; Petit, H.; Loumeau, P.; Fakhoury, H.; Le Guillou, Y.; Paquelet, S. Statistical analysis of noise in broadband and high resolution ADCs. In Proceedings of the 2014 21st IEEE International Conference on Electronics, Circuits and Systems (ICECS), Marseille, France, 7–10 December 2014; pp. 490–493.
25. Frangi, A.; Rosell, J. A theoretical analysis of noise in electrical impedance tomographic images. In Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 'Magnificent Milestones and Emerging Opportunities in Medical Engineering' (Cat. No. 97CH36136), Chicago, IL, USA, 30 October–2 November 1997; Volume 1, pp. 433–436.
26. Porter, T.; Duff, T. Compositing digital images. In Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques, Minneapolis, MN, USA, 23–27 July 1984; pp. 253–259.
27. Inoue, H. Data augmentation by pairing samples for images classification. *arXiv* **2018**, arXiv:1801.02929.
28. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.

29. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4905–4913.
32. Azulay, A.; Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.* **2019**, *20*, 1–25.
33. Zhang, R. Making convolutional networks shift-invariant again. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7324–7334.
34. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
35. Liu, J.; Sun, Y.; Xu, X.; Kamilov, U.S. Image restoration using total variation regularized deep image prior. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7715–7719.
36. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software. Available online: tensorflow.org (accessed on 10 February 2022).
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Conference Track Proceedings; Bengio, Y., LeCun, Y., Eds.; DBLP: Trier, Germany, 2015.
38. Adler, A.; Arnold, J.H.; Bayford, R.; Borsic, A.; Brown, B.; Dixon, P.; Faes, T.J.; Frerichs, I.; Gagnon, H.; Gärber, Y.; et al. GREIT: A unified approach to 2D linear EIT reconstruction of lung images. *Physiol. Meas.* **2009**, *30*, S35. [CrossRef] [PubMed]
39. Muders, T.; Luepschen, H.; Meier, T.; Reske, A.W.; Zinserling, J.; Kreyer, S.; Pikkemaat, R.; Maripu, E.; Leonhardt, S.; Hedenstierna, G.; et al. Individualized positive end-expiratory pressure and regional gas exchange in porcine lung injury. *Anesthesiology* **2020**, *132*, 808–824. [CrossRef] [PubMed]

diagnostics

MDPI

Article

Generalization Challenges in Drug-Resistant Tuberculosis Detection from Chest X-rays

Manohar Karki [1,*], Karthik Kantipudi [2,*], Feng Yang [1], Hang Yu [1], Yi Xiang J. Wang [1,3], Ziv Yaniv [2] and Stefan Jaeger [1,*]

[1] Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD 20894, USA; feng.yang2@nih.gov (F.Y.); hang.yu@nih.gov (H.Y.); yixiang_wang@cuhk.edu.hk (Y.X.J.W.)
[2] Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20894, USA; zivyaniv@nih.gov
[3] Department of Imaging and Interventional Radiology, Faculty of Medicine, The Chinese University of Hong Kong, Prince of Wales Hospital, New Territories, Hong Kong
* Correspondence: mkarki2@gmail.com (M.K.); karthik.kantipudi@nih.gov (K.K.); stefan.jaeger@nih.gov (S.J.)

Abstract: Classification of drug-resistant tuberculosis (DR-TB) and drug-sensitive tuberculosis (DS-TB) from chest radiographs remains an open problem. Our previous cross validation performance on publicly available chest X-ray (CXR) data combined with image augmentation, the addition of synthetically generated and publicly available images achieved a performance of 85% AUC with a deep convolutional neural network (CNN). However, when we evaluated the CNN model trained to classify DR-TB and DS-TB on unseen data, significant performance degradation was observed (65% AUC). Hence, in this paper, we investigate the generalizability of our models on images from a held out country's dataset. We explore the extent of the problem and the possible reasons behind the lack of good generalization. A comparison of radiologist-annotated lesion locations in the lung and the trained model's localization of areas of interest, using GradCAM, did not show much overlap. Using the same network architecture, a multi-country classifier was able to identify the country of origin of the X-ray with high accuracy (86%), suggesting that image acquisition differences and the distribution of non-pathological and non-anatomical aspects of the images are affecting the generalization and localization of the drug resistance classification model as well. When CXR images were severely corrupted, the performance on the validation set was still better than 60% AUC. The model overfitted to the data from countries in the cross validation set but did not generalize to the held out country. Finally, we applied a multi-task based approach that uses prior TB lesions location information to guide the classifier network to focus its attention on improving the generalization performance on the held out set from another country to 68% AUC.

Keywords: Tuberculosis (TB); drug resistance; deep learning; chest X-rays; generalization; localization

Citation: Karki, M.; Kantipudi, K.; Yang, F.; Yu, H.; Wang, Y.X.J.; Yaniv, Z.; Jaeger, S. Generalization Challenges in Drug-Resistant Tuberculosis Detection from Chest X-rays. *Diagnostics* **2022**, *12*, 188. https://doi.org/10.3390/diagnostics12010188

Academic Editors: Philippe A. Grenier, Henk A. Marquering, Sameer Antani and Sivaramakrishnan Rajaraman

Received: 1 December 2021
Accepted: 5 January 2022
Published: 13 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

According to the 2020 World Health Organization (WHO) report [1], it is estimated that in 2019 about 10 million people fell ill with Tuberculosis (TB) and about 1.4 million died from the disease. Based on the same report, it is estimated that in 2019 about 0.5 million individuals were infected with rifampicin-resistant TB out of which about 400,000 were multidrug-resistant.

Drug-resistant TB (DR-TB) is a growing public health concern requiring longer and more complex treatment than drug-sensitive TB (DS-TB), in addition to incurring higher financial costs. Treatment for DR-TB requires a course of second-line drugs for at least 9 months and up to 20 months, supported by counselling and monitoring for adverse events. In comparison, treatment of DS-TB only lasts between 6–9 months. Early diagnosis of DR-TB is crucial for selecting appropriate, patient-specific, treatment regimens. Thus,

improving early decision making has the potential to increase favorable patient outcomes, combat the spread of infection and reduce the overall financial costs associated with the disease.

Currently, the diagnostic methods for identifying DR-TB infections require culture and drug susceptibility testing. These procedures are not feasible globally, especially for countries unable to scale up their testing capacities. An automated, low cost, computational approach that utilizes readily available resources such as medical images and other clinical information is thus desirable.

In the context of TB diagnosis, automated deep learning based systems which only utilize Chest X-rays (CXRs) have seen significant success, with multiple commercial offerings available [2,3]. In one evaluation study, these systems classified CXRs as TB/not-TB with an Area Under the Curve (AUC) of above 0.9 [2]. In another study, they outperformed radiologists, with two of the systems meeting the WHO's target product profile for triage tests [3].

Currently, discrimination between DR-TB and DS-TB using readily available clinical images and possibly additional clinical side information is still an open problem. In this work, we used both deep and classical machine learning algorithms to classify drug-resistant (DR) and drug-sensitive (DS) tuberculosis in chest X-ray images, radiological features, and clinical patient information. Specifically, we have:

- analyzed a state-of-the-art classifier in terms of its capability to generalize on unseen data from another country. This has been an issue largely neglected by the research community in the past. However, we show that a high classification performance is not sufficient for practical usefulness. The capability to provide consistent performance across different datasets, hospitals, and countries, is essential;
- investigated the problem of poor generalization to unseen data by comparing the performance of the deep learning based classifier with other classifiers trained on texture features extracted from X-ray images. We also explore if these generalization problems exist in other clinical data, and if non-disease related attributes such as the origin of chest X-rays can influence the drug resistance detection performance;
- studied explicit and implicit ways to steer the attention of the classifier. We use segmented lungs as a means to guide the network to learn explicitly from the lungs. We also propose a novel multi-task approach that uses prior information (TB lesions' locations) to implicitly focus on the important regions and improve DR/DS classification performance.

2. Previous Work

Attempts to utilize images and clinical data to distinguish between DR-TB and DS-TB have been previously described in multiple publications. These works are either based on the utilization of radiological findings identified in the image by a clinician or via fully automated methods which receive as input the image and potentially other available clinical data and output the likelihood for each of the two classes.

Several studies have shown that radiological findings based on a radiologist reading of a CT or CXR have the potential to differentiate between the two classes. A literature review from 2018 [4] concluded that the presence of thick-walled multiple cavities in the images is a useful predictor for DR-TB, with good specificity but low sensitivity. Another study [5] compared 183 DR-TB cases and 183 DS-TB cases from a single hospital. This study concluded that there were substantial differences in findings between the two classes in terms of lesion size and morphology. A slightly larger study [6], which compared 468 DR-TB cases and 223 DS-TB cases concluded that a combination of the number and size of consolidated nodules is a good predictor for DR-TB. Another, small study [7], utilized data from 144 patients and found that the presence of multiple cavities is a good predictor for DR-TB. A much larger study [8], compared 516 DR-TB and 1030 DS-TB cases, obtaining an AUC of 0.83 using a regression model. This study observed that the co-existence of multiple findings (multiple cavities, thick-walled cavities, disseminated lesions along the

bronchi, whole-lung involvement) was indicative of DR-TB. Finally, more recent work [9] compared 1455 DR-TB and 782 DS-TB cases, using two clinical features and 23 types of radiological findings. A support vector machine was used to distinguish between DR-TB and DS-TB with an AUC of 0.78. It should be noted that reliance on a radiologist reading is a significant limitation. The lack of consensus on radiological findings for drug resistance further hinders the clinical usefulness of these approaches. Because of these reasons, fully automated solutions, described next, are more desirable.

Several fully automated solutions were presented as part of the ImageCLEF 2017 and 2018 evaluation challenge forums [10]. These challenges included a subtask, differentiating between DR-TB and DS-TB using thoracic Computed Tomography (CT) images. This classification task included 259 training images and 236 test images with about half of the cases DR-TB and half DS-TB. Proposed solutions included Gentili et al. [11] who reformatted the CT images to the coronal plane and used a pre-trained ResNet50 Convolutional Neural Network (CNN). For the same challenge, Ishay et al. [12] used an ensemble of 3D CNNs and Cid et al. [13] used a 3D texture-based graph model and support vector machines (SVM). Allaouzi et al. [14] replaced the softmax function of a 3D CNN architecture with an SVM to tackle this classification task. All entries had limited success, resulting in AUCs of about 0.6. After two editions, the organizers removed the subtask from the competition with the conclusion that "the MDR subtask was not possible to solve based only on the image". While these challenges did not yield the desired results, the results obtained using radiologist readings are more favorable, suggesting that the sub-optimal performance may be due to the small number of images available for training. It should be noted that increasing the number of CTs for this task is not trivial as the use of CT imaging in DS-TB cases is uncommon, with the standard imaging modality being CXR. The rare use of CT imaging in standard practice, and the consequential lack of data to analyze, limits the usage of CT images to train a model to distinguish between DR and DS-TB.

On CXR images, [15] utilized a customized CNN architecture to classify DR-TB and DS-TB from 2973 images from the TB portals dataset. They achieved a classification performance of 66%, which improved to 67% when follow up images were also included. Our group has previously proposed fully automated methods utilizing CXRs as described in [16,17]. In [16], we utilized 135 CXRs from a single source. Using a shallow neural network we obtained an AUC of 0.66. In [17], we utilized a much larger dataset, 3642 images from multiple sources. Using a deep neural network, InceptionV3 pre-trained on ImageNet, we obtained an AUC of 0.85. This result is the current state-of-the-art performance achieved on the TB portals data. This is a significant improvement of results from other approaches. However, even though a 10-fold cross validation was performed, the capability of the trained network to classify chest X-rays from unseen domains was not evaluated. In fact, the common weakness of all of these automated methods is that they have not been evaluated for generalization by separating the source of the data. As different medical imaging technologies and devices produce different standards and quality of images, it is important for our models to be robust to these changes.

An underlying assumption of most machine learning algorithms is that the population, test, and training data are independent and identically distributed. If the two distributions are different, then the learned parameters will not yield a good performance. That is, the model will not generalize well to unseen data. While CXR imaging is a low cost modality that is in widespread use, the variations in the standards of the acquired images is significant [18,19], bringing into question the utility of any proposed method which is not evaluated on its generalization capability. More specifically, Harris et al. [20] found that 80% of published works on using CXR for TB diagnosis either used the same databases to train and test their software, or did not comment on databases they used for testing their models. Sathitratanacheewin et al. [21] also observed that a model for CXR-based TB diagnosis performed well with 0.85 AUC when tested on images within their intramural dataset with significant performance deterioration when tested on extramural images, yielding an AUC of 0.7. For domain shift, when the change in image distribution between the training and

testing sets is inevitable, it has been shown that these effects can be ameliorated if training is formulated using a multi-task approach [22].

In addition to the generalization issues due to domain shift, the generalization of deep learning algorithms can also deteriorate if they learn irrelevant features. This is a specific shortcoming of deep learning algorithms as they do not preclude the algorithm from learning features present in the training set that are arbitrarily correlated with the disease, yet are completely irrelevant. These can stem from characteristics of the imaging devices or clinical practices such as patient positioning [23–25] used at the specific locations. If a model implicitly learns such features it will not generalize well when presented with data obtained on different imaging devices or using different clinical workflows, both of which are irrelevant to disease diagnosis.

In this work, we explore various strategies to improve the generalization of models for classification of CXRs as DR-TB or DS-TB using various normalization and attention mechanisms, both explicit (segmentation based) and implicit (multi-task based).

3. Data

3.1. TB Portals Data

The primary data source used in this work is from the NIAID TB Portals program (https://tbportals.niaid.nih.gov (accessed on 10 January 2022)), with a public data release date of October 2020. The dataset contains clinical data and CXR images that are anonymized and made available for public use [26]. Each patient record is manually annotated with clinical information and radiological findings based on the associated CXR image. For this work, data from 1756 patients from ten countries were used. Table 1 shows the data distribution based on country of origin and gender. It should be noted that the TB portals data were collected with a primary focus on acquisition of drug-resistant cases and cases that reflect the specific research interest at the country of origin. As a result, the data are imbalanced in terms of the ratio between drug-resistant and drug-sensitive cases, which does not necessarily reflect the prevalence of TB from either class in the contributing country. Interestingly, we also see that the data are not balanced in terms of gender with about double the number of males to females. This does reflect known differences in TB prevalence in females versus males and has been linked to both societal and biological differences between the sexes [27–29].

Table 1. Patient distribution from different countries and genders for the chest X-ray data used in this work.

Country	Number of Patients			
	Drug-Sensitive	*Drug-Resistant*	*Male*	*Female*
Belarus	118	344	294	168
Georgia	399	236	472	163
Romania	15	114	91	38
Azerbaijan	0	32	24	8
India	197	21	165	53
Moldova	12	32	37	7
Kyrgyzstan	0	18	11	7
Ukraine	8	25	25	8
Kazakhstan	15	53	36	32
South Africa	114	3	72	45
Total	878	878	1227	529

3.2. Clinical Data

The clinical data contain an extensive set of features associated with each patient. This includes demographic data, radiologists' findings for each CXR, different diagnostic tests and treatment information. Additionally, it includes demographic features such as age of onset, gender, patient type (New, Relapse or Failure), body mass index, country of

origin, education, employment, number of daily contacts, number of children, prescription drug usage, laboratory tests, treatment period, treatment status and outcome. The radiologists' findings include chest radiography patterns such as nodules, cavities, collapses and infiltrates and their location in the lungs. Due to financial constraints and the size of the TB portals CXR dataset, radiological findings are obtained using a single experienced radiologist-reading per image. The whole dataset was annotated by multiple radiologists from the countries contributing data to the program. Consequentially, the radiological findings are not biased towards a single radiologist. Table 2 lists all finding types used by the radiologists to annotate the images. These are abnormalities commonly associated with TB. In addition to the type of abnormality, the findings are further differentiated based on their size (small, medium, large) and number of occurrences (single, multiple).

Table 2. Twenty features derived from the presence of abnormalities that are localized to different sextants.

Types of Abnormalities	
collapse	small nodules
small cavities	medium nodules
medium cavities	large nodules
large cavities	huge nodules
large cavity belonging to multiple sextants	non-calcified nodule
multiple cavities	clustered nodules
low ground glass density, active fresh nodules	multiple nodules
medium density stabilized fibrotic nodules	infiltrate: low ground glass density
high density calcification, typically sequella	infiltrate: medium density
calcified or partially calcified nodule	infiltrate: high density

3.3. Chest X-ray Images

All TB Portals CXRs used in this work are from a frontal, AP or PA, view and have varied resolutions (206 × 115 to 4453 × 3719). The intensity range found in the images also varies, with 1177 images having a low dynamic, intensities in the 0–255 range, and 579 images having a high dynamic, intensities in the 0–65,536 range.

It should be noted that the drug susceptibility label associated with each image is obtained via drug susceptibility testing and is not derived from the image. Additionally, the usage of radiological findings for predicting drug susceptibility has shown moderate success. Thus, the question of whether good performance for predicting drug susceptibility from CXRs from unseen sources is possible remains open.

In addition to the CXRs from the TB Portals program, we use a publicly available TB CXR dataset collected from a hospital in China [30] (Download from http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip (accessed on 10 January 2022)). This dataset contains 662 frontal chest X-rays, of which 326 are labeled as non-TB cases and 336 are labeled as TB. There are two sets of annotations where each abnormal TB image has been manually annotated by two radiologists. Figure 1 shows one such segmentation.

Sextant Division

To further differentiate between radiological findings, we associate them with their spatial location in the lungs. To do so, we define lung sextants by dividing each lung into three equal sections from apex to base, as shown in Figure 2. The division of the sextants can be subjective for findings close to sextant borders, when the division boundaries may not be strictly adhered to by the annotating radiologist. In this work, we say a sextant is affected by TB if at least one of the abnormalities listed in Table 2 is present in the sextant.

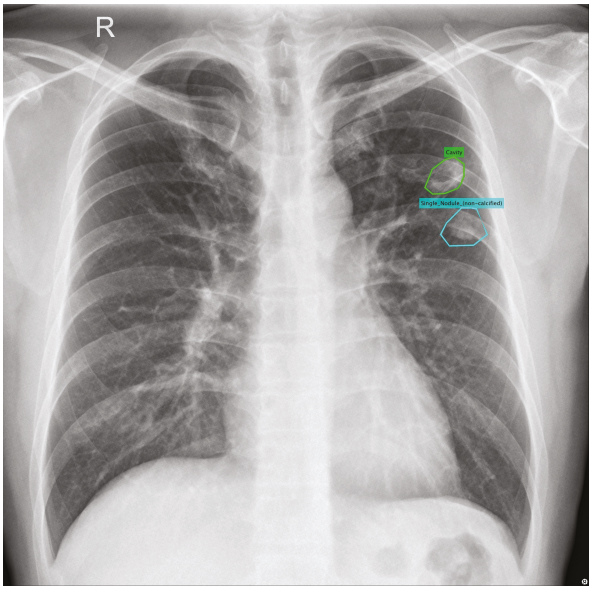


Figure 1. Example of a lung segmentation for a nodule and a cavity.

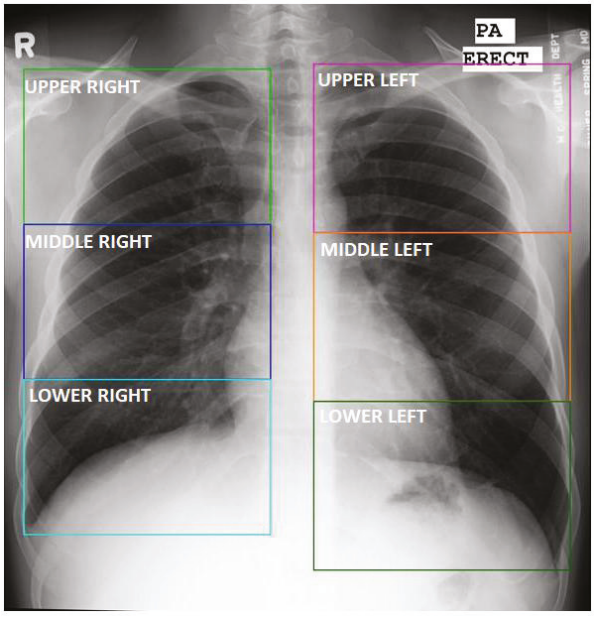


Figure 2. Definition of six lung sextants. Abnormal annotations are assigned to one or more of these sextant divisions.

3.4. Dataset Definitions

For our experiments, we only select the first image taken in the clinical process for a patient; hence, the number of images is equal to the number of patients. All of our drug resistance classification experiments feature an equal number of DR and DS patients in the

training set. For data balancing, we use a conservative approach, excluding images from the majority class. The subsets used for training and evaluation are listed below:

- **Generalization Dataset (Gen. Dataset)**: A total of 1520 samples are selected for this set, 760 samples for each class. All samples originating from the country Belarus are excluded.
- **Dataset with sextant annotations (Sext. Dataset)**: This set contains a total of 1118 samples, 559 samples for each class. This set also does not include any data from Belarus.
- **Validation Set**: This is a cross validation set, which varies for each fold. It contains a randomly selected set containing 20% of the dataset (5-fold CV) for each cross validation training. The numbers reported for the validation set are the average performance values of all cross validation folds. Because, this would be a subset of the above two datasets, no samples from Belarus will be present in this set either.
- **Belarus Dataset**: This dataset contains a maximum of 118 samples from each class with 236 samples in total. The Belarus dataset is used as the test set for most experiments. When sextant-based data should be required, five samples from each class are removed as they do not contain sextant information.

3.5. Data Standardization

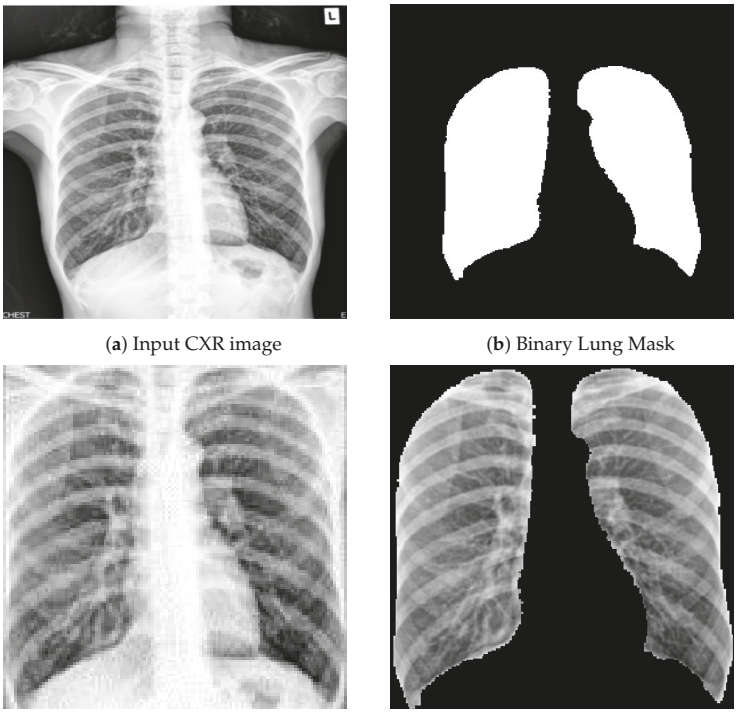
Lung segmentation is used to explicitly address the challenges associated with generalization due to domain shift and the possible existence of confounding factors due to class-correlated yet irrelevant features. Segmentation enables us to limit the input images for the binary DR/DS classifier so that they only contain regions relevant for classification of pulmonary tuberculosis, meaning the lungs. Additionally, the lungs are scaled to a uniform size and position within the image, removing potential confounding factors such as lung size and patient placement that are often correlated with the clinical sites and thus with the local prevalence of TB types. Once the lung regions are segmented, the image is cropped to the lung bounding box, Figure 3c, and all information outside the lung is removed, Figure 3d.

For lung segmentation we initially utilized a publicly available U-Net model which was trained on two datasets with a total of 385 images and corresponding manual lung segmentations [30,31] (https://github.com/imlab-uiip/lung-segmentation-2d (accessed on 10 January 2022)). Unfortunately, this model failed frequently when applied to the TB portals images. Often, one or both sides of the lung were not segmented.

Furthermore, segmentation using this model failed on pathological lung regions in a significant number of images, which is detrimental for disease analysis.

To address these performance limitations a U-Net based [32] segmentation model with a ResNet50 backbone [33] was trained using the publicly available v7 COVID-19 X-ray dataset, which contains 6500 images and corresponding manual lung segmentations (https://github.com/v7labs/covid-19-xray-dataset (accessed on 10 January 2022)).

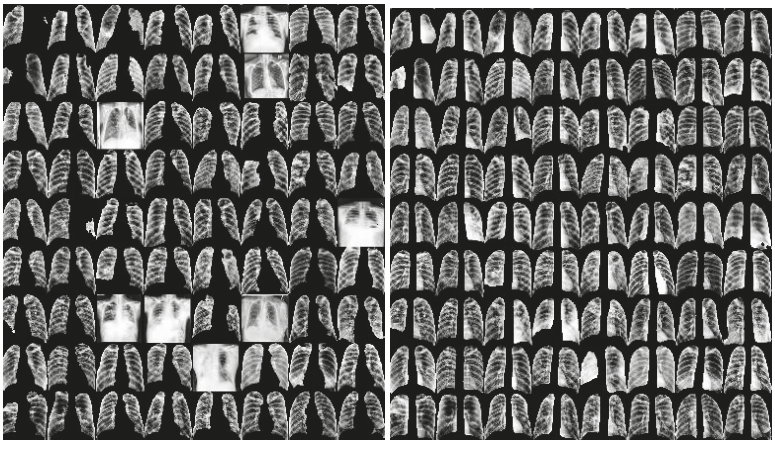
As the TB portals dataset does not provide ground truth lung segmentations, results were visually evaluated as either failure or success. The segmentation failure rates of this model and the previous model were 0.06% and 3% respectively. Aside from that, the old model segmented one of the lungs with less than 10% of the corresponding ground truth pixels in 0.8% of the cases. No such cases were observed in the new model. Figure 4 illustrates the difference between the two models applied to the same set of 72 images.



(**a**) Input CXR image (**b**) Binary Lung Mask

(**c**) Cropped Image (**d**) Lung Segmented

Figure 3. Original CXR (**a**) is fed to the U-Net, which outputs a binary lung mask (**b**) with which the original CXR is cropped(**c**) and the lungs are segmented (**d**) in the cropped bounding box.



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

Figure 4. Cropped images based on the lung segmentation results obtained using a publicly available UNet model trained on the combined JSRT and Montgomery datasets (**a**), and using a customized UNet model trained on the v7 COVID-19 X-ray dataset (**b**). The performance of the customized model is clearly better. Note that if the segmentation fails, the entire image is used.

4. Drug Resistance Classification

For classifying between drug-resistant and drug-sensitive TB, we primarily use the chest X-rays but also utilize text data to assist with the classification and to compare the performance when using just the images. Classic machine learning algorithms and CNNs with pretrained weights were used on the clinical text data and chest X-ray images respectively. Figure 5 shows the setup of our classification network where the preprocessed chest X-ray image is the input and the prediction is either drug-resistant (DR) or drug-sensitive (DS).

As the focus of this work is to evaluate, understand and propose solutions to the issue of generalization to unseen data, we describe in the following subsections: (a) the need of domain adaptation for a network to generalize to unseen data, (b) the use of radiomics features derived from chest X-rays, (c) multi-task learning as a means to provide implicit attention to the main task of DR/DS classification, (d) classifying per-sextant abnormality, and (e) segmenting abnormal regions.

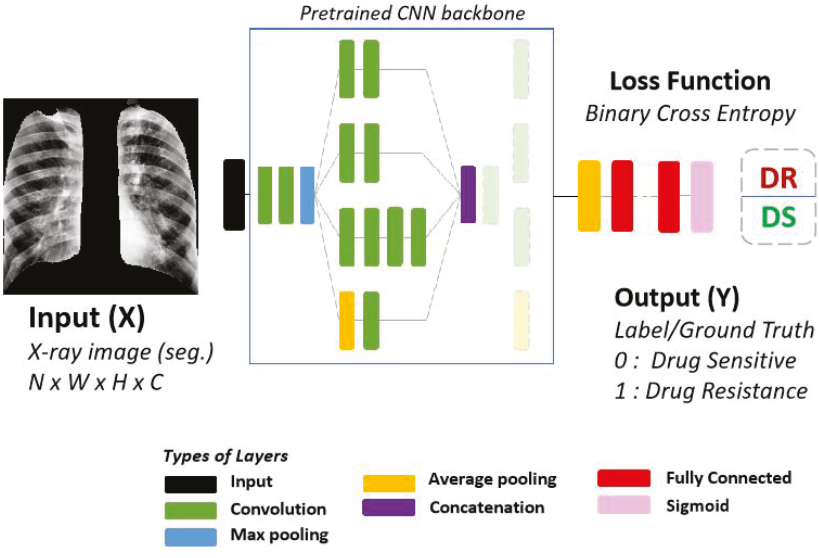


Figure 5. Standard CNN architecture for drug resistance classification. The Input (X) is a preprocessed X-ray image with segmented lungs. The Output (Y) is one of two classes, DR-TB or DS-TB. For this work, we use the ResNet18 [33] architecture as the backbone for all of our experiments.

4.1. Domain Adaptation

The distribution for which a trained model is tested can often be significantly different from the distribution that it was trained on. There is no guarantee that a trained model will be robust to data it has not seen before. Different acquisition standards [34], equipment, and even personnel can create vastly different looking images for medical images. Even after acquisition, other processing and storage differences can create differences in the images. For a human, these variations may be easier to overcome but a machine learning model needs to be trained to understand the differences. Either smart features and algorithms need to be employed or a large and diverse set of data is required to train such a model.

Evaluation of models on unseen data from different domains is the logical way to evaluate such models. Besides that, interpreting the model's decision can also be valuable to understand a prediction. For drug resistance TB classification, localizing the prediction decisions is worthwhile, as tuberculosis itself is frequently observed in certain regions of the lung.

Furthermore, it is worth exploring how easily images from different domains can be discriminated. Easily distinguishable domains in the input coupled with an imbalanced dataset can readily result in failure to generalize.

The usage of transfer learning enables a model to adapt to the new domain, but is less desirable when compared to a fixed model which does not require additional training per domain. Starting from pretrained weights allows for the high-level features to be consistent and not overly dependent on the domain of the training images. This explains why networks initialized with pretrained weights consistently outperformed networks randomly initialized [17].

4.2. Radiomics Features

Usage of explicit, engineered features can be used as a counter measure to prevent a network from learning correlated, yet irrelevant, features within a dataset. Radiomic features have been used to extract patterns that may have been missed by radiologists to identify abnormalities present in medical images [35].

The features used in the paper include 2D-shape based features (e.g. axis lengths), first order statistics (e.g. skewness), gray-level co-occurence matrix features (GLCM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM) and neighbouring gray-tone difference matrix (NGTDM) features.

4.3. Multi-Task Learning

To encourage a network to focus on desirable localities, such that the network is generalizing based on the actual abnormalities within the provided anatomical regions, adding a secondary task sharing some of the features is a promising approach. When networks have been trained to predict different but related tasks in tandem, performance on each of the tasks have benefited [36]. The auxiliary information from a secondary task can be beneficial to the main task and is useful to regularize the network as well.

A big motivation behind using multi-task learning for this work is the availability of the radiologists' annotations for different abnormalities in the lung, localized to the six divisions (sextants) of the lungs. While this information will not be available during testing, it can be utilized to regularize the main model and to focus the attention of the network to supposedly relevant areas of the image. For our drug resistance classification, the network consists of a pretrained CNN network that is trained with binary cross entropy loss. For multi-tasked networks, the combined loss [37] for the two tasks is as follows:

$$\mathcal{L} = \frac{1}{\sigma_1^2} \times \mathcal{L}_1(W) + \frac{1}{\sigma_2^2} \times \mathcal{L}_2(W) + \log \sigma_1 + \log \sigma_2, \tag{1}$$

where W represents the weights of the network, and σ_1 and σ_2 are the noise parameters for the respective tasks, which are used to determine the relative weights given to each of the losses.

4.4. Abnormal Sextant Classification

The abnormal sextant information provides the locations of TB-related abnormalities to the network. These are expert-annotated features that provide additional context and information during DR/DS classification. Figure 6 shows the architecture for this type of multi-task learning. The classification model is modified such that the output of the last convolution layer is diverged into two stacks of fully-connected layers. The first path is the same as the normal architecture where the network decides if the X-ray image shows manifestations of drug resistance or drug sensitivity. The second path outputs a vector of length 6. Each of the six values represents the presence or absence of any of the 20 abnormality features described in the Data section above. Hence, for each sextant, if one of the abnormalities is present, the sextant is considered 'abnormal,' whereas if none of the abnormalities is present, it is considered 'normal.' In Equation (1), $\mathcal{L}_1$ and $\mathcal{L}_2$ are both

binary cross entropy losses in this case. For the secondary task, the loss is the average loss among all six outputs.

4.5. Abnormality Segmentation

Segmenting abnormalities provides location information as the locations of each of the sextants are also available. For this task, the losses from Equation 1 are modified such that $\mathcal{L}_1$ is the binary cross entropy loss and $\mathcal{L}_2$ is the combination of Jaccard loss and binary cross entropy loss for the segmentation of abnormal regions. Figure 7 shows the modification of the base model into an encoder-decoder U-Net style architecture for the additional task of abnormality segmentation. Two approaches are taken into account to determine the abnormal ground truth regions.

4.5.1. Sextant Segmentation from Radiologist Annotation

The sextant annotations from the radiologist are converted into masks such that the location information of each sextant is also available to the network. Each pixel in the sextant with presence of any abnormality is set to 1, and 0 otherwise. This is similar to the sextant classification with added information of the location of each of the sextants.

4.5.2. TB Abnormality Segmentation

Instead of using the abnormalities from clinical text data, we alternatively use the TB abnormality segmentation network to derive the ground truth. The Shenzhen dataset with annotations has a finer segmentation of abnormalities. In this approach, the chest X-ray images are segmented for lesions using the network trained on the Shenzhen data [30]. The advantage of this approach is that even images without annotations can be used for training.

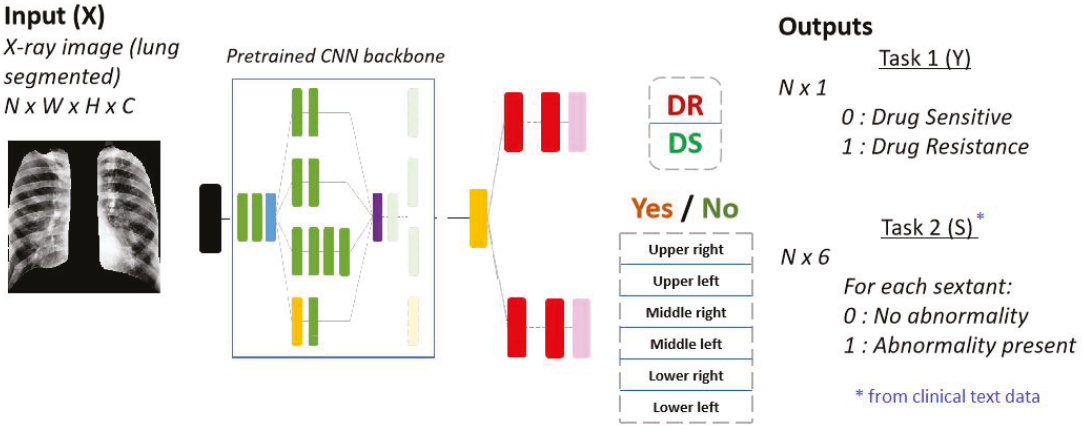


Figure 6. Multi-output network with the same pretrained backbone (ResNet18) for the additional task of abnormal sextant classification. The data used for this task is multi-modal. The inputs to the network are the chest X-ray images, whereas the labels for abnormal sextants are derived from the clinical text data described in Section 3.2.

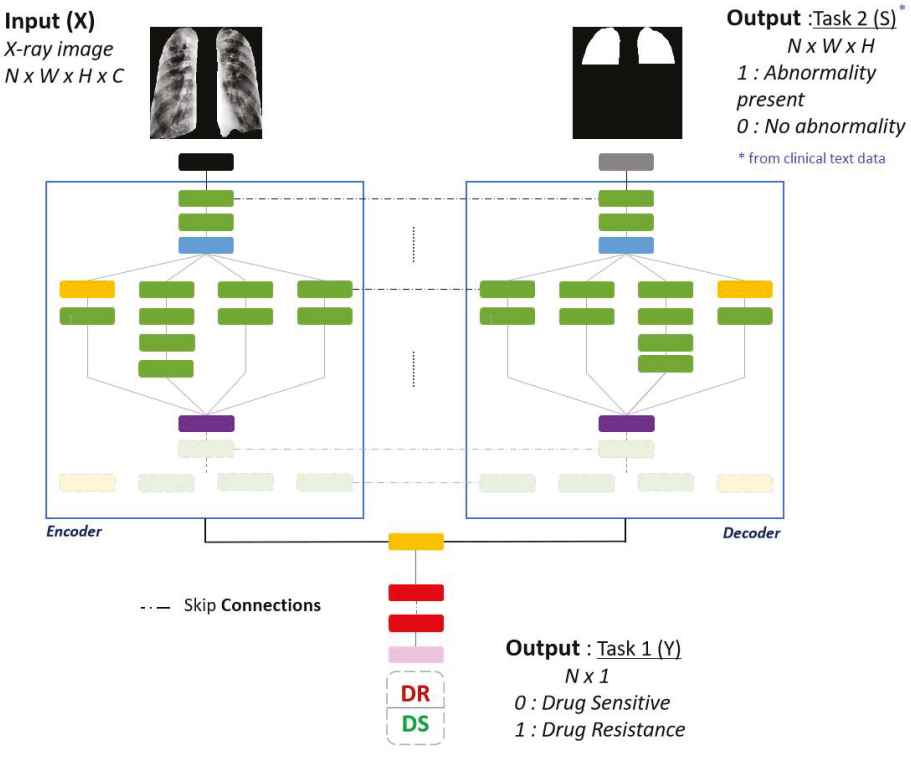


Figure 7. Multi-output network for the additional task of abnormality segmentation. The pretrained backbone (ResNet18) is modified to be a U-Net with encoder and decoder. The inputs for this network are chest X-ray images whereas the segmentation output masks are derived from the clinical text data.

5. Experimental Results

We perform our experiments with a 5-fold cross validation stratification. We also separate 7.5% of the training data, to check inter-epoch performance and stop the model early once the performance degrades for a long period on this set. The pretrained network backbone, as described in Figure 5, is the ResNet18 [33] architecture. The number of parameters for ResNet18 (11 million) are half of that of InceptionV3 (22.3 million), which we previously used [17]. Even with the smaller network and smaller dataset (since samples are held out), the performance on the validation set was 79% AUC. As we convert these networks to a U-Net style segmentation network for secondary tasks, the difference in parameters is increased even more. With the choice of ResNet18, we are able to transfer the pretrained weights from ImageNet [38] and keep the number of total trainable parameters small while having a consistent network to compare different approaches.

5.1. CNN-Based Drug Resistance Classification

To examine if our network is robust against domain changes and if it generalizes well to unseen data, we exclude the data from one country before cross validation stratification and use it as a held out set. As we see from Table 1, the only two countries that have more than 100 samples in each class are Belarus and Georgia. Aside from these two countries, every other country has less than 25 samples in the minority class. Choosing other countries would lead to a highly imbalanced testing set or a testing set with very few samples per class. When we excluded Georgia to use it as a held out set for evaluating generalization, the total number of samples decreased by almost half. There were only 479 samples per

class for the balanced training set. The AUC performance on the validation set was 78% but the performance on the held out Georgia data was at 52%. Also, less than 25% of the Georgia patients had sextant information available and hence the evaluation for the multi-task learning was not feasible. The data from Belarus has been used previously in [16], to both train and evaluate the classification of DR/DS TB from chest X-rays. Because of these reasons, we only used the data from Belarus as our held out set and used Georgia's data as part of the training sets.

For our classification training, we experimented with two different sets of initialization weights. The first set of weights is from the ImageNet classification task and the second set is from the network trained for TB-abnormality segmentation described in Section 5.6. When we trained the model with the cropped images (Figure 3c), similar to the previous approach [17], the performance on the validation set was 73% AUC and on the Belarus dataset it was 55% AUC.

In an effort to improve generalization, explicit attention on the lung regions was provided by setting the areas outside the segmented lung to 0 (Figure 3d). This approach improved the classification performance on each of the datasets. Table 3 shows that the best AUC performance was observed on the validation set, using both the ImageNet classification and TB abnormality segmentation weights, with 79% AUC. On the Belarus dataset, the best AUC of 65% was observed with the dataset with sextant information and with ImageNet weights. Achieving a much better performance with this approach, we use the segmented lungs as an input to the following experiments.

Table 3. DR-TB/DS-TB Classification Performance on the Validation Set and the Belarus Set.

		Validation Set		**Belarus Dataset**	
Trained on	***Initialization Weights***	**AUC**	**Accuracy**	**AUC**	**Accuracy**
Gen. Dataset	*ImageNet classification [38]*	0.79 ± 0.03	**0.72 ± 0.04**	0.60 ± 0.01	0.55 ± 0.03
	TB abnormality segmentation	**0.79 ± 0.02**	**0.72 ± 0.03**	0.60 ± 0.02	0.57 ± 0.02
Sext. Dataset	*ImageNet classification*	0.77 ± 0.03	**0.72 ± 0.03**	**0.65 ± 0.02**	**0.62 ± 0.02**

5.2. Classification with Radiomic Features

With the usage of non-learnable features, some acquisition-specific details can be hidden, which may be easily identified by a sufficiently large deep network. For this purpose, 104 radiomic features are extracted with the aid of the pyradiomics (https://pyradiomics.readthedocs.io/en/latest/features.html (accessed on 10 January 2022)) library [35]. The library calculates the features based on the X-ray image and the mask of the object of interest. We evaluate these features on both the lungs and the rest of the image by providing the lung masks and the complement of the lung masks, respectively. For our classifiers we use standard machine learning algorithms such as support vector classifiers (SVC), k-nearest neighbors (k-NN), Random Forest (RF) and multi-layer Perceptron (MLP).

The support vector classifier achieved the best performance on the validation set and the Belarus dataset, as seen in Table 4. Surprisingly, the best validation performance (74.5%) was computed when the lung region was excluded, that is, only non-lung parts of the image were used to derive these features. The performance on the Belarus dataset was 62.8%.

Table 4. AUC Performance with radiomic features.

Testing Data	**Input Data**	***SVC***	***k-NN***	***RF***	***MLP***
Validation Set (Gen. dataset)	Lung only	0.722	0.681	0.725	0.720
	Lungs excluded	**0.745**	0.688	0.732	0.725
Belarus dataset	Lung only	**0.628**	0.577	0.620	0.620
	Lung excluded	0.583	0.563	0.621	0.530

5.3. Classification with Sextant Divisions

The location of abnormalities are useful for classifying tuberculosis [39,40]. The sextant-based annotations are localized features that show different abnormalities within the lung. We also divide the chest X-rays into six divisions similar to how they were annotated by radiologists.

We classify DR-TB and DS-TB from the annotations acquired and our divided chest X-ray images. As described in Table 2, there are 20 such features for each of the sextants. Hence, there are 120 features in total. Figure 8 shows how abnormalities are more frequent in the apex of the lung.

Figure 9 shows the classification performance when individual sextants, the entire lung, and top, bottom, and middle regions were evaluated regarding DR-TB vs DS-TB classification. On the validation set, the CNN classifier trained on chest X-rays performed indiscriminately to the lung location used for training. With the annotated data and classical machine learning classifiers, the top sextants were more discriminatory than the bottom sextants. On the Belarus dataset, however, classical machine learning classifiers were not able to discriminate between the two classes with much success. An MLP (multi-layer Perceptron) classifier was able to achieve 60% AUC performance. When a single sextant was used, they all performed similarly at around 60%. Training on the entire image yielded the best results (65%) on the CXR images.

When we reduce the number of features to use just the location information or the type of abnormality, providing the location yielded better AUC performance (62.9%) on the Belarus dataset. However, on the validation set, providing the type of abnormality performed better (AUC of 70.0%) as shown in Table 5. 'Location' refers to the presence of any abnormality in sextants whereas 'Type' refers to the presence of one of the 20 abnormalities listed in Table 2 in any area of the lung.

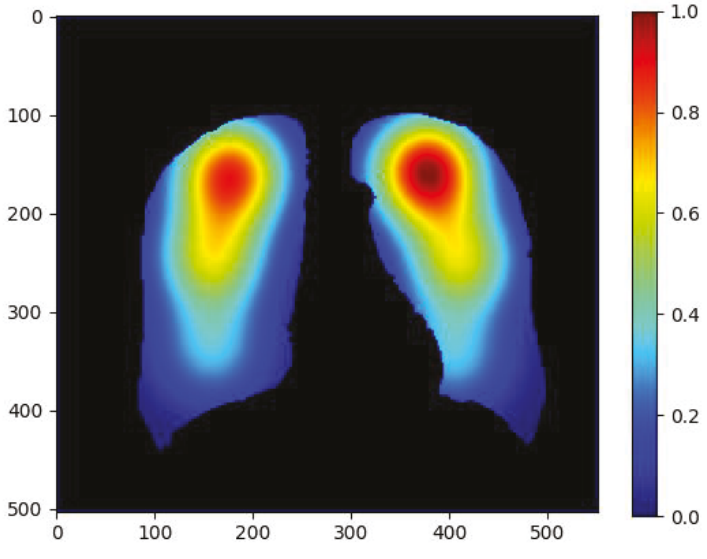


Figure 8. Abnormality occurrence heatmap in different regions of the lung derived from radiologists' annotated sextant data.

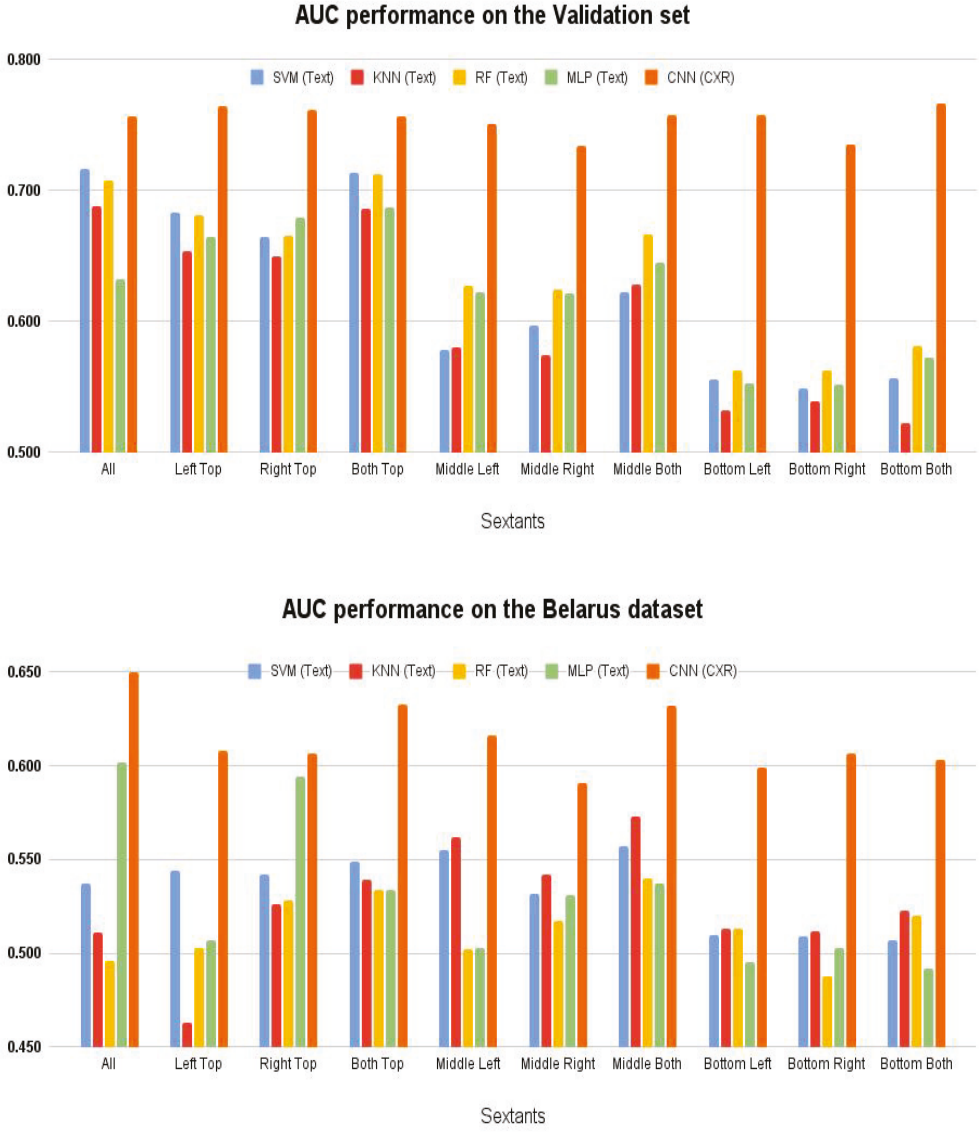


Figure 9. Drug Resistance Classification—AUC performance on sextants.

Table 5. Performance when location and type are separated for annotated radiologic features.

Classifier:	Validation Set		Belarus Dataset	
	Location	Type	Location	Type
SVC	0.573	0.700	**0.629**	0.539
KNN	0.501	0.670	0.484	0.506
RF	0.547	0.687	0.527	0.497
MLP	0.505	0.663	0.596	0.504

5.4. Classification with Data and Network Capacity Limitations

The classification performance with radiomic features derived from the non-lung region also prompted us to further examine the performance of our CNN classifier with limited information in Table 6. The limitations we added are regarding the input data and the training networks.

To further investigate the bias in the data that is supposedly not related to the underlying disease manifestations, we apply limitations to the information received by the network or limit the capacity of the network itself. This was achieved by modifying the data as well as the training network. Figure 10 shows examples of different ways the X-ray images were manipulated to reduce the information input to the classifier network. The information from the chest X-rays were limited or diminished by randomizing pixel locations. For example, in a particular experiment, entire image intensities were randomized. This would conceal the spatial relationships between pixels but still preserve the histograms and first order statistics of image intensity values. For further experiments, only certain regions of the image were randomized and the rest were set to 0. Lung masks were also used as input where the pixel intensity values are lost but the shape of the lungs are still intact. Another approach was subtracting the mean of the background (non-lung) and re-normalizing each image.

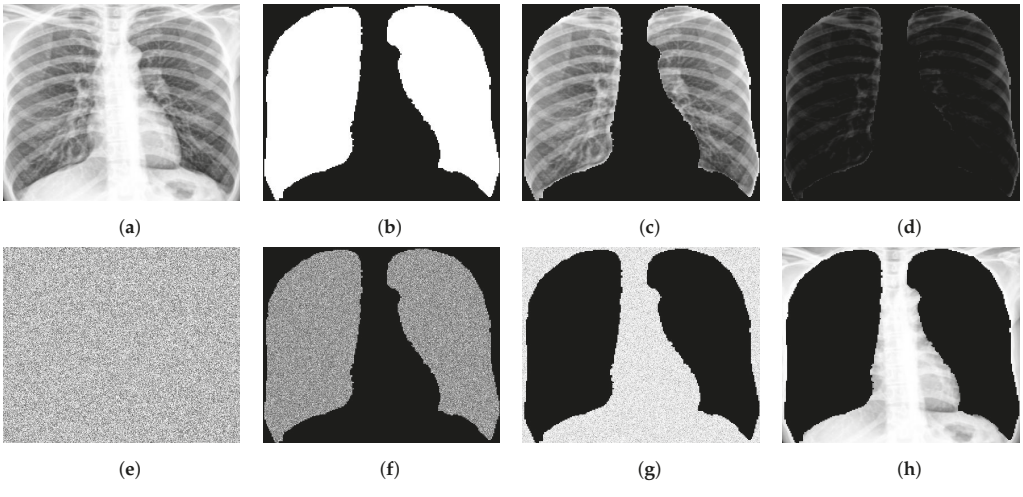


Figure 10. Different ways in which a chest X-ray image is modified to limit the information the classifier relies on. Top (left to right): (**a**) Cropped X-ray image, (**b**) lung mask, (**c**) segmented lung, (**d**) segmented lung with mean of background subtracted. Bottom (left to right): (**e**) Entire image randomized, (**f**) lung pixels randomized with non-lung area set to 0, (**g**) non-lung pixels randomized with lung pixels set to 0, and (**h**) lung pixels set to 0.

Table 6. Performance with input and model limitations.

Model Input	Validation Set	Belarus Dataset
Lung excluded	0.79	0.59
Histogram normalized (lung excluded)	0.74	0.58
Lung mask	0.61	0.55
Randomized pixels (entire image)	0.63	0.50
Randomized pixels (lung only)	0.64	0.50
Randomized pixels (lung excluded)	0.64	0.55
Frozen conv. layers (lung only)	0.66	0.62
Frozen conv. layers (lung excluded)	0.66	0.50
Background normalized (lung only)	0.75	0.61

To limit the network and its capacity, all the convolutional layers of the CNN were frozen (set to become non-trainable). These layers were used as a feature extractor and the fully-connected layers acted as a trainable classifier. When the lung pixel values were set to 0 (lung excluded), the performance on the verification test set (Belarus data) with the CNN network was 59%. Histogram normalization did not yield better results for this. As expected, when we completely randomized the pixel locations of the entire image or of the lung regions the performance was random (50%) on the balanced Belarus dataset. Using the shape of the lungs (lung masks) without the intensity values was enough to improve that performance to 55%. When intensity values of the non-lung regions (background) were provided, the performance further improved to 59%. On the same images, if the background pixels were randomized again, the performance dropped back to 55%. This series of results hints that the shape of the lung itself carries useful information. However, it was interesting that the non-lung regions had a small contribution to the identification of drug resistance. The performance on both datasets improved when the local information of the non-lung regions was retained.

Freezing the convolutional layer weights (ImageNet weights) but allowing the fully-connected layer weights to be trainable, the performance achieved was comparable at 62%. Here, the frozen convolutional layers are acting as fixed feature extractors and the dense layers are learning to interpret them. The approach and the performance were comparable to using the radiomic features. Normalizing by subtracting the mean of the non-lung regions from the lungs also had a similar performance of 61%.

5.5. Country Classification

Being able to identify the origin of the chest X-rays can be insightful in understanding the extent of bias in the data based on the data origin and the acquisition standards within a country. As seen in Table 1, the distribution of samples from different countries is not uniform and neither is the distribution of samples of each class. The chest X-ray images originate from a variety of imaging devices from hospitals in several different countries with their own imaging protocols and other variances that affect image content. These types of artifacts can often be identified by a deep neural network but may not be visible to a human observer. The entire dataset was used for the country classification experiments. Because the number of samples was imbalanced, we used weights based on the number of samples for the categorical cross entropy loss function during our training.

The mean intensities of the input images to the training network were plotted to observe the distribution differences across various countries in Figure 11. The width of each violin plot represents the frequency of the mean intensity. Generally, the wider the violin plot, the higher is the probability that the images of the respective country have the corresponding mean intensity.The histogram equalization centers the mean of the images intensities to zero for each of the countries and helps to reduce some of the bias present in the dataset due to the images' country of origin. The variance in the intensity distribution is still present.

The multi-class country-of-origin classification from the X-ray images achieves an accuracy of 85.7%. When histogram equalization is applied to the same images to remove some of the intensity-based biases, performance decreased to 82.6%. These results show that the models are very efficient in deriving the country of origin from the chest X-ray images. Even with histogram equalization, the country classification performance did not decrease sharply. This points to other biases in the data (potentially other acquisition biases) that are not accessible like the country of origin.

We also performed a multi-class country-of-origin classification based on clinical text data. Demographic features such as gender, age, and education, were included along with radiological findings from chest X-rays (such as nodules, cavities, infiltrates, collapses, etc.). The multi-class country-of-origin classification with seven demographic features and 20 radiological features resulted in an accuracy of 59.3%, and the classification with just 20 radiological features showed an accuracy of 35.2%. The confusion matrices in Figure 12

show that the image-derived features have many fewer false predictions when classifying the country of origin compared to the model trained on clinical features. The performance of this classifier is based on clinical findings whereas the deep learning classifier is using the image content which potentially includes information not related to the disease and not visible to the human observer but detectable by the network, allowing it to obtain better performance. Consequently, the DR/DS classifier also has access to this non-disease specific information which may introduce confounding features into that model, improving its performance on the trained domain but harming its generalizability.

5.6. Tuberculosis Abnormality Segmentation

Transfer learning with pretrained weights has been effectively used not only to reduce training time compared to random initialization but also to obtain a better performance. Intuitively, it makes sense to use TB abnormalities as priors to the drug resistance classification as well. Hence, we utilize the TB abnormality segmentation network to aide the classification of drug resistance, using its weights and output for the multi-task classification. We use the Shenzhen dataset with TB lung lesions [30], which has two sets of annotations for the same image.

The average segmentation overlap between the two sets of annotations was 0.538 (Dice score). For the TB abnormality segmentation network, the cross validation Dice score was 0.636. The weights for this network are used to initialize the classification and multi-task models. For the classification tasks, only the encoder weights were used, whereas for the segmentation tasks all the convolutional layer weights were used.

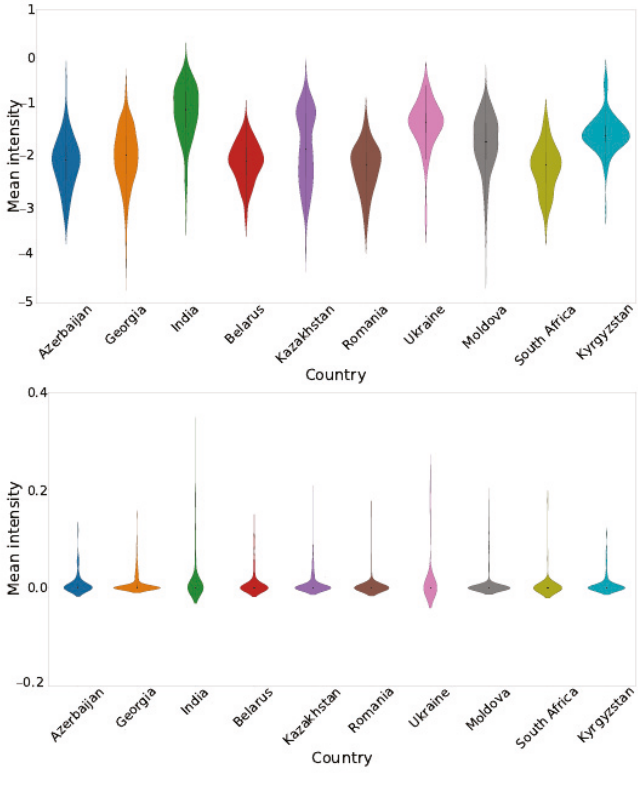


Figure 11. Mean intensity distributions per country for the normalized cropped X-ray images (**top**) and the same images after histogram equalization is applied (**bottom**).

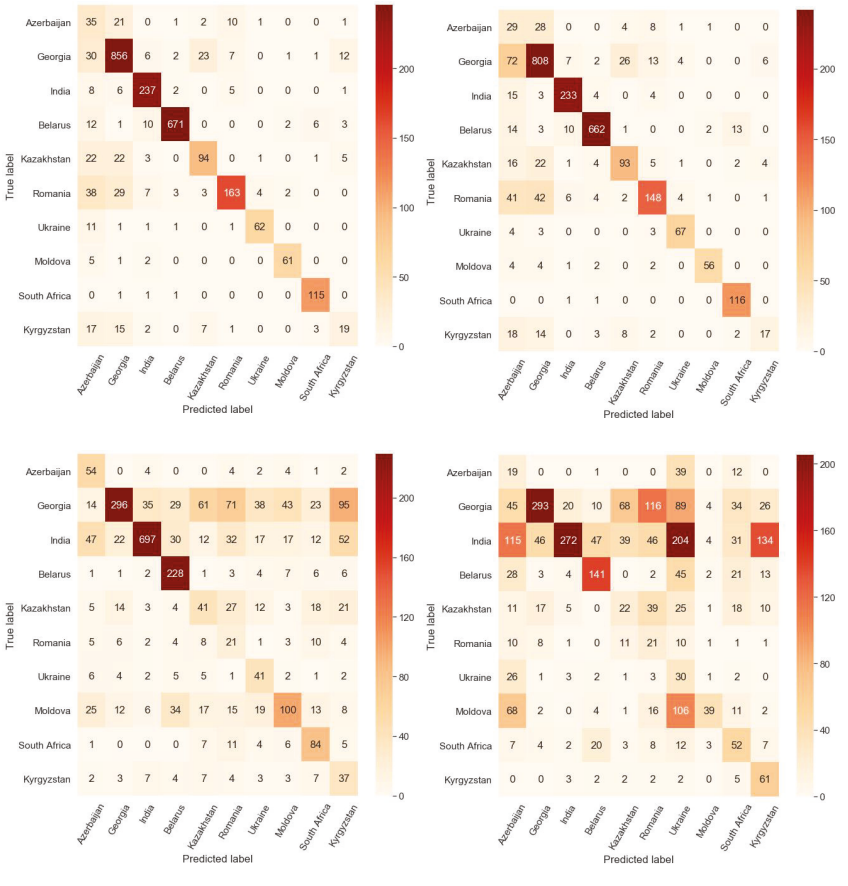


Figure 12. Summation of confusion matrices on test sets from 5 folds on the country-of-origin classification task. (**top-left**) Lung-segmented images, (**top-right**) histogram-equalized lung-segmented images, (**bottom-left**) radiological and demographic features and (**bottom-right**) radiological features. Note that the image-derived features have a significantly better performance than the disease-specific radiological findings.

Table 7 shows that the annotations by radiologists matched with the abnormalities identified by the TB abnormality segmentation model in each of the sextants and overall. The average GradCAM [41] map values for each of the sextants was poorly correlated with sextant annotations. Abnormality predictions were also derived from the mean of the GradCAM heatmaps (with a threshold of 0.7) and compared with the abnormalities from the TB segmentation model. On both of these, the top sextants had very little overlap compared with the middle and bottom sextants. This result is contrary to expectation as TB abnormalities are often seen at the top of the lung [42].

Figure 13 shows some examples of DR-TB and DS-TB. GradCAM heatmaps show the average of activations from the last layer of a CNN. In these examples, it can be seen that for the DR-TB class, activations are more frequent. This is consistent with previous results [9], where DR-TB patients are shown to have more abnormalities in their lungs.

Table 7. Comparison (AUC) of localized abnormalities from different sources tested on Belarus dataset.

Ground Truth:	**Sextant Annotations**	**Sextant Annotations**	**GradCAM (Predictions)**
Prediction:	**TB Abnormal Seg. Prob. Map**	**GradCAM (Heatmap)**	**TB Abnormal Seg. Prob. Map**
Top Right	0.788	0.469	0.518
Top Left	0.742	0.504	0.567
Middle Right	0.759	0.627	0.692
Middle Left	0.759	0.630	0.716
Bottom Right	0.808	0.642	0.699
Bottom Left	0.758	0.571	0.665
Overall	0.769	0.574	0.643

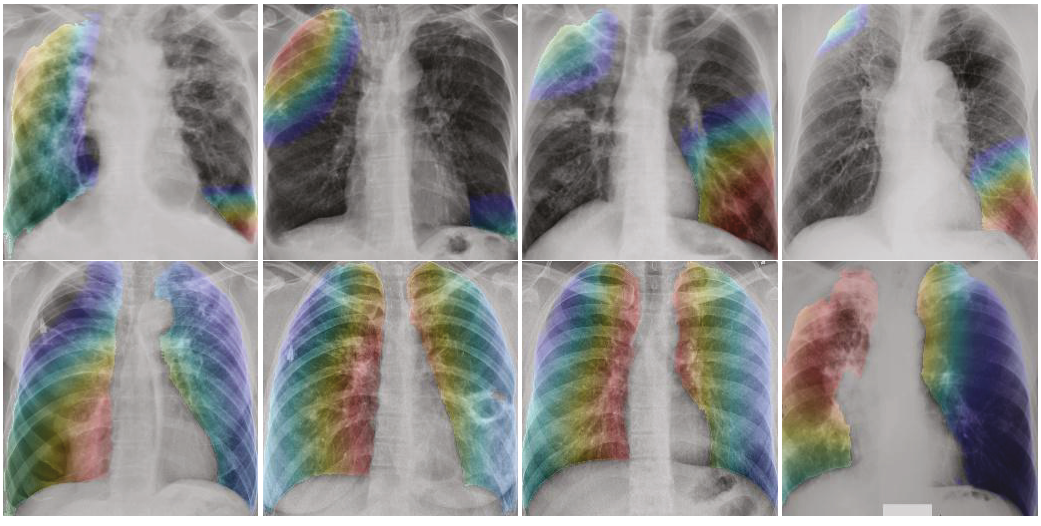


Figure 13. GradCAM visualizations on X-ray images. **Top row** shows examples of drug-sensitive TB and **bottom row** shows examples of drug-resistant TB.

5.7. Multi-Task Based Classification

As an approach to assist the drug resistance classifier, a secondary task was added to further incentivize the network to focus on relevant areas. The second task acts as a regulator to constraint the neural network to focus on the interesting regions. ResNet18 was used as the primary backbone for the networks and layers were added to generate output for the secondary tasks. Instead of experimentally determining the best loss weights for the model, we allow them to be learnt by the model itself.

Initially, both tasks were given equal weights to allow the network to determine which task needs to be focused. To restrain the scope of the work, while we monitored the performance on the secondary class, we only used the performance on the main task to determine our experimental setup and hence we report the performance on the main task only.

While adding a related secondary task did not improve the drug resistance classification on the validation set, the AUC performance on the Belarus dataset improved by about 2%–3% and accuracy improved by 1%. An AUC of 68% (±1%) was the best performance achieved on the Belarus dataset as shown in Table 8.

Table 8. Classification performance with an additional task.

		Validation Set		Belarus Dataset	
Trained on	**Secondary Task**	**AUC**	**Accuracy**	**AUC**	**Accuracy**
Sext. Dataset	Abnormal Sextant Classification	0.77 ± 0.02	**0.70 ± 0.02**	0.64 ± 0.01	0.61 ± 0.03
	Abnormal Sextant Segmentation	**0.78 ± 0.02**	**0.70 ± 0.02**	0.67 ± 0.01	**0.63 ± 0.01**
Gen. Dataset	TB Abnormalities Segmentation	0.77 ± 0.02	0.69 ± 0.02	**0.68 ± 0.01**	**0.63 ± 0.02**

6. Conclusions

This paper explores the cross validation and generalization performance achieved for drug-sensitive and drug-resistant classification on chest X-rays from different countries. By excluding data from one country of origin from training and using it for testing, we evaluate classifier performance on unseen data. The generalization performance was much lower (65% AUC) compared to the cross validation performance (79% AUC). The same CNN architecture was able to classify the country of origin from a chest X-ray image. Evaluations with radiomic features from X-ray images, and experimental limitations to the data and classifier, indicated that the model based its decisions on other artifacts present in the images. TB lesions annotated by radiologists were utilized to see if the location information was useful for discriminating between drug-resistant and drug-sensitive cases. While GradCAM heatmaps from the X-ray image-based CNN model did not overlap significantly with the TB lesions and the annotations from radiologists, adding a secondary task related to the localization of lesions did improve the classification performance to 68% AUC. Because of an imbalanced dataset, insufficient amount of samples of one of the two classes, and the lack of clinical text data describing the radiological findings for all the patients, we only excluded one single country for our generalization evaluation. A solution that does not require annotations by radiologists to improve the generalization performance would be more valuable. Procedures and methods that allow the model to pick up only the manifestations of disease are a direction for future research. In general, we believe that experiments addressing generalization to new datasets should be standard practice in medical image analysis with deep learning.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Data usage is exempt from local institutional review board review as it is publicly available from the TB portals program. The TB portals program participants are responsible for ensuring compliance with their countries laws, regulations, and ethics considerations.

Data Availability Statement: Links to datasets used in this study are provided in Section 3.

Acknowledgments: This work was supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF), under Interagency Agreement #750119PE080057, and by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. This project has also been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases under BCBB Support Services Contract HHSN316201300006W/HHSN27200002.

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

References

1. World Health Organization. *Global Tuberculosis Report* ; World Health Organization: Geneva, Switzerland, 2020; p. xiii.
2. Qin, Z.Z.; Sander, M.S.; Rai, B.; Titahong, C.N.; Sudrungrot, S.; Laah, S.N.; Adhikari, L.M.; Carter, E.J.; Puri, L.; Codlin, A.J.; et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci. Rep.* **2019**, *9*, 15000. [CrossRef]
3. Qin, Z.Z.; Ahmed, S.; Sarker, M.S.; Paul, K.; Adel, A.S.S.; Naheyan, T.; Barrett, R.; Banu, S.; Creswell, J. Tuberculosis detection from chest X-rays for triaging in a high tuberculosis-burden setting: An evaluation of five artificial intelligence algorithms. *Lancet Digit. Health* **2021**, *3*, e543–e554. [CrossRef]
4. Wàng, Y.X.J.; Chung, M.J.; Skrahin, A.; Rosenthal, A.; Gabrielian, A.; Tartakovsky, M. Radiological signs associated with pulmonary multi-drug resistant tuberculosis: An analysis of published evidences. *Quant. Imaging Med. Surg.* **2018**, *8*, 161–173. [CrossRef]
5. Icksan, A.G.; Napitupulu, M.R.S.; Nawas, M.A.; Nurwidya, F. Chest X-ray findings comparison between multi-drug-resistant tuberculosis and drug-sensitive tuberculosis. *J. Nat. Sci. Biol. Med.* **2018**, *9*, 42.
6. Huang, X.L.; Skrahin, A.; Lu, P.X.; Alexandru, S.; Crudu, V.; Astrovko, A.; Skrahina, A.; Taaffe, J.; Harris, M.; Long, A.; et al. Prediction of multiple drug resistant pulmonary tuberculosis against drug sensitive pulmonary tuberculosis by CT nodular consolidation sign. *bioRxiv* **2019**. [CrossRef]
7. Flores-Trevino, S.; Rodriguez-Noriega, E.; Garza-Gonzalez, E.; Gonzalez-Diaz, E.; Esparza-Ahumada, S.; Escobedo-Sanchez, R.; Perez-Gomez, H.R.; Leon-Garnica, G.; Morfin-Otero, R. Clinical predictors of drug-resistant tuberculosis in Mexico. *PLoS ONE* **2019**, *14*, e0220946.
8. Cheng, N.; Wu, S.; Luo, X.; Xu, C.; Lou, Q.; Zhu, J.; You, L.; Li, B. A Comparative Study of Chest Computed Tomography Findings: 1030 Cases of Drug-Sensitive Tuberculosis versus 516 Cases of Drug-Resistant Tuberculosis. *Infect. Drug Resist.* **2021**, *14*, 1115–1128. [CrossRef]
9. Yang, F.; Yu, H.; Kantipudi, K.; Karki, M.; Kassim, Y.M.; Rosenthal, A.; Hurt, D.E.; Yaniv, Z.; Jaeger, S. Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features. *Quant. Imaging Med. Surg.* **2022**, *12*, 675–687. [CrossRef]
10. Ionescu, B.; Müller, H.; Villegas, M.; de Herrera, A.G.S.; Eickhoff, C.; Andrearczyk, V.; Cid, Y.D.; Liauchuk, V.; Kovalev, V.; Hasan, S.A.; et al. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 309–334.
11. Gentili, A. ImageCLEF2018: Transfer Learning for Deep Learning with CNN for Tuberculosis Classification. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
12. Ishay, A.; Marques, O. ImageCLEF 2018 Tuberculosis Task: Ensemble of 3D CNNs with Multiple Inputs for Tuberculosis Type Classification. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
13. Cid, Y.D.; Müller, H. Texture-based Graph Model of the Lungs for Drug Resistance Detection, Tuberculosis Type Classification, and Severity Scoring: Participation in ImageCLEF 2018 Tuberculosis Task. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
14. Allaouzi, I.; Ahmed, M.B. A 3D-CNN and SVM for Multi-Drug Resistance Detection. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
15. Ureta, J.; Shrestha, A. Identifying drug-resistant tuberculosis from chest X-ray images using a simple convolutional neural network. *J. Phys. Conf. Ser.* **2021**, *2071*, 012001. [CrossRef]
16. Jaeger, S.; Juarez-Espinosa, O.H.; Candemir, S.; Poostchi, M.; Yang, F.; Kim, L.; Ding, M.; Folio, L.R.; Antani, S.; Gabrielian, A.; et al. Detecting drug-resistant tuberculosis in chest radiographs. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1915–1925. [CrossRef]
17. Karki, M.; Kantipudi, K.; Yu, H.; Yang, F.; Kassim, Y.M.; Yaniv, Z.; Jaeger, S. Identifying Drug-Resistant Tuberculosis in Chest Radiographs: Evaluation of CNN Architectures and Training Strategies. In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Guadalajara, Mexico, 1–5 November 2021; IEEE: Piscataway, NJ, USA, 2021.
18. Pooch, E.H.; Ballester, P.L.; Barros, R.C. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv* **2019**, arXiv:1909.01940.
19. Castro, D.C.; Walker, I.; Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **2020**, *11*, 3673. [CrossRef] [PubMed]
20. Harris, M.; Qi, A.; Jeagal, L.; Torabi, N.; Menzies, D.; Korobitsyn, A.; Pai, M.; Nathavitharana, R.R.; Ahmad Khan, F. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest X-rays for pulmonary tuberculosis. *PLoS ONE* **2019**, *14*, e0221339. [CrossRef] [PubMed]
21. Sathitratanacheewin, S.; Sunanta, P.; Pongpirul, K. Deep learning for automated classification of tuberculosis-related chest X-Ray: Dataset distribution shift limits diagnostic performance generalizability. *Heliyon* **2020**, *6*, e04614. [CrossRef]

22. Rajpurkar, P.; Joshi, A.; Pareek, A.; Chen, P.; Kiani, A.; Irvin, J.; Ng, A.Y.; Lungren, M.P. CheXpedition: Investigating generalization challenges for translation of chest X-ray algorithms to the clinical setting. *arXiv* **2020**, arXiv:2002.11379.
23. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **2018**, *15*, e1002683. [CrossRef]
24. Badgeley, M.A.; Zech, J.R.; Oakden-Rayner, L.; Glicksberg, B.S.; Liu, M.; Gale, W.; McConnell, M.V.; Percha, B.; Snyder, T.M.; Dudley, J.T. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit. Med.* **2019**, *2*, 31. [CrossRef]
25. Ahmed, K.B.; Goldgof, G.M.; Paul, R.; Goldgof, D.B.; Hall, L.O. Discovery of a Generalization Gap of Convolutional Neural Networks on COVID-19 X-Rays Classification. *IEEE Access* **2021**, *9*, 72970–72979. [CrossRef]
26. Rosenthal, A.; Gabrielian, A.; Engle, E.; Hurt, D.E.; Alexandru, S.; Crudu, V.; Sergueev, E.; Kirichenko, V.; Lapitskii, V.; Snezhko, E.; et al. The TB portals: An open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis. *J. Clin. Microbiol.* **2017**, *55*, 3267–3282. [CrossRef] [PubMed]
27. Dodd, P.J.; Looker, C.; Plumb, I.D.; Bond, V.; Schaap, A.; Shanaube, K.; Muyoyeta, M.; Vynnycky, E.; Godfrey-Faussett, P.; Corbett, E.L.; et al. Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection. *Am. J. Epidemiol.* **2015**, *183*, 156–166.
28. Yates, T.A.; Atkinson, S.H. Ironing out sex differences in tuberculosis prevalence. *Int. J. Tuberc. Lung Dis.* **2017**, *21*, 483–484. [CrossRef]
29. Hertz, D.; Schneider, B. Sex differences in tuberculosis. *Semin. Immunopathol.* **2019**, *41*, 225–237. [CrossRef]
30. Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.X.J.; Lu, P.X.; Thoma, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475.
31. Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.I.; Matsui, M.; Fujita, H.; Kodera, Y.; Doi, K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AJR Am. J. Roentgenol.* **2000**, *174*, 71–74. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Williams, M.B.; Krupinski, E.A.; Strauss, K.J.; Breeden, W.K., III; Rzeszotarski, M.S.; Applegate, K.; Wyatt, M.; Bjork, S.; Seibert, J.A. Digital radiography image quality: Image acquisition. *J. Am. Coll. Radiol.* **2007**, *4*, 371–388. [CrossRef] [PubMed]
35. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
36. Argyriou, A.; Evgeniou, T.; Pontil, M. Convex multi-task feature learning. *Mach. Learn.* **2008**, *73*, 243–272. [CrossRef]
37. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.
38. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
39. Kim, H.Y.; Song, K.S.; Goo, J.M.; Lee, J.S.; Lee, K.S.; Lim, T.H. Thoracic sequelae and complications of tuberculosis. *Radiographics* **2001**, *21*, 839–858. [CrossRef] [PubMed]
40. Nachiappan, A.C.; Rahbar, K.; Shi, X.; Guy, E.S.; Mortani Barbosa, E.J., Jr.; Shroff, G.S.; Ocazionez, D.; Schlesinger, A.E.; Katz, S.I.; Hammer, M.M. Pulmonary tuberculosis: Role of radiology in diagnosis and management. *Radiographics* **2017**, *37*, 52–72. [CrossRef] [PubMed]
41. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
42. Bennett, J.E.; Dolin, R.; Blaser, M.J. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases E-Book*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2019.

diagnostics

MDPI

Article

Deep Learning Supplants Visual Analysis by Experienced Operators for the Diagnosis of Cardiac Amyloidosis by Cine-CMR

Philippe Germain [1,*], Armine Vardazaryan [2,3], Nicolas Padoy [2,3], Aissam Labani [1], Catherine Roy [1], Thomas Hellmut Schindler [4] and Soraya El Ghannudi [1,5]

1 Department of Radiology, Nouvel Hopital Civil, University Hospital, 67000 Strasbourg, France; aissam.labani@chru-strasbourg.fr (A.L.); catherine.roy@chru-strasbourg.fr (C.R.); soraya.elghannudi-abdo@chru-strasbourg.fr (S.E.G.)
2 ICube, University of Strasbourg, CNRS, 67000 Strasbourg, France; vardazaryan@unistra.fr (A.V.); npadoy@unistra.fr (N.P.)
3 IHU (Institut Hopitalo-Universitaire), 67000 Strasbourg, France
4 Mallinckrodt Institute of Radiology, Division of Nuclear Medicine, Washington University School of Medicine, Saint Louis, MO 63110, USA; thschindler@wustl.edu
5 Department of Nuclear Medicine, Nouvel Hopital Civil, University Hospital, 67000 Strasbourg, France
* Correspondence: germain.philippe7@gmail.com

Abstract: Background: Diagnosing cardiac amyloidosis (CA) from cine-CMR (cardiac magnetic resonance) alone is not reliable. In this study, we tested if a convolutional neural network (CNN) could outperform the visual diagnosis of experienced operators. Method: 119 patients with cardiac amyloidosis and 122 patients with left ventricular hypertrophy (LVH) of other origins were retrospectively selected. Diastolic and systolic cine-CMR images were preprocessed and labeled. A dual-input visual geometry group (VGG) model was used for binary image classification. All images belonging to the same patient were distributed in the same set. Accuracy and area under the curve (AUC) were calculated per frame and per patient from a 40% held-out test set. Results were compared to a visual analysis assessed by three experienced operators. Results: frame-based comparisons between humans and a CNN provided an accuracy of 0.605 vs. 0.746 ($p < 0.0008$) and an AUC of 0.630 vs. 0.824 ($p < 0.0001$). Patient-based comparisons provided an accuracy of 0.660 vs. 0.825 ($p < 0.008$) and an AUC of 0.727 vs. 0.895 ($p < 0.002$). Conclusion: based on cine-CMR images alone, a CNN is able to discriminate cardiac amyloidosis from LVH of other origins better than experienced human operators (15 to 20 points more in absolute value for accuracy and AUC), demonstrating a unique capability to identify what the eyes cannot see through classical radiological analysis.

Keywords: cardiac amyloidosis; AL/TTR amyloidosis; hypertrophic cardiomyopathy; left ventricular hypertrophy; deep learning; convolutional neural network

Citation: Germain, P.; Vardazaryan, A.; Padoy, N.; Labani, A.; Roy, C.; Schindler, T.H.; El Ghannudi, S. Deep Learning Supplants Visual Analysis by Experienced Operators for the Diagnosis of Cardiac Amyloidosis by Cine-CMR. *Diagnostics* **2022**, *12*, 69. https://doi.org/10.3390/diagnostics12010069

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 9 December 2021
Accepted: 27 December 2021
Published: 29 December 2021

1. Introduction

Cardiac amyloidosis (CA) is a specific cardiomyopathy caused by the deposition of misfolded amyloid fibrils in the extracellular myocardial space. Light-chain (AL) and transthyretin (TTR) are the most common subtypes. Cardiac amyloidosis is a fatal disease requiring rapid diagnosis for patients to benefit from recently released medications [1–3]. Its diagnosis has gained significant improvements in recent years, in particular with the recognition of diphosphonate SPECT imaging for the identification of the TTR form of the disease [4].

MRI plays an important role in this field thanks to gadolinium injections providing quite a specific pattern of myocardial late-enhancement [5] and demonstrating highly relevant extracellular volume (ECV) increase [6]. Despite recent relief in the restrictions

on the use of gadolinium chelates [7], caution needs to be exercised in case of renal impairment, and a diagnostic approach without injection would be beneficial. Steady-state free precession (SSFP) cine-CMR is a basic method in cardiac MRI, offering a good quality morphological and functional depiction of important cardiac features [8]. Myocardial wall thickening, atrial enlargement and pericardial or pleural effusion constitute the hallmarks of amyloid cardiac involvement [9]. However, these signs are very nonspecific since they are also seen in many other etiologies of left ventricular hypertrophy such as advanced hypertensive disease, aortic stenosis and other overload diseases such as Fabry disease and sarcomeric hypertrophic cardiomyopathies, which is why cine-CMR alone is not recognized as effective for diagnosing cardiac amyloidosis.

Machine learning and, particularly, deep learning applied to imaging quickly established themselves in most pathological areas, and these methods are now recognized as having diagnostic capacities similar to experienced radiologists, particularly in cardiomyopathies [10] and cardiac amyloidosis [11]. An even more interesting fact concerns the superior diagnostic capacities of deep learning over human readers in some fields, such as breast cancer [12], especially its ability to identify pathologies invisible to the naked eye, such as abnormalities discernible only in immunohistochemistry or through genetic analysis. For example, deep learning was reported to be efficient in improving mutation prediction in hypertrophic cardiomyopathy using MR-cine images [13].

This innovative concept led us to initiate the present study in which we compared the performance of commonly available deep learning methods to experienced radiologists to discriminate cardiac amyloidosis from other myocardial hypertrophies based on cine-CMR alone. Moreover, we explored the capacity of deep learning to differentiate AL from TTR amyloidosis, which is not reliably achievable visually with cine-CMR.

2. Materials and Methods

2.1. Study Population

We retrospectively analyzed the cine-CMR sequences of patients performed between 2010 and 2020 in the radiology department of our hospital. This study was registered and approved by the Institutional Review Board of our university hospital, and all datasets were obtained and de-identified, with waived consent in compliance with the rules of our institution. The cine-CMR exams of 241 patients were studied, including 119 with histologically proven amyloidosis and 122 with left ventricular hypertrophy without amyloidosis (LVH). The patients' characteristics are listed in Table 1.

The left ventricular hypertrophy without amyloidosis group (n = 122) consisted of patients referred to CMR for suspected cardiac amyloidosis due to several suggestive features such as a heart failure episode, thickening of the myocardial walls on ultrasound examination, restrictive transmitral Doppler filling pattern, reduced longitudinal strain with apical sparring, monoclonal gammopathy or dubious Perugini grade 1 bone scintigraphy. Other cases presented a CMR of concentric left ventricular hypertrophy (left ventricular wall thickness $\geq$13 mm in diastole). The clinical context was consistent with hypertension, aortic stenosis or non-obstructive hypertrophic cardiomyopathy. Late-enhancement imaging obtained in all cases never demonstrated circumferential subendocardial or diffuse late-enhancement patterns suggestive of amyloid involvement.

For the amyloidosis group (n = 119), the selection criteria for amyloidosis diagnosis were based on typical CMR features confirmed by clinical, biological, bone scintigraphic and anatomo-histological findings. Left ventricular wall thickening ($\geq$13 mm in diastole), left $\pm$ right atrial dilatation, increased native myocardial T1 relaxation time and/or extracellular volume (ECV), pericardial or pleural effusion and typical subendocardial late-enhancement pattern (circumferential, diffuse or not related to a coronary territory) were the main diagnostic clues for amyloidosis.

Table 1. Clinical and CMR characteristics of the study population.

	Amyloidosis	**LVH**	***p***
N patients	119	122	
Age (years)	74.65 ± 9.53	59.50 ± 14.34	0.0001
Sex (F/M)	31/88	39/83	0.31
Weight (kg)	70.80 ± 15.16	82.95 ± 20.50	0.0001
Height (m)	169.9 ± 8.84	170.36 ± 10.05	0.78
BSA (m^2)	1.84 ± 0.22	2.00 ± 0.27	<0.0001
IVS (mm)	18.11 ± 3.54	18.38 ± 3.54	0.56
LVMI (g/m^2)	115.96 ± 29.08	116.58 ± 31.43	0.88
LVDVI (mL/m^2)	69.88 ± 22.21	74.51 ± 20.82	0.36
LVEF (%)	58.96 ± 10.93	67.33 ± 12.18	<0.0001
LA surface (cm^2)	31.55 ± 5.23	25.47 ± 5.96	0.0002
Systolic time (ms)	321 ± 39	332 ± 40	0.095
T1 (ms)	1138.5 ± 48.1	1038.0 ± 56.2	<0.0001
ECV (%)	53.97 ± 11.17	26.89 ± 4.00	<0.0001
N long axis frames/patient	2.24 ± 0.93	2.22 ± 0.94	0.93
N short axis frames/patient	3.41 ± 1.45	3.59 ± 1.27	0.49
N frames/patient	5.68 ± 1.85	5.47 ± 1.81	0.58
N frame post-gadolinium	171/676	167/667	0.96
N patient with pericard.	54 (45%)	27 (22%)	0.00013
N patients with pleural.	45 (38%)	10 (8%)	0.00001
N patients with both.	24 (20%)	3 (2.5%)	0.00001

The characteristics of patients with amyloidosis and left ventricular hypertrophy were included in this study. The number of observations, (integer) or average values ± standard deviation, are listed: BSA: body surface area; IVS: interventricular septum thickness; LVMI: left ventricular mass index; LVDVI: left ventricular diastolic volume index; LVEF: left ventricular ejection fraction; LA: left atrial; systolic time: the time of the systolic image; and ECV: extracellular volume. Between the parentheses is the percentage. Pericard. is for pericardial effusion, pleural is for pleural effusion and both is for pericardial + pleural effusions.

The characteristics of AL and TTR patients can be found in the supplemental material (Table S1). TTR amyloidosis was defined in 38 patients without monoclonal gammopathy and with a ^{99m}Tc-diphosphonate SPECT Perugini score of >1 or with amyloid deposits on an extracardiac and/or endomyocardial biopsy. AL amyloidosis was reported in 59 cases, based on the detection of a kappa/lambda free light-chain with monoclonal gammopathy and an extracardiac and/or endomyocardial biopsy. Among the 22 patients who were not categorized as AL or TTR, three were AA type, three had uncertain immunostaining, one had Perugini 1 and no gammopathy, four elderly patients died and 11 were lost to follow-up.

For the cine-CMR acquisitions, all images were obtained at 1.5 Tesla, using three Siemens (Erlangen, Germany) and one Philips (Eindhoven, The Netherlands) scanners. Steady-state free precession (SSFP) cine sequences were obtained with TE/TR 1.6/3.5 ms, 8 to 32 elements cardiac coil and 6 to 8 mm thick slices. End-systole (with the smallest left ventricular dimension) was visually selected (systolic time in Table 1). Orientation planes were long axis (4-chamber and vertical 2-chamber views) and short axis views. Table 1 lists the summary of acquisition parameters.

2.2. Image Preparation

The image preparation of cine studies exported from the PACS of our hospital was carried out with a dedicated Visual C software. All images were first de-identified and resampled (bilinear) in order to obtain a normalized homogeneous pixel size of 1.5 mm. The images' intensity windowing was manually focused on the central cardiac region of interest. Diastolic and systolic frames were selected. Epicardial contours (ROI_epi) and myocardial contours (ROI_myo) were manually drawn.

Finally, five pairs of images (cropped to 128 and 160 pixels, full view 256 pixels, ROI_epi and ROI_endo), as illustrated in Figure 1, were stored. The purpose of these tests (especially for ROIs) was to determine if a focused analysis led to better classification performance. Labeling (orientation plane, pathology, presence of effusion and gadolinium injection) was carried out simultaneously and saved in the labeled file.

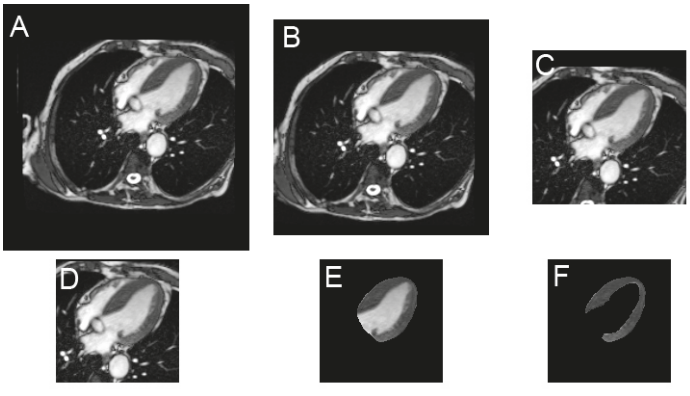


Figure 1. Example of input shapes submitted to the CNN, with native 256 × 256 full image format (**A**), 224 × 224 cropped image (**B**), 160 × 160 cropped image (**C**), 128 × 128 cropped image (**D**), epicardial region of interest (ROI) image (**E**) and myocardial ROI (**F**).

2.3. Deep Learning Process

CNN implementation was performed in Python 3.7.6, with Keras library and TensorFlow backend. According to CLAIM recommendations [14], the data were distributed in order to ensure that images of the same patient always lie in either the train set, the validation set or the test set (no mixture between these sets).

For hyperparameter trimming, data processing was performed according to the diagram shown in Figure 2. A 40% test set (538 pairs of frames and 96 patients) was isolated and stored as a held-out test set. With the 60% remaining data, a three-fold cross-validation training was performed in order to trim hyperparameters (batch size, optimizer, learning rate, decay, number of trainable layers, dropout rate and parameters of the image data generator). This was done to avoid the influence of individual training and validation examples on the choice of hyperparameters.

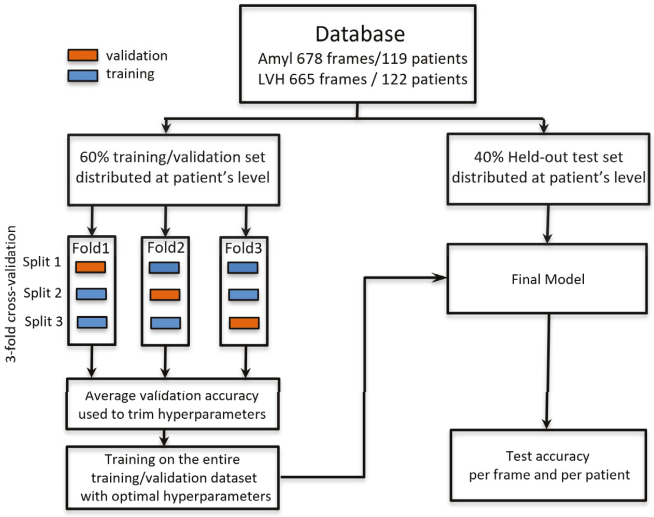


Figure 2. Schematic view of the processing method used in order to strictly separate training/validation data and test data.

With optimal hyperparameters, a final model was built with all training data and evaluated on the test set. Patient-based metrics were calculated from the average of the predicted probability corresponding to all frames of a unique patient.

A VGG16 [15] base model was used and trained from scratch for diastolic and systolic frames. The two outputs (diastole and systole) were concatenated and followed by the following layers, where ReLU non-linearity was used after each Dense layer except the last one: Flatten, Dense 256, Dropout 0.40, Dense 128, Dropout 0.45, Dense 64, Dropout 0.50, Dense 1 and output Sigmoid activation layer. In the final model, training was done with batch size: 32; number of epochs: 150; optimizer: SGD; LR 4 $10^{-5.}$; and decay: 10^{-6}. Binary cross entropy was used as a loss function. The parameters of data augmentation applied during training were zoom range <0.15, 15% height and width shift range and up to 20° rotation.

The Grad-CAM algorithm [16] was used to visualize class activation maps. With this algorithm, the identification of the most contributive pixels involved for each class is related to the gradient information flowing into the final convolutional layer of the network.

2.4. Experienced Radiologists/Cardiologists Blind Reading

The blind reading of diastolic and systolic images was performed by one radiologist and two cardiologists (>10 years' experience of CMR analysis and reporting). Frame-based reading was obtained from the pairs of images corresponding to the test set. Patient-based reading was obtained from the whole dataset (241 patients), and paired comparisons were made with the 40% held-out test set (average number of frame pairs, 5.5 per patient).

2.5. Evaluation and Statistical Analysis

The performance metrics—computed on a frame-basis and a patient-basis—were test accuracy, sensitivity, specificity, confusion matrices, receiver operating characteristic (ROC) curves and precision-recall curves with the corresponding area under the curve (AUC) values. Testing the relationship between categorical variables (e.g., accuracy comparisons) was carried out with a Chi-square test. A comparison of the quantitative values was performed with Student's *t*-test, and a comparison of the AUC of ROC curves was performed with the Delong test. MedCalc 12.1.4 (MedCalc Software, Ostend, Belgium) was used for statistical analyses.

3. Results

3.1. Amyloidosis vs. LVH Classification Obtained with the Held-Out Test Set According to the Input Shape

Table 2 lists the results obtained with the various input shapes illustrated in Figure 1. Patient-based results were always better than frame-based results.

Table 2. Accuracy and AUC of the ROC curve for classification of amyloidosis vs. LVH in the 40% held-out test group, according to the input shape.

	Frame-Based		Patient-Based	
Input Shape	**Accuracy**	**ROC AUC**	**Accuracy**	**ROC AUC**
160 × 160/D + S	0.759	0.836 [0.786–0.878]	0.812	0.937 [0.828–0.987]
160 × 160/D	0.760 (ns)	0.820 (ns) [0.769–0.864]	0.833 (ns)	0.918 (ns) [0.802–0.978]
160 × 160/S	0.733 (ns)	0.801 (0.04) [0.749–0.848]	0.833 (ns)	0.890 (ns) [0.767–0.962]
256 × 256/D + S	0.710 (ns)	0.790 (0.03) [0.735–0.836]	0.771 (ns)	0.803 (0.02) [0.663–0.904]
224 × 224/D + S	0.728 (ns)	0.823 (ns) [0.772–0.867]	0.812 (ns)	0.852 (ns) [0.720–0.938]
128 × 128/D + S	0.740 (ns)	0.808 (ns) [0.756–0.853]	0.812 (ns)	0.922 (ns) [0.807–0.979]
Epicardial ROI	0.722 (ns)	0.787 (0.01) [0.762–0.810]	0.791 (ns)	0.888 (ns) [0.839–0.927]
Myocard. ROI	0.662 (0.05)	0.719 (0.01) [0.693–0.745]	0.714 (ns)	0.814 (0.03) [0.756–0.863]

Results obtained with the 40% held-out test set after hyperparameters tuning. 160 × 160 indicates the cropping size of input frames. D and S indicate diastole and systole. Between brackets is the confidence interval of AUC. Values between parentheses indicate the level of significance of the difference as compared to the 160 × 160 D + S result (assessed with Chi-square test from the number of observations for accuracy and assessed by Delong test for AUC comparisons).

Optimal performance was obtained with 160 × 160 cropped diastolic and systolic images in which per frame analysis provided a test accuracy of 0.759 and an AUC of 0.836, whereas per patient analysis provided a test accuracy of 0.812 and an AUC of 0.937.

Combining diastole and systole did not improve the results. Full field 256 × 256 frames and focused myocardial ROI images provided significantly weaker results.

3.2. Amyloidosis vs. LVH Classification Obtained with the Held-Out Test Set by Human Readers and by CNN

The comparison between classification by experienced radiologists/cardiologists and the CNN is given in Table 3. The CNN provided a largely superior performance when compared to human readers.

Frame-based comparisons of human vs. CNN classification led to an accuracy of 0.605 vs. 0.746 ($p < 0.0008$) and an AUC of 0.630 vs. 0.824 ($p < 0.0001$).

Patient-based comparisons provided an accuracy of 0.660 vs. 0.825 (0.008) and an AUC of 0.727 vs. 0.895 ($p < 0.002$). The ROC curves of these comparisons are plotted in Figure 3.

Table 3. Accuracy and AUC of the ROC curve for classification of amyloidosis vs. LVH in the held-out test group for human readers vs. CNN.

	Frame-Based			Patient-Based		
Metric	**Accur.**	**Sensitiv. Specific.**	**ROC AUC**	**Accur.**	**Sensitiv. Specific.**	**ROC AUC**
CNN	0.746	77.0 71.0	0.824 [0.770–0.869]	0.825	85.7 77.6	0.895 [0.816–0.948]
Read 1	0.585 (0.001)	66.4 50.85	0.570 [0.506–0.632] (0.0001)	0.629 (0.004)	67.4 58.8	0.654 [0.550–0.747] (0.001)
Read 2	0.623 (0.005)	69.6 54.5	0.623 [0.560–0.684] (0.0001)	0.649 (0.009)	69.6 60.8	0.712 [0.611–0.799] (0.0002)
Read 3	0.585 (0.001)	66.4 50.9	0.587 [0.523–0.649] (0.0001)	0.660 (0.013)	71.1 61.5	0.731 [0.631–0.816] (0.002)
Read (avg)	0.605 (0.0008)	69.2 52.7	0.630 [0.567–0.690] (0.0001)	0.660 (0.008)	72.1 61.1	0.727 [0.627–0.813] (0.002)

Frame-based and patient-based results obtained with the held-out test set by human readers and by CNN. Accur. is for accuracy, Sensitiv. and Specific. are for sensitivity and specificity. Values between parentheses indicate the level of significance of the difference between human reader and CNN (assessed with Chi-square test from the number of observations for accuracy and assessed by Delong test for AUC comparisons).

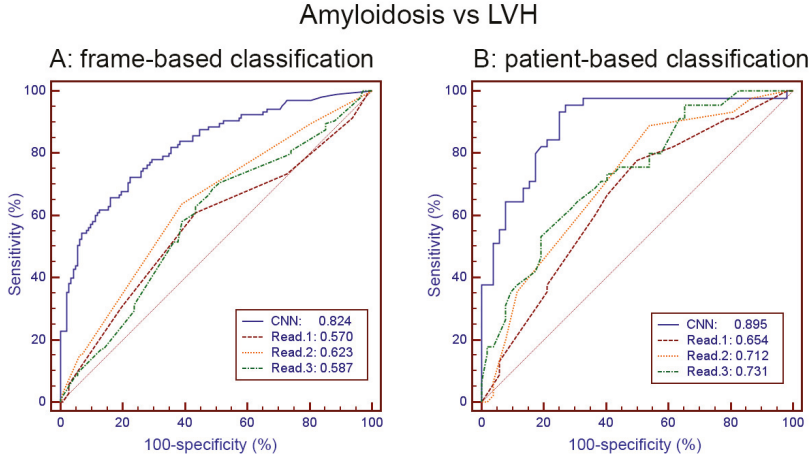


Figure 3. ROC curves and AUC for frame-based (**A**) and patient-based (**B**) classification of amyloidosis vs. LVH by CNN and by three human readers (Read. 1 to 3).

3.3. CNN Classification of AL vs. TTR Amyloidosis

The frame-based accuracy and AUC obtained by the CNN classification of AL vs. TTR cardiac amyloidosis were 0.662 and 0.703 [0.664–0.741]. The corresponding patient-based values were 0.711 and 0.752 [0.654–0.834]. No comparison was performed here with human classification, but the comparison between the AUC values of the CNN and the simple left ventricular septal wall thickness measurement (per-patient AUC 0.735) did not show a statistically significant difference.

3.4. Analysis of the Saliency Maps

Saliency maps, which reveal the pixel areas responsible for classification, show that cardiac regions contribute to CNN decisions in only 25% of cases (Figure 4). Among the extracardiac targeted regions, the lungs are the most frequent, followed by the subcutaneous fat and liver. Distribution is quite similar for correct classification (concordant) and erroneous classification (discordant).

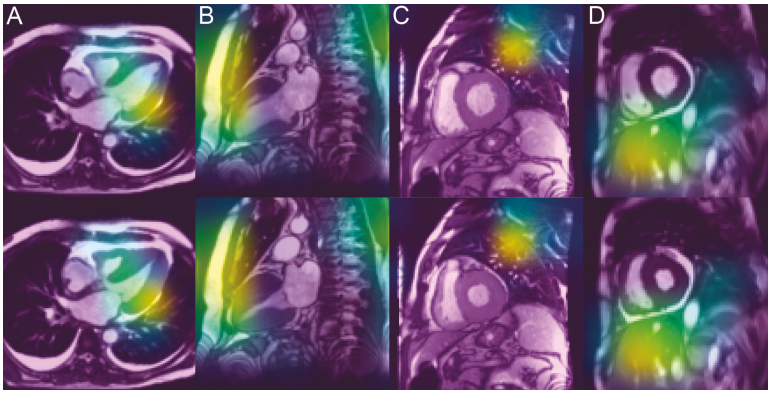


Figure 4. Saliency maps targeting cardiac region (**A**) but also frequently subcutaneous fat (**B**), lung (**C**) or liver (**D**). Diastolic frames are shown in the upper row and systolic frames in the lower row.

4. Discussion

The most important result of this study is the possibility of discriminating cardiac amyloidosis and LVH from other origins by simple cine-CMR images, which was significantly better with the CNN than by the physicians' visual analysis. The comparison carried out on slightly more than 100 patients of the two groups shows that, for frame-based and patient-based analysis, binary classification accuracy is approximately 15 absolute points higher with the CNN than with experienced radiologists/cardiologists. The same significant difference is also found by considering the AUC of the ROC curve, with a little less than 20 points absolute value improvement with the CNN as compared with experienced human readers.

4.1. Methodological Considerations

Patient-based analysis constitutes a much more relevant assessment because this is how the clinical diagnosis is carried out, and it should be noted that the transposition of the results from the image-level to the patient-level (by taking the average of the elementary predictions per frame) leads to an improvement of the accuracy in the range of 5 points (absolute value) and in the range of 10 points for the AUC. This phenomenon, which is observed for the human reader and CNN, may be explained thanks to the "averaging process" in the mind of the physician who examines the whole set of pictures of the patient.

The influence of methodological choices should be stressed: (1) The distribution of patients' images in a distinct train or validation/test data sets is mandatory; otherwise, the results would be clearly biased because we would have trained on images that are—for some features—similar to test images. Processing this way (without frame distribution for a unique patient) with our data set led to a misleading "improvement" of almost 10 points (absolute value) for accuracy and AUC results (data not listed here). (2) The strict separation of the train and validation set for hyperparameter tuning and the test set has been done. This method, based on a separate test set, schematized in Figure 2, is required to avoid information leakage related to hyperparameter tuning.

4.2. Superiority of CNN Capacities over Human Diagnosis

The aim of this study was not to propose making the diagnosis of cardiac amyloidosis solely on the cine-CMR data because much more relevant CMR indices are available thanks to gadolinium injection. Actually, late-enhancement and ECV allow the diagnosis of the presence of CA with a high sensitivity of 95% and an even higher specificity of 98% [5], and deep learning was demonstrated to be efficient in this field [11]. Our goal was to show that deep learning is able to extract diagnostic clues clearly surpassing visual analysis (15 to 20 points in the present study).

Excellent performances of the CNN are often reported in the literature, but their interest is limited if they are not compared to human performance. Among human–machine comparisons, many studies have reported that CNN diagnosis is on par with human visual assessment in multiple areas [17]. For example, for malignancy risk estimation of pulmonary nodules using thoracic CT, Venkadesh et al. [18] reported that the DL algorithm had an AUC of 0.96, which was significantly better than the average AUC of the clinicians (0.90) but comparable to that of thoracic radiologists. Our model was able to discriminate between AL and TTR CA with interesting values of patient-based accuracy (0.711) and AUC (0.752); however, this was no better than the classification obtained with the simple measurement of the septal thickness, already reported in previous publications [19–21] and resulting from the known increased amyloid burden in this subtype.

Of more interest is to show significant machine-over-human superiority in routine areas, where "clinical" visual analysis is the classic benchmark. Our study provides an interesting demonstration in this direction for diagnosing cardiac amyloidosis from cine-CMR. A small number of other publications could demonstrate that AI systems are capable of surpassing human experts in disease prediction. Such is the case for the distinction between low-grade and high-grade glioma by radiologists, which lacks accuracy (40–45% of non-enhancing MR lesions are found subsequently to be malignant glioma), whereas, in contrast, CNN-based grading provides > 90% accuracy [22]. Resnet-50 CNN outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task [23]. For the diagnosis of breast cancer, in a large multicenter study, Mc Kinney et al. [12] found that the AI system exhibited specificity and sensitivity superior to that of radiologists practicing in an academic medical center and exceeded the average performance of radiologists by a significant improvement in the area under the ROC curve (ΔAUC = +0.115). Similarly, in differentiating benign from malignant renal tumors, Xu et al. [24] reported higher AUC with the CNN model (0.906, based on T2-weighted images) as compared to the AUC obtained by two radiologists (0.724).

4.3. Unveiling the Invisible

One more step in this diagnostic quest is the possibility of discriminating pathological conditions that clinicians are not able to predict at all using the naked eye. Subtyping molecular markers, histological or immune-histochemical and genetic classes is impossible to ascertain from radiologic data. These identifications were initially proposed from radiomic signatures, for instance, to discriminate between hypertensive heart disease and hypertrophic cardiomyopathy [25] or between recent infarction vs. old infarction [26]. However, several comparative studies have demonstrated that deep learning based on radiologic data is superior to radiomics. This has been demonstrated for renal cancer [24], subtyping different types of cerebral glioma [27], diagnosis of breast cancer [28] and predicting axillary lymph node metastasis of breast cancer [29].

This may be explained because radiomics' features are handcrafted in advance and, thus, may not always fit to discriminate particular tasks. In contrast, the CNN is more flexible, adaptive and dynamic. As a data-driven tool, it is able to automatically learn to extract and select task-specific features if the amount of training data is large enough. Further evidence for the power of deep learning to make a histological diagnosis from radiological data has been provided by Zhao et al. for renal cell carcinomas Fuhrman-grading [30] and by Yuan et al. for prostate cancer Gleason score staging (accuracy 0.87) [31].

4.4. Explanation of Classification Remains Unsatisfactory

Deep neural networks operate through a multilayer nonlinear structure, making their predictions difficult to interpret. They are able to pick up a number of features that cannot be interpreted by humans but which are relevant for making a diagnosis. These automatically-learned discriminative features are unfortunately presently not clearly identifiable.

Grad-CAM helps identify the areas of pixels that are most responsible for class prediction [16]. This should provide valuable clues to understand the algorithm's decision. In principle, the salient areas should be located in the cardiac region, which only appeared in a quarter of the cases in our study. Two explanations may be advanced for this anomaly.

(1) Technically, our network uses only fully connected layers in the last phase, which is where the classification happens, but saliency cannot be obtained from fully connected layers. As a solution, we should try replacing some of the fully connected layers that come right after VGG with convolution. This way, the spatial information would be preserved longer in the network, and we might see more meaning in the saliency maps.

(2) Amyloidosis is not a disease confined to the heart since the involvement of the lungs, fatty tissues and other organs is also common. Liver and, moreover, spleen amyloid deposits have been reported in 41% of patients with systemic amyloidosis (almost only in AL type), and CMR-derived ECV measurement showed good diagnostic capability in this field [32]. This is why the diagnosis is also based on extracardiac biopsies, and it is interesting to note that the texture analysis was able to show specificities in the architecture of ultrasound images within abdominal fat [33], resulting in increased echogenicity and a loss of the normal structure of the fat layer, consistent with histopathological amyloid deposition in the fat.

This ubiquitous aspect of the disease may also explain why the input shape submitted to the CNN (from the full field image to the small region of interest focused on the sole myocardium illustrated in Figure 1) hardly modifies the performance of our model as shown in Table 2. It can also be noted in Table 2 that the combination of diastole and systole does not provide any diagnostic benefit, unlike for other cardiomyopathies [10], because the global LV systolic function is generally preserved in the early stage of amyloidosis.

4.5. Study Limitations

Two types of confounding factors must be mentioned. First, plane orientation and the presence of gadolinium in the sets of images could have influenced the results, but Table 1 shows a perfect equivalence between the two groups. Second, the presence of pericardial or pleural effusion constitutes a more important bias because the prevalence (slightly higher than in the study of Binder et al. [9]) is very different in the two groups. Pericardial effusion is observed in almost 50% of CA, i.e., two times more often than in hypertrophies unrelated to amyloidosis. Pleural effusions are observed in just over a third of CA, i.e., four times more than in other hypertrophies, and mixed effusions are 10 times more frequent in the amyloidosis group than in the LVH group. This disparity probably contributes to the classification made by CNNs (although heat maps rarely focus on areas of effusion) but also influences clinical judgment, so that the bias is the same for the machine and for the human, which, therefore, does not explain the diagnostic superiority of the algorithm.

A multiparametric approach is needed. Only cine-CMR data has been used here, and it is likely that one could significantly improve performances by combining the analysis with other CMR sequences such as T1 mapping, ECV assessment and late gadolinium-enhancement imaging. Based on gadolinium-enhanced images—and not on cine-MR images—Martini et al. obtained an accuracy of 0.88 and AUC of 0.98 [11], but remember that our aim was not to develop the best model to optimize cardiac amyloidosis diagnosis but to compare CNN and human reader performance. For the distinction between AL and TTR CA, it has been reported that transmural patterns of late gadolinium enhancement may differentiate these two types of the disease [21] but with relatively low performance. Recently, the use of a logistic regression model integrating T2 mapping (slightly increased

in the AL subtype) and right ventricular ejection fraction combined with age was reported to discriminate between these two subtypes with an AUC of 0.92 [34]. The performance of AI integrating such multiparametric CMR features, especially for the distinction between AL and TTR cardiac amyloidosis, should be explored in the future.

Technical improvements should be implemented. The leverage of more sophisticated CNN models (not limited to the classical VGG model used here) and, moreover, the combination (concatenation) of several multiparametric inputs, with possible additional categorical clinical input variables (e.g., [30]), should improve performance. Orientation plane specific models [11] should also be tested since images of different views were classified here by the same network, which makes learning relevant features from images potentially much harder as it increases variability unrelated to any disease. Significant work also remains to be done to improve the explainability of the results. Finally, the relatively limited number of observations and the monocentric nature of this study constitute another limitation. Multicenter studies could be of interest for the further validation and generalization of our findings.

5. Conclusions

In this study, based on cine-CMR images alone, we could demonstrate the ability of CNNs to discriminate cardiac amyloidosis from LVH of other origins significantly better than experienced human operators. The diagnostic accuracy and AUC were 15 to 20 points higher (in absolute value) for the VGG convolutional network used here than for human readers. This diagnostic superiority of the CNN results from the unique capability of the algorithm to identify features invisible to the naked eye, indiscernible through the classical radiological analysis. This scientific novelty, already reported in a few recent articles concerning other pathological fields, opens up promising prospects for improving diagnostic capacities in routine clinical practice. The astonishing potential of CNNs to improve the recognition of pathologies that are imperfectly detectable in radiology and to reveal invisible clues such as the histological type of lesions will certainly constitute a large field of future research.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics12010069/s1; Table S1: Clinical and CMR characteristics of AL and TTR cardiac amyloidosis.

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

Funding: This work was partially supported by French state funds managed by the ANR under reference ANR-10-IAHU-02, without any involvement in the study design, data gathering, analysis/interpretation of data or writing of the report.

Institutional Review Board Statement: This retrospective study was registered and approved by the Institutional Review Board of the university hospital of Strasbourg (ref 20–072/sept 2020). All datasets were obtained and de-identified, with waived consent in compliance with the Institutional Review Board of our institution.

Informed Consent Statement: All datasets were obtained and de-identified, with waived consent in compliance with the Institutional Review Board of our institution. No protected health information for any subject is given in this manuscript.

Data Availability Statement: The database and code can be made available by reasonable request after the agreement of the Clinical Research Department of our hospital.

Conflicts of Interest: Nicolas Padoy serves as a consultant for Caresyntax and has received research support from Intuitive Surgical, unrelated to this work. Soraya El Ghannudi serves as a consultant for Pfizer.

References

1. Garcia-Pavia, P.; Rapezzi, C.; Adler, Y.; Arad, M.; Basso, C.; Brucato, A.; Burazor, I.; Caforio, A.L.P.; Damy, T.; Eriksson, U.; et al. Diagnosis and treatment of cardiac amyloidosis. A position statement of the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases. *Eur. J. Heart Fail.* **2021**, *23*, 512–526. [CrossRef] [PubMed]
2. Kittleson, M.M.; Maurer, M.S.; Ambardekar, A.V.; Bullock-Palmer, R.P.; Chang, P.P.; Eisen, H.J.; Nair, A.P.; Nativi-Nicolau, J.; Ruberg, F.L. American Heart Association Heart Failure and Transplantation Committee of the Council on Clinical Cardiology. Cardiac Amyloidosis: Evolving Diagnosis and Management: A Scientific Statement From the American Heart Association. *Circulation* **2020**, *142*, 7–22. [CrossRef] [PubMed]
3. Papathanasiou, M.; Carpinteiro, A.; Rischpler, C.; Hagenacker, T.; Rassaf, T.; Luedike, P. Diagnosing cardiac amyloidosis in every-day practice: A practical guide for the cardiologist. *Int. J. Cardiol. Heart Vasc.* **2020**, *28*, 100519. [CrossRef]
4. Maurer, M.S.; Bokhari, S.; Damy, T.; Dorbala, S.; Drachman, B.M.; Fontana, M.; Grogan, M.; Kristen, A.V.; Lousada, I.; Nativi-Nicolau, J.; et al. Expert Consensus Recommendations for the Suspicion and Diagnosis of Transthyretin Cardiac Amyloidosis. *Circ. Heart Fail.* **2019**, *12*, e006075. [CrossRef] [PubMed]
5. Chatzantonis, G.; Bietenbeck, M.; Elsanhoury, A.; Tschöpe, C.; Pieske, B.; Tauscher, G.; Vietheer, J.; Shomanova, Z.; Mahrholdt, H.; Rolf, A.; et al. Diagnostic value of cardiovascular magnetic resonance in comparison to endomyocardial biopsy in cardiac amyloidosis: A multi-centre study. *Clin. Res. Cardiol.* **2021**, *110*, 555–568. [CrossRef] [PubMed]
6. Wang, T.K.M.; Brizneda, M.V.; Kwon, D.H.; Popovic, Z.B.; Flamm, S.D.; Hanna, M.; Griffin, B.P.; Xu, B. Reference Ranges, Diagnostic and Prognostic Utility of Native T1 Mapping and Extracellular Volume for Cardiac Amyloidosis: A Meta-Analysis. *J. Magn. Reason. Imaging* **2021**, *53*, 1458–1468. [CrossRef]
7. Weinreb, J.C.; Rodby, R.A.; Yee, J.; Wang, C.L.; Fine, D.; McDonald, R.J.; Perazella, M.A.; Dillman, J.R.; Davenport, M.S. Use of Intravenous Gadolinium-based Contrast Media in Patients with Kidney Disease: Consensus Statements from the American College of Radiology and the National Kidney Foundation. *Kidney Med.* **2021**, *3*, 142–150. [CrossRef] [PubMed]
8. Martinez-Naharro, A.; Treibel, T.A.; Abdel-Gadir, A.; Bulluck, H.; Zumbo, G.; Knight, D.S.; Kotecha, T.; Francis, R.; Hutt, D.F.; Rezk, T.; et al. Magnetic Resonance in Transthyretin Cardiac Amyloidosis. *J. Am. Coll. Cardiol.* **2017**, *70*, 466–477. [CrossRef] [PubMed]
9. Binder, C.; Duca, F.; Binder, T.; Rettl, R.; Dachs, T.M.; Seirer, B.; Camuz Ligios, L.; Dusik, F.; Capelle, C.; Qin, H.; et al. Prognostic implications of pericardial and pleural effusion in patients with cardiac amyloidosis. *Clin. Res. Cardiol.* **2021**, *110*, 532–543. [CrossRef] [PubMed]
10. Germain, P.; Vardazaryan, A.; Padoy, N.; Labani, A.; Roy, C.; Schindler, T.H.; El Ghannudi, S. Classification of Cardiomyopathies from MR Cine Images Using Convolutional Neural Network with Transfer Learning. *Diagnostics* **2021**, *11*, 1554. [CrossRef] [PubMed]
11. Martini, N.; Aimo, A.; Barison, A.; Della Latta, D.; Vergaro, G.; Aquaro, G.D.; Ripoli, A.; Emdin, M.; Chiappino, D. Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance. *J. Cardiovasc. Magn. Reson.* **2020**, *22*, 84. [CrossRef] [PubMed]
12. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A.; et al. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [CrossRef]
13. Zhou, H.; Li, L.; Liu, Z.; Zhao, K.; Chen, X.; Lu, M.; Yin, G.; Song, L.; Zhao, S.; Zheng, H.; et al. Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images. *Eur. Radiol.* **2021**, *31*, 3931–3940. [CrossRef] [PubMed]
14. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef] [PubMed]
15. Simonyan, K.; Zisserman, A. Very Deep Convolutional Network for Large Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
16. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
17. Shen, J.; Zhang, C.J.P.; Jiang, B.; Chen, J.; Song, J.; Liu, Z.; He, Z.; Wong, S.Y.; Fang, P.H.; Ming, W.K. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Med. Inform.* **2019**, *7*, e10010. [CrossRef]
18. Venkadesh, K.V.; Setio, A.A.A.; Schreuder, A.; Scholten, E.T.; Chung, K.; Wille, M.M.W.; Saghir, Z.; van Ginneken, B.; Prokop, M.; Jacobs, C. Deep Learning for Malignancy Risk Estimation of Pulmonary Nodules Detected at Low-Dose Screening CT. *Radiology* **2021**, *300*, 438–447. [CrossRef] [PubMed]
19. Itzhaki Ben Zadok, O.; Vaturi, M.; Vaxman, I.; Iakobishvili, Z.; Rhurman-Shahar, N.; Kornowski, R.; Hamdan, A. Differences in the characteristics and contemporary cardiac outcomes of patients with light-chain versus transthyretin cardiac amyloidosis. *PLoS ONE* **2021**, *16*, e0255487. [CrossRef] [PubMed]
20. Quarta, C.C.; Solomon, S.D.; Uraizee, I.; Kruger, J.; Longhi, S.; Ferlito, M.; Gagliardi, C.; Milandri, A.; Rapezzi, C.; Falk, R.H. Left ventricular structure and function in transthyretin-related versus light-chain cardiac amyloidosis. *Circulation* **2014**, *129*, 1840–1849. [CrossRef]
21. Dungu, J.N.; Valencia, O.; Pinney, J.H.; Gibbs, S.D.; Rowczenio, D.; Gilbertson, J.A.; Lachmann, H.J.; Wechalekar, A.; Gillmore, J.D.; Whelan, C.J.; et al. CMR-based differentiation of AL and ATTR cardiac amyloidosis. *JACC Cardiovasc. Imaging* **2014**, *7*, 133–142. [CrossRef] [PubMed]

22. Zlochower, A.; Chow, D.S.; Chang, P.; Khatri, D.; Boockvar, J.A.; Filippi, C.G. Deep Learning AI Applications in the Imaging of Glioma. *Top. Magn. Reason. Imaging* **2020**, *29*, 115–121. [CrossRef]
23. Brinker, T.J.; Hekler, A.; Enk, A.H.; Klode, J.; Hauschild, A.; Berking, C.; Schilling, B.; Haferkamp, S.; Schadendorf, D.; Holland-Letz, T.; et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer* **2019**, *113*, 47–54. [CrossRef] [PubMed]
24. Xu, Q.; Zhu, Q.; Liu, H.; Chang, L.; Duan, S.; Dou, W.; Li, S.; Ye, J. Differentiating Benign from Malignant Renal Tumors Using T2- and Diffusion-Weighted Images: A Comparison of Deep Learning and Radiomics Models Versus Assessment from Radiologists. *J. Magn. Reason. Imaging* **2021**. [CrossRef]
25. Neisius, U.; El-Rewaidy, H.; Nakamori, S.; Rodriguez, J.; Manning, W.J.; Nezafat, R. Radiomic Analysis of Myocardial Native T1 Imaging Discriminates Between Hypertensive Heart Disease and Hypertrophic Cardiomyopathy. *JACC Cardiovasc. Imaging* **2019**, *12*, 1946–1954. [CrossRef] [PubMed]
26. Larroza, A.; Materka, A.; López-Lereu, M.P.; Monmeneu, J.V.; Bodí, V.; Moratal, D. Differentiation between acute and chronic myocardial infarction by means of texture analysis of late gadolinium enhancement and cine cardiac magnetic resonance imaging. *Eur. J. Radiol.* **2017**, *92*, 78–83. [CrossRef] [PubMed]
27. Li, Y.; Wei, D.; Liu, X.; Fan, X.; Wang, K.; Li, S.; Zhang, Z.; Ma, K.; Qian, T.; Jiang, T.; et al. Molecular subtyping of diffuse gliomas using magnetic resonance imaging: Comparison and correlation between radiomics and deep learning. *Eur. Radiol.* **2021**, *52*. [CrossRef] [PubMed]
28. Truhn, D.; Schrading, S.; Haarburger, C.; Schneider, H.; Merhof, D.; Kuhl, C. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology* **2019**, *290*, 290–297. [CrossRef]
29. Sun, Q.; Lin, X.; Zhao, Y.; Li, L.; Yan, K.; Liang, D.; Sun, D.; Li, Z.C. Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region. *Front. Oncol.* **2020**, *10*, 53. [CrossRef] [PubMed]
30. Zhao, Y.; Chang, M.; Wang, R.; Xi, I.L.; Chang, K.; Huang, R.Y.; Vallières, M.; Habibollahi, P.; Dagli, M.S.; Palmer, M.; et al. Deep Learning Based on MRI for Differentiation of Low- and High-Grade in Low-Stage Renal Cell Carcinoma. *J. Magn. Reason. Imaging* **2020**, *52*, 1542–1549. [CrossRef]
31. Yuan, Y.; Qin, W.; Buyyounouski, M.; Ibragimov, B.; Hancock, S.; Han, B.; Xing, L. Prostate cancer classification with multiparametric MRI transfer learning model. *Med. Phys.* **2019**, *46*, 756–765. [CrossRef]
32. Chacko, L.; Boldrini, M.; Martone, R.; Law, S.; Martinez-Naharrro, A.; Hutt, D.F.; Kotecha, T.; Patel, R.K.; Razvi, Y.; Rezk, T.; et al. Cardiac Magnetic Resonance-Derived Extracellular Volume Mapping for the Quantification of Hepatic and Splenic Amyloid. *Circ. Cardiovasc. Imaging* **2021**, *14*, e012506. [CrossRef] [PubMed]
33. Misumi, Y.; Ueda, M.; Yamashita, T.; Masuda, T.; Kinoshita, Y.; Tasaki, M.; Nagase, T.; Ando, Y. Novel screening for transthyretin amyloidosis by using fat ultrasonography. *Ann. Neurol.* **2017**, *81*, 604–608. [CrossRef] [PubMed]
34. Slivnick, J.A.; Tong, M.S.; Nagaraja, H.N.; Elamin, M.B.; Wallner, A.; O'Brien, A.; Raman, S.V.; Zareba, K.M. Novel Cardiovascular Magnetic Resonance Model to Distinguish Immunoglobulin Light Chain From Transthyretin Cardiac Amyloidosis. *JACC Cardiovasc. Imaging* **2021**, *14*, 302–304. [CrossRef]

diagnostics

MDPI

Article

VGG19 Network Assisted Joint Segmentation and Classification of Lung Nodules in CT Images

Muhammad Attique Khan [1], Venkatesan Rajinikanth [2], Suresh Chandra Satapathy [3], David Taniar [4], Jnyana Ranjan Mohanty [5], Usman Tariq [6] and Robertas Damaševičius [7,*]

1 Department of Computer Science, HITEC University, Taxila 47080, Pakistan; attique@ciitwah.edu.pk
2 Department of Electronics and Instrumentation Engineering, St. Joseph's College of Engineering, Chennai, Tamilnadu 600119, India; v.rajinikanth@ieee.org
3 School of Computer Engineering, Kalinga Institute of Industrial Technology (Deemed to Be University), Bhubaneswar, Odisha 751024, India; suresh.satapathyfcs@kiit.ac.in
4 Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia; David.Taniar@monash.edu
5 School of Computer Applications, Kalinga Institute of Industrial Technology (Deemed to Be University), Bhubaneswar, Odisha 751024, India; jmohantyfca@kiit.ac.in
6 College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia; u.tariq@psau.edu.sa
7 Faculty of Applied Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland
* Correspondence: robertas.damasevicius@polsl.pl

Abstract: Pulmonary nodule is one of the lung diseases and its early diagnosis and treatment are essential to cure the patient. This paper introduces a deep learning framework to support the automated detection of lung nodules in computed tomography (CT) images. The proposed framework employs VGG-SegNet supported nodule mining and pre-trained DL-based classification to support automated lung nodule detection. The classification of lung CT images is implemented using the attained deep features, and then these features are serially concatenated with the handcrafted features, such as the Grey Level Co-Occurrence Matrix (GLCM), Local-Binary-Pattern (LBP) and Pyramid Histogram of Oriented Gradients (PHOG) to enhance the disease detection accuracy. The images used for experiments are collected from the LIDC-IDRI and Lung-PET-CT-Dx datasets. The experimental results attained show that the VGG19 architecture with concatenated deep and handcrafted features can achieve an accuracy of 97.83% with the SVM-RBF classifier.

Keywords: lung CT images; nodule detection; VGG-SegNet; pre-trained VGG19; deep learning

Citation: Khan, M.A.; Rajinikanth, V.; Satapathy, S.C.; Taniar, D.; Mohanty, J.R.; Tariq, U.; Damaševičius, R. VGG19 Network Assisted Joint Segmentation and Classification of Lung Nodules in CT Images. *Diagnostics* **2021**, *11*, 2208. https://doi.org/10.3390/diagnostics11122208

Academic Editor: Sameer Ántani

Received: 28 October 2021
Accepted: 24 November 2021
Published: 26 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Lung cancer/nodule is one of the severe abnormalities in the lung, and a World Health Organization (WHO) report indicated that around 1.76 million deaths have occurred globally in 2018 due to lung cancer [1]. Lung cancer/nodule is due to abnormal cell growth in the lung and, in most cases, the nodule may be cancerous/non-cancerous. The Olson report [2] confirmed that lung nodules can be categorized into benign/malignant based on their dimension (5 to 30 mm fall into the benign class and >30 mm is malignant). When a lung nodule is diagnosed using the radiological approach, a continuous follow-up is recommended to check its growth rate. The follow-up procedure can continue for up to two years and, along with non-invasive radiographic imaging procedures, other invasive methodologies, such as bronchoscopy and/or tissue biopsy, can also be suggested to confirm the condition and harshness of the lung nodules in a patient [3].

Noninvasive radiological techniques are commonly adopted in initial level lung nodule detection using CT images, and, therefore, several lung nodule detection works are already proposed in the literature [4–6] which involve the use of traditional signal processing and texture analysis techniques combined with machine learning classification [7],

deep learning models [8,9], neural networks combined with nature-inspired optimization techniques [10,11] and ensemble learning [12]. The aims of this research are to construct a Deep Learning (DL) supported scheme to segment the lung nodule segment from the CT image slice with better accuracy and classify the considered CT scan images into normal/nodule class with improved accuracy using precisely selected deep and handcrafted features.

The recent article by Rajinikanth and Kadry [13] proposed a framework with VGG16 neural network model for the automated segmentation and classification of lung nodules from CT images. In their paper, a threshold filter technique is implemented to remove artifacts from CT images, and the artifact-eliminated images are then considered to test the proposed disease detection framework. The proposed scheme is tested using the LIDC-IDRI database [14–16] and the classification task implemented with the combined deep and handcrafted features helped to achieve a classification accuracy of 97.67% with a Random Forest (RF) classifier.

In this paper, we suggest a framework to support automated segmentation and classification of lung nodules with improved accuracy. The proposed scheme includes the following stages: (i) image collection and resizing, (ii) implementing the pre-trained VGG supported segmentation; (iii) deep feature-based classification, (iv) extracting the essential handcrafted features such as Gray Level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP) and Pyramid Histogram of Oriented Gradients (PHOG), (v) implementing a serial feature concatenation to unite the deep and handcrafted features and (vi) implementing and validating the performance of the classifiers using a 10-fold cross validation.

The images used for the experiments are collected from the LIDC-IDRI [15] and Lung-PET-CT-Dx [17] datasets. All these works are realized using the MATLAB® (MathWorks, Inc., Natick, MA, USA), and the attained result is then compared and validated with the earlier results presented in the literature.

The major contribution of the proposed work is as follows:

i. Implementation of VGG19 to construct the VGG-SegNet scheme to extract lung nodule.
ii. Deep learning feature extraction based on VGG19.
iii. Combining handcrafted features and deep features to improving nodule detection accuracy.

The proposed work is organized as follows. Section 2 presents and discusses earlier related research. Section 3 presents the implemented methodology. Section 4 shows the experimental results and discussions and, finally, the conclusions of the present research study are given in Section 5.

2. Related Work

Due to its impact, a significant amount of lung nodule detection from CT images is proposed using a variety of image databases, and summarizing the presented schemes will help to obtain an idea of the advantages and limitations of the existing lung nodule detection procedures. Traditional methods of machine learning (ML) and deep learning (DL) were proposed to examine lung nodules using CT image slices, and the summary of the selected DL-based lung nodule detection systems is presented in Table 1; all the considered works in this table discuss the lung nodule detection technique using a chosen methodology. Furthermore, all these works considered the LIDC-IDRI database for examination.

Table 1. Summary of existing lung nodule detection system with LIDC-IDRI database.

Reference	Lung Nodule Detection Technique	Accuracy (%)	Sensitivity (%)	Specificity (%)
Bhandary et al. [4]	A modified AlexNet with the Support Vector Machine (SVM) based binary classification helped to achieve improved result.	97.27	97.80	98.09
Choi and Choi [5]	An automated Computer-Aided-Detection scheme is proposed to examine the lung nodules using CT images.	97.60	95.20	96.20
Tran et al. [6]	A novel 15-layer DL architecture is implemented by considering the cross entropy/focal as the loss functions.	97.20	96.00	97.30
Rajinikanth and Kadry [13]	Implemented VGG16 DL scheme to segment and classify the lung nodules using deep and handcrafted features.	97.67	96.67	98.67
Kuruvilla and Gunavathi [18]	This research implemented Neural-Network (NN) supported recognition of lung nodules in CT images.	93.30	91.40	100
Nascimento et al. [19]	This work implemented a lung nodule classification based on Shannon and Simpson-Diversity Indices and SVM classifier.	92.78	85.64	97.89
Khehrah et al. [20]	Improved lung nodule detection is achieved with the help of statistical and shape features.	92.00	93.75	91.18
Wang et al. [21]	Deep NN (DNN) and 6G communication network supported lung nodule detection is proposed and implemented in this work using the CT images.	91.70	92.23	91.17
Li et al. [22]	This work implements a Convolutional-Neural-Network (CNN) supported lung nodule detection using the lung CT images.	86.40	87.10	n/a
Kaya and Can [23]	The lung nodule classification is implemented in this work and the ensemble random-forest classifier provided enhanced classification result.	84.89	83.11	92.09
Song et al. [24]	This work implemented a DNN scheme to classify the cropped lung nodule sections from the CT image slices.	82.37	80.66	83.90

The summary (see Table 1) presents a few similar methods implemented using CT images of the LIDC-IDRI database, and the highest categorization accuracy achieved is 97.67% [13].

In addition, a detailed evaluation of various lung nodule recognition practices existing in the literature is available in the following references [25–27]. Some of the works discussed in Table 1 recommended the need for a competent lung nodule detection system that can support both segmentation of the nodule section and classification of lung nodules from normal (healthy) CT images. The works discussed in Table 1 implemented either a segmentation or classification technique using deep features only. Obtaining better detection accuracy is difficult with existing techniques and, hence, the combination of deep features (extracted by a trained neural network model) and handcrafted features is necessary.

In this paper, the pre-trained VGG-16 supported segmentation (VGG-SegNet) is initially executed to extract the lung nodule section from CT images, and then the CT image classification is executed using deep features as well as combined deep and handcrafted features. A detailed assessment among various two-class classifiers, such as SoftMax, Decision-Tree (DT), RF, K-Nearest Neighbor (KNN) and SVM-RBF are also presented using a 10-fold cross-validation to validate the proposed scheme.

3. Methodology

In the literature, several lung abnormality detection systems based on DL are proposed and implemented using clinical-level two-dimensional (2D) CT images as well as benchmark images. Figure 1 shows the proposed system to segment and classify the lung nodule section of the CT images. Initially, the CT images are collected from the benchmark data set and, later, the conversion from 3D to 2D is implemented using ITK-Snap [28]. The ITK-Snap converts the 3D images into 2D slices of planes, such as axial, coronal and sagittal and, in this work, only the axial plane is considered for the assessment. Finally, all test images are resized to 224 × 224 × 3 and then used for the segmentation and classification task. The resized 2D CT images are initially considered for the segmentation task; where the lung nodule segment is mined using the VGG-SegNet scheme implemented with the VGG19 architecture. Later, the essential features are extracted with GLCM, LBP and PHOG, and then these features are combined with the learned features of the pre-trained DL scheme. Finally, the serially concatenated deep features (DF) and handcrafted features (HCF) are used to train, test and confirm the classifier. Based on the attained performance values, the performance of the proposed system is validated.

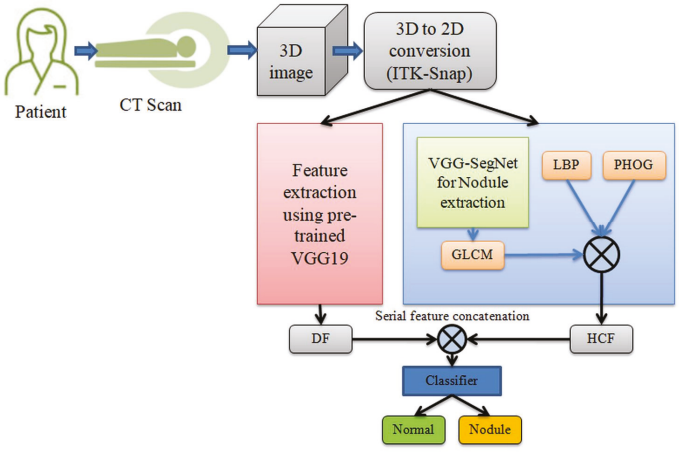


Figure 1. Structure of the proposed lung-nodule segmentation and classification system.

3.1. Image Database Preparation

The CT images are collected from LIDC-IDRI [15] and Lung-PET-CT-Dx [17] databases. These data sets have the clinically collected three-dimensional (3D) lung CT images with the chosen number of slices.

The assessment of the 3D CT images is quite complex and, hence, 3D to 2D conversion is performed to extract the initial image with a dimension of 512 × 512 × 3 pixels, and these images are then resized to 224 × 224 × 3 pixels to decrease the assessment complexity. In this work, only the axial view of 2D slices is used for the estimation and the sample test images of the considered image data set are depicted in Figure 2 and the total images for investigation are given in Table 2.

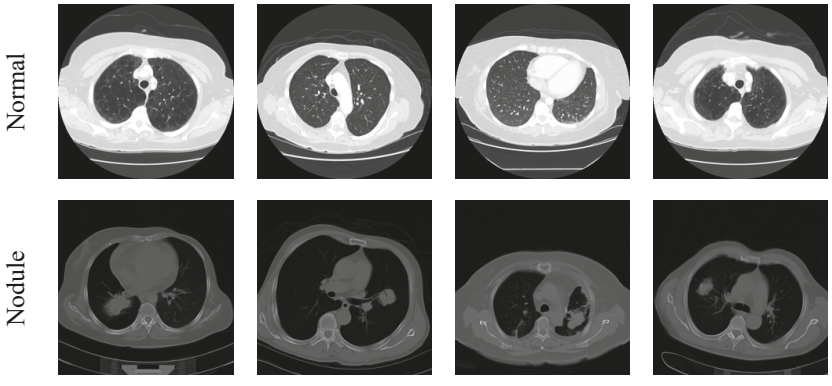


Figure 2. Sample test images considered in this study.

Table 2. The lung CT images analyzed in the experiments.

Image Class	Dimension	Total Images	Training Images	Validation Images
Normal	224 × 224 × 3	1000	750	250
Nodule	224 × 224 × 3	1000	750	250

3.2. Nodule Segmentation

Evaluation of the shape and dimension of the abnormality in medical images is widely preferred during the image-supported disease diagnosis and treatment implementation process [29,30]. Automated segmentation is widely used to extract the infected section from the test image and the mined fragment is further inspected to verify the disease and its severity level. In the assessment of the lung nodule with CT images, the dimension of the lung nodule plays a vital role and, therefore, the extraction of the nodule is very essential. In this work, the VGG-SegNet scheme is implemented with the VGG19 scheme to extract the CT image nodule. Information on the traditional VGG-SegNet model can be found in [29].

The proposed VGG-SegNet model consists of the following specification; traditional VGG19 scheme is considered as the encoder section and its associated structure forms the decoder unit. Figure 3 illustrates the construction of the VGG19-based segmentation and classification scheme in which the traditional VGG19 scheme (first 5 layers) works as the encoder region and the inverted VGG19 with up-sampling facility is then considered as the decoder region. The pre-tuning of this scheme for the CT image is performed using the test images considered for training along with the essential image enhancement process [31]. The preliminary constraints for training the VGG-SegNet are allocated as follows: batch size is equal for encoder-decoder section, initialization uses a normal weight, learning rate

is fixed as 1e-5, Linear Dropout Rate (LDR) is assigned, and Stochastic Gradient-Descent (SGD) optimization is selected. The final SoftMax layer uses a sigmoid activation function.

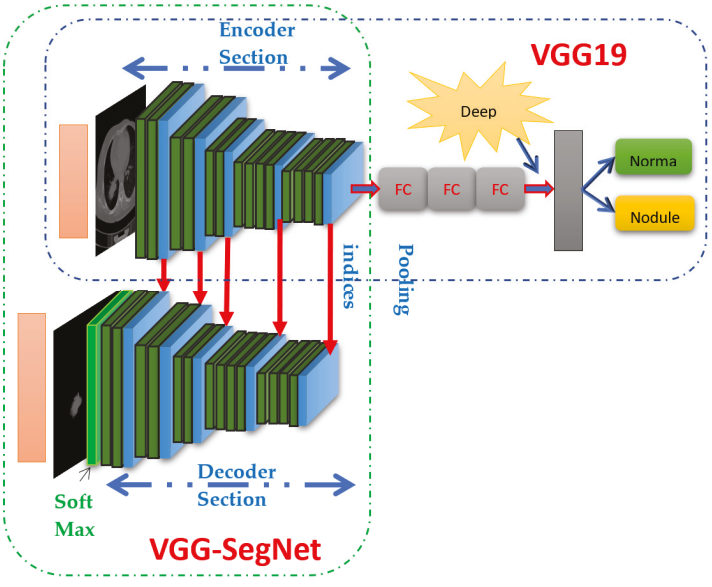


Figure 3. Structure of VGG19 supported segmentation (VGG-SegNet) and classification scheme.

3.3. Nodule Classification

In the medical domain, automated disease classification plays an important role during the mass data assessment and a perfectly tuned disease classification system further reduces the diagnostic burden of physicians and acts as an assisting system during the decision-making process [32–35]. Therefore, a considerable number of disease detection systems assisted by DL are proposed and implemented in the literature [36–40]. Recent DL schemes implemented in the LIDC-IDRI with fused deep and HCF helped achieve a classification accuracy of >97% [13].

Figure 3 presents the assisted classification of using the VGG19 of lung CT images (dimension 224 × 224 × 3 pixels) using the DF using the SoftMax classifier, and then the performance of VGG19 is validated with VGG16, ResNet18, ResNet50 and AlexNet (images with dimension of 227 × 227 × 3 pixels) [41–46] and the performance is compared and validated. The performance of the implemented VGG19 is validated using DF, concatenated DF + HCF and well-established binary classifiers existing in the literature [47–50].

3.3.1. Deep Features

Initially, the proposed scheme is implemented by considering the DF attained at fully connected layer 3 (FC3). After possible dropout, FC3 helps to provide a feature vector of dimension 1 × 1024, whose value is mathematically represented as in Equation (1).

$$FV_{VGG19\ (1\times1024)} = VGG19_{(1,1)}, VGG19_{(1,2)}, \ldots, VGGgg19_{(1,1024)} \quad (1)$$

Other essential information on VGG19 and the related issues can be found in [41].

3.3.2. Handcrafted Features

The features extracted from the test image using a chosen image processing methodology are known as Machine Learning Features (MLF) or handcrafted features (HCF).

Previous research in the literature already confirmed the need for the precision of HCF to progress the categorization accuracy in a class of ML and DL-based disease detection systems [46,50,51]. In the proposed work, the essential HCF from the considered test images is extracted using well-known methods such as GLCM [13,36,42], LBP [13,46] and PHOG [48].

The GLCM features are commonly used due to their high performance and, in this paper, the GLCM features are extorted from the lung nodule section segmented with the VGG-SegNet. The entire feature used in this work can be found in Equation (2).

$$FV1_{\text{GLCM }(1\times25)} = GLCM_{(1,1)}, GLCM_{(1,2)}, \ldots, GLCM_{(1,25)} \quad (2)$$

In this work, the LBP with varied weight (weights with values; W = 1, 2, 3, and 4) is considered to mine the important features from the considered test images and the proposed LBP is already implemented in the works of Gudigar et al. [52] and Rajinikanth and Kadry [13]. The LBP features for the varied weights are depicted in Equations (3)–(6) and Equation (7) depicts the overall LBP features.

$$FV_{LBP1\ (1\times59)} = LBP1_{(1,1)}, LBP1_{(1,2)}, \ldots, LBP1_{(1,59)} \quad (3)$$

$$FV_{LBP2\ (1\times59)} = LBP2_{(1,1)}, LBP2_{(1,2)}, \ldots, LBP2_{(1,59)} \quad (4)$$

$$FV_{LBP3\ (1\times59)} = LBP3_{(1,1)}, LBP3_{(1,2)}, \ldots, LBP3_{(1,59)} \quad (5)$$

$$FV_{LBP4\ (1\times59)} = LBP4_{(1,1)}, LBP4_{(1,2)}, \ldots, LBP4_{(1,59)} \quad (6)$$

$$FV2_{LBP\ (1\times236)} = FV_{LBP1\ (1\times59)} + FV_{LBP2\ (1\times59)} + FV_{LBP3\ (1\times59)} + FV_{LBP4\ (1\times59)} \quad (7)$$

Along with the above said features, the PHOG features are also extracted and considered along with GLCM and LBP. The total information on the PHOG can be found in the article by Murtza et al. [48]. In this work, 255 features are extracted by assigning number of bins = 3 and levels (L) = 3. The PHOG features of the proposed work are depicted in Equation (8).

$$FV3_{PHOG\ (1\times255)} = PHOG_{(1,1)}, PHOG_{(1,2)}, \ldots, PHOG_{(1,255)}, \quad (8)$$

3.3.3. Features Concatenation

In this work, a serial features concatenation is realized to unite the DF and HCF, and this technique helps to improve the feature dimension to a higher level. The serial features concatenation implemented in this work is depicted in Equation (9) and Final-Feature-Vector (FFV) is presented in Equation (10).

$$Concatenated\ features = DF_{(1\times1024)} + HCF_{(1\times516)}, \quad (9)$$

$$FFV_{(1\times1540)} = FV_{VGG19\ (1\times1024)} + FV1_{GLCM(1\times25)} + FV2_{LBP(1\times236)} + FV3_{PHOG(1\times255)}, \quad (10)$$

The FFV is then used to train, test and validate the classifier considered in the proposed methodology for the automated classification of lung nodules using CT images.

3.3.4. Classifier Implementation

The performance of the DL-based automated disease detection arrangement depends chiefly on the performance of the classifier implemented to categorize the considered test images based on the need. In this paper, a binary classification is initially implemented using the SoftMax classifier and, later, the well-known classifiers, such as Decision Trees (DT), RF, KNN and Support Vector Machine-Radial Basis Function (SVM-RBF) [13,53–56], are also considered to improve the classification task. In this paper, a 10-fold cross-validation process is implemented, and the finest result attained is then considered as the final classification

result. The performance of the classifier is then authenticated and confirmed based on the Image Performance Values (IPV) [57–59].

3.4. Performance Computation and Validation

The overall eminence of the proposed method is validated by computing the essential IPV measures, such as True-Positive (TP), False-Negative (FN), True-Negative (TN), False-Positive (FP), Accuracy (ACC), Precision (PRE), Sensitivity (SEN), Specificity (SPE), Negative-Predicted-Value (NPV), F1-Score (F1S), Jaccard Index and Dice coeeficient, which are calculated in percentages, presented in Equations (11)–(16). The necessary information regarding these values can be found in [45–47].

$$Accuracy = ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{11}$$

$$Precision = PRE = \frac{TP}{TP + FP} \times 100\% \tag{12}$$

$$Sensitivity = SEN = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

$$Specificity = SPE = \frac{TN}{TN + FP} \times 100\% \tag{14}$$

$$Negative\ Predictive\ Value = NPV = \frac{TN}{TN + FN} \times 100\% \tag{15}$$

$$F1 - Score = F1S = \frac{2TP}{2TP + FN + FP} \times 100\% \tag{16}$$

$$Jaccard = \frac{TP}{TP + FN + FP} \times 100\% \tag{17}$$

$$Dice = \frac{2TP}{2TP + FN + FP} \times 100\% \tag{18}$$

4. Results and Discussions

This section demonstrates the results and discussions attained using a workstation with an Intel i5 2.5GHz processor, with 16GB RAM and 2GB VRAM equipped with MATLAB® (version R2018a). Primarily, lung CT images are used as presented in Table 2 and then each image is resized into $224 \times 224 \times 3$ pixels to perform the VGG19-supported segmentation and classification task. Initially, the VGG-SegNet-based lung nodule extraction process is executed on the test images considered, and the sample result obtained for the normal/nodule class image is represented in Figure 4. Figure 4 presents the experimental result of the trained VGG-SegNet with CT images. Figure 4a shows the sample images of the normal/nodule class considered for the assessment; Figure 4b depicts the outcome attained with the final layer of the encoder unit; Figure 4c,d depicts the results of the decoder and the SoftMax classifier, respectively. For the normal (healthy) class image, the decoder will not provide a positive outcome for localization and segmentation, and this section will provide the essential information only for the nodule class.

In this paper, the extracted lung-nodule section with the proposed VGG-SegNet is compared to the ground truth (GT) image generated using ITK-Snap [28] and the essential image measures are calculated as described in previous works [4,13]. The performance of VGG-SegNet is also validated against the existing SegNet and UNet schemes in the literature [24,25,48,49]. The result achieved for the trial image is depicted in Figure 5 and Table 3, respectively. Note that the performance measures [50,51] achieved with VGG-SegNet are superior compared to other approaches.

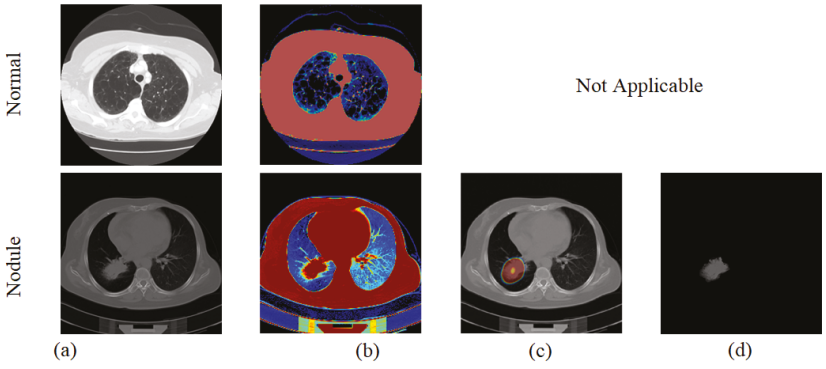


Figure 4. Results obtained with proposed VGG-SegNet scheme: (**a**) text image, (**b**) lung section enhanced by encoder, (**c**) localization of nodule by decoder and (**d**) extracted nodule by SoftMax unit.

(**a**) GT (**b**)VGG-SegNet (**c**) SegNet (**d**) UNet

Figure 5. Segmentation results attained with considered CNN models.

Table 3. Performance evaluation of CNN models on sample lung CT image. Best values are shown in bold.

Approach	Jaccard (%)	Dice (%)	ACC (%)	PRE (%)	SEN (%)	SPE (%)
VGG-SegNet	**82.6464**	**90.4988**	**99.6811**	**98.4496**	83.7363	**99.9756**
SegNet	73.1898	84.5198	99.4539	96.6408	75.1004	99.9471
UNet	79.2308	88.4120	99.6233	93.1525	**84.1307**	99.8925

The segmentation performance of the proposed scheme is then tested using the lung nodules with various dimensions, such as small, medium and large, and the attained results are depicted in Figure 6. This figure confirms that the VGG-SegNet provides a better segmentation on the medium and large nodule dimension and provides reduced segmentation accuracy on the images having lesser lung nodule due to the smaller test image dimension.

After collecting the essential DF with VGG19, the other HCFs, such as GLCM, LBP and PHOG are collected. The GLCM features for the normal (healthy) class image are collected from the whole CT image, and for the abnormal class image it is collected from the binary image of the extracted nodule segment. Figure 7 shows the LBP patterns generated for the normal/nodule class test images with various weight values. During LBP feature collection, each image is treated with the LBP algorithm with various weights (ie, W = 1 to 4) and the 1D features obtained from each image are combined to obtain a 1D feature vector of dimension 1×236.

The PHOG features for the CT images are then extracted by assigning a bin size (L) of 3 and this process helped to obtain a 1×255 vector of features. The sample PHOG features collected for a sample CT image are seen in Figure 8. All these features (GLCM+LBP+PHOG) are then combined to form a HCF vector with a dimension of 1×516 features, following which they are then combined with the DF to improve the lung

nodule detection accuracy. After collecting the essential features, the image classification task is implemented using DF and DF + HCF separately.

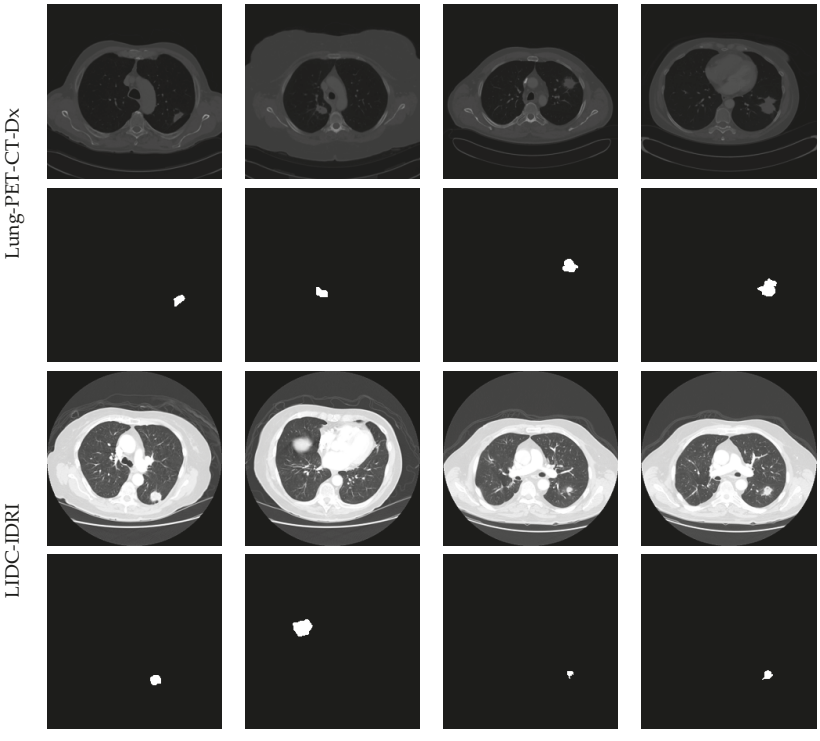


Figure 6. Segmentation of nodule from chosen images of Lung-PET-CT-Dx and LIDC-IDRI dataset.

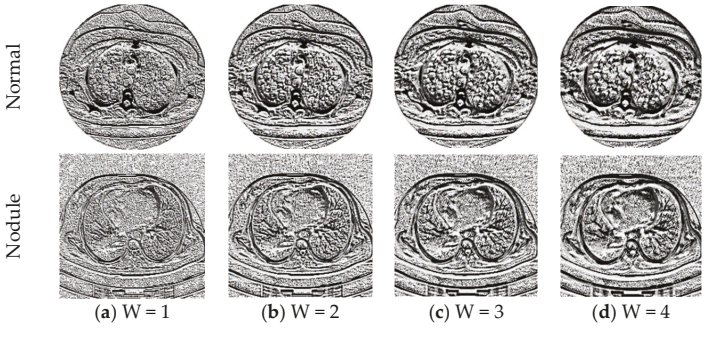


Figure 7. LBP patterns generated from the sample image with various LBP weights.

Initially, the DF-based sorting is executed with the considered CNN schemes and the classification performance obtained with the SoftMax is depicted in Table 4. Figure 9 presents the spider plot for the features considered, and the result of Table 4 and the dimension of the glyph plot confirm that VGG19 helps achieve an enhanced IPV compared to other CNN schemes. VGG19 is chosen as the suitable scheme to examine the considered CT images, and then an attempt is made to enhance the performance of VGG19 using DF + HCF.

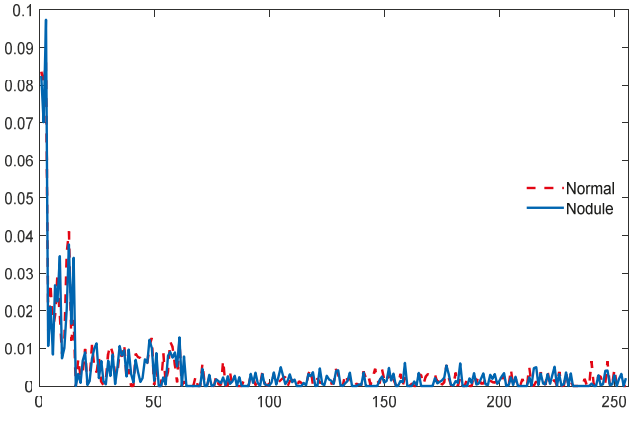


Figure 8. PHOG features obtained with the sample test images of Normal/Nodule class.

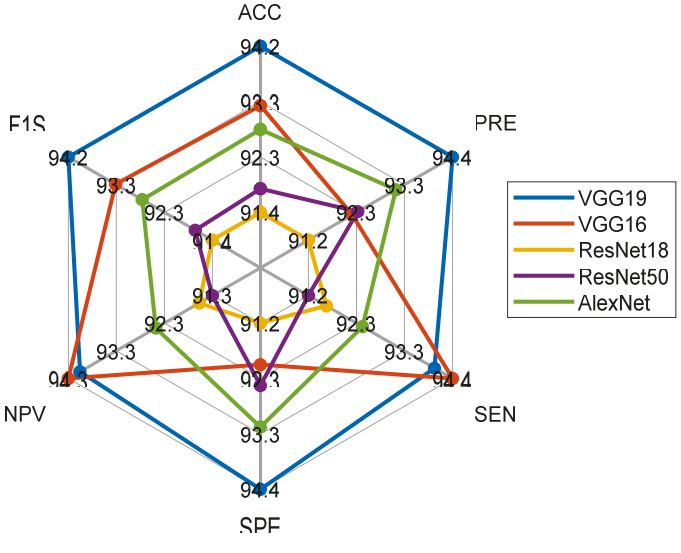


Figure 9. Spider plot to compare the CT image classification performance of CNN models.

Table 4. Classification performance attained with pre-trained DL scheme with DF and SoftMax classifier. Here TP—true positives, FN—false negatives, TN—true negatives, FP—false positives, ACC—accuracy, PRE—precision, SEN—sensitivity, SPE—specificity, NPV—negative predictive value and F1S—F1-score.

DL Scheme (Image Size)	TP	FN	TN	FP	ACC (%)	PRE (%)	SEN (%)	SPE (%)	NPV (%)	F1S (%)
VGG19 (224 × 224 × 3)	235	15	236	14	94.20	94.38	94.00	94.40	94.02	94.19
VGG16 (224 × 224 × 3)	236	14	230	20	93.20	92.19	94.40	92.00	94.26	93.28
ResNet18 (224 × 224 × 3)	229	21	228	22	91.40	91.23	91.60	91.20	91.57	91.42
ResNet50 (224 × 224 × 3)	228	22	231	19	91.80	92.31	91.20	92.40	91.30	91.75
Ale × Net (2274 × 227 × 3)	231	19	233	17	92.80	93.14	92.40	93.20	92.46	92.77

The experiment is then repeated using the VGG19 scheme with the DF + HCF (1 × 1540 features) using classifiers, such as SoftMax, DT, RF, KNN and SVM-RBF; the outcomes are depicted in Table 5. Figure 10 shows the performance of VGG19 with SVM-RBF, in which a 10-fold cross validation is implemented and the best result attained among the 10-fold validation is demonstrated. The result demonstrated in Table 5 confirms that the SVM-RBF classifier offers superior outcome contrast to other classifiers and a graphical illustration in Figure 11 (Glyph-Plot) also confirmed the performance of SVM-RBF. The Receiver-Operating-Characteristic curve (ROC) presented in Figure 12 also confirms the merit of proposed technique.

Table 5. Disease detection performance of VGG19 with DF + HCF with different classifiers. Best values are shown in bold.

Classifier	TP	FN	TN	FP	ACC (%)	PRE (%)	SEN (%)	SPE (%)	NPV (%)	F1S (%)
SoftMax	237	13	244	6	96.20	97.53	94.80	97.60	94.94	96.14
DT	238	12	241	9	95.80	96.36	95.20	96.40	95.25	95.77
RF	240	10	238	12	95.60	95.24	96.00	95.20	95.97	95.62
KNN	241	9	242	8	96.60	96.79	96.40	96.80	96.41	96.59
SVM-RBF	**243**	**7**	**246**	**4**	**97.83**	**98.38**	**97.20**	**98.40**	**97.23**	**97.79**

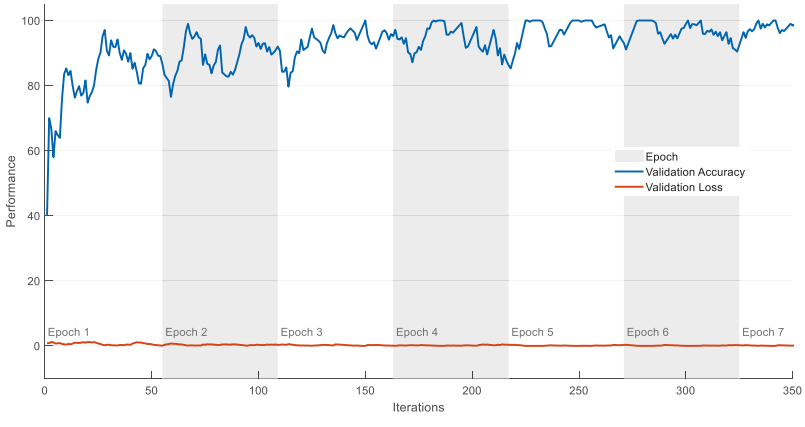


Figure 10. Training performance of the VGG19 with SVM-RBF for lung CT image slices.

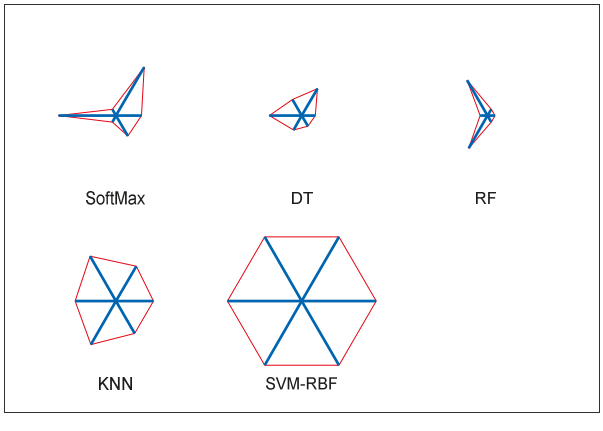


Figure 11. Overall performance of VGG19 with various classifiers summarized as glyph-plots.

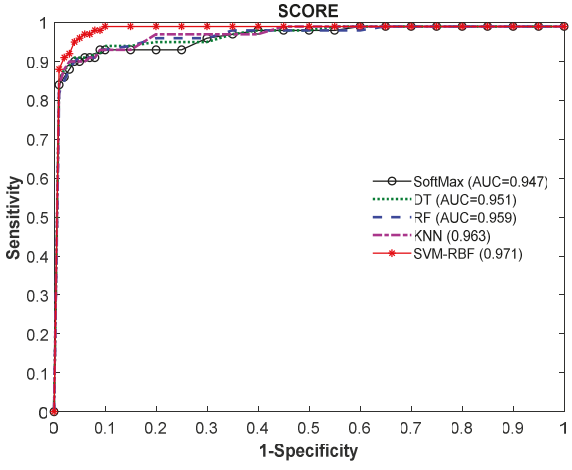


Figure 12. ROC curve attained for VGG19 with DF + HCF.

The above-shown result confirms that the disease detection performance of VGG19 can be enhanced by using both the DF with the HCF. The eminence of the proposed lung nodule detection system is then compared with other methods found in the literature. Figure 13 shows the comparison of the classification precision existing in the literature and the accuracy obtained with the proposed approach (97.83%) is superior compared to other works considered for the study. This confirms the superiority of the proposed approach compared to the existing works.

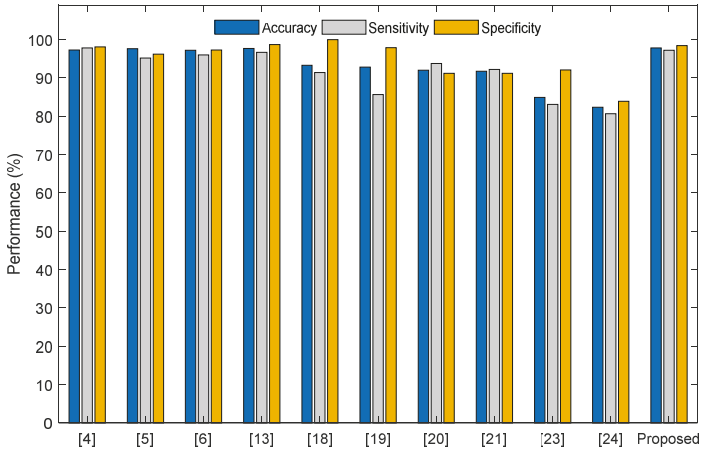


Figure 13. Validation of the disease detection accuracy of the proposed system with existing approaches.

The major improvement of the proposed technique compared to other works, such as Bhandary et al. [4] and Rajinikanth and Kadry [13], is as follows: this paper proposed the detection of lung nodules using CT images without removing the artifact. The number of stages in the proposed approach is lower compared to existing methods [4,10].

The future work includes: (i) considering other hand-made characteristics, such as HOG [48] and GLDM [43], to improve disease detection accuracy, (ii) considering the other variants of the SVM classifiers [43] to achieve better image classification accuracy and (iii) implementing a selected procedure to enhance the segmentation accuracy in lung CT having a lesser nodule size.

5. Conclusions

Due to its clinical significance, several automated disease detection systems have been proposed in the literature to detect lung nodules from CT images. This paper proposes a pre-trained VGG19-based automated segmentation and classification scheme to examine lung CT images. This scheme is implemented in two stages: (i) VGG-SegNet supported extraction of lung nodules from CT images and (ii) classification of lung CT images using deep learning schemes with DF and DF + HCF. The initial part of this work implemented the VGG-SegNet architecture with VGG19-based Encoder-Decoder assembly and extracted the lung nodule section using the SoftMax classifier. Handcrafted features from the test images are extracted using GLCM (1×25 features), LBP with varied weights (1×236 features) and PHOG with an assigned bin = L = 3 (1×255 features), and this combination helped to obtain the chosen HCF with a dimension of 1×516 features. The classification task is initially implemented with the DF and SoftMax, and the result confirmed that the VGG19 provided better result compared to the VGG16, ResNet18, ResNet50 and AlexNet models. The CT image classification performance of VGG19 is once again verified using DF + HCF and the obtained result confirmed that the SVM-RBF classifier helped to obtain better classification accuracy (97.83%).

The limitation of the proposed approach is the dimension of concatenated features (1×1540) which is rather large. In the future, a feature reduction scheme can be considered to reduce this set of features. Also, the performance of the proposed system can be improved by considering other HCFs that are known from the literature.

Author Contributions: All authors have contributed equally to this manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Statement: Not applicable.

Data Availability Statement: The image dataset of this study can be accessed from; https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.

Acknowledgments: The authors of this paper would like to thank The Cancer Imaging Archive for sharing the clinical grade lung CT images for research purpose.

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

References

1. WHO. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 21 September 2021).
2. Olson, E.J. Available online: https://www.mayoclinic.org/diseases-conditions/lung-cancer/expert-answers/lung-nodules/faq-20058445 (accessed on 21 September 2021).
3. Girvin, F.; Ko, J.P. Pulmonary nodules: Detection, assessment, and CAD. *Am. J. Roentgenol.* **2008**, *191*, 1057–1069. [CrossRef] [PubMed]
4. Bhandary, A.; Prabhu, G.A.; Rajinikanth, V.; Thanaraj, K.P.; Satapathy, S.C.; Robbins, D.E.; Shasky, C.; Zhang, Y.D.; Tavares, J.M.R.S.; Raja, N.S.M. Deep-learning framework to detect lung abnormality—A study with chest X-Ray and lung CT scan images. *Pattern Recognit. Lett.* **2020**, *129*, 271–278. [CrossRef]
5. Choi, W.J.; Choi, T.S. Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. *Entropy* **2013**, *15*, 507–523. [CrossRef]
6. Tran, G.S.; Nghiem, T.P.; Nguyen, V.T.; Luong, C.M.; Burie, J.C. Improving accuracy of lung nodule classification using deep learning with focal loss. *J. Healthc. Eng.* **2019**, *2019*, 5156416. [CrossRef] [PubMed]

7. Akram, T.; Attique, M.; Gul, S.; Shahzad, A.; Altaf, M.; Naqvi, S.S.R.; Damaševičius, R.; Maskeliūnas, R. A novel framework for rapid diagnosis of COVID-19 on computed tomography scans. *Pattern Anal. Appl.* **2021**, *24*, 951–964. [CrossRef] [PubMed]
8. Rajaraman, S.; Folio, L.R.; Dimperio, J.; Alderson, P.O.; Antani, S.K. Improved semantic segmentation of tuberculosis—Consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations. *Diagnostics* **2021**, *11*, 616. [CrossRef]
9. Rehman, N.; Zia, M.S.; Meraj, T.; Rauf, H.T.; Damaševičius, R.; El-Sherbeeny, A.M.; El-Meligy, M.A. A self-activated cnn approach for multi-class chest-related COVID-19 detection. *Appl. Sci.* **2021**, *11*, 9023. [CrossRef]
10. Połap, D.; Woźniak, M.; Damaševičius, R.; Wei, W. Chest radiographs segmentation by the use of nature-inspired algorithm for lung disease detection. 2018 IEEE Symposium Series on Computational Intelligence. *SSCI* **2018**, *2019*, 2298–2303. [CrossRef]
11. Capizzi, G.; Sciuto, G.L.; Napoli, C.; Polap, D.; Wozniak, M. Small lung nodules detection based on fuzzy-logic and probabilistic neural network with bioinspired reinforcement learning. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 1178–1189. [CrossRef]
12. Rajaraman, S.; Kim, I.; Antani, S.K. Detection and visualization of abnormality in chest radiographs using modality-specific convolutional neural network ensembles. *PeerJ* **2020**, *8*, e8693. [CrossRef] [PubMed]
13. Rajinikanth, V.; Kadry, S. Development of a framework for preserving the disease-evidence-information to support efficient disease diagnosis. *Int. J. Data Warehous. Min.* **2021**, *17*, 63–84. [CrossRef]
14. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [CrossRef]
15. Armato, S.G., III; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef] [PubMed]
16. Armato, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. Data from LIDC-IDRI, 2015. The Cancer Imaging Archive. Available online: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI (accessed on 24 November 2021).
17. Li, P.; Wang, S.; Li, T.; Lu, J.; Huangfu, Y.; Wang, D. A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis [Data set], 2020. The Cancer Imaging Archive. Available online: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70224216 (accessed on 24 November 2021).
18. Kuruvilla, J.; Gunavathi, K. Lung cancer classification using neural networks for CT images. *Comput. Methods Programs Biomed.* **2014**, *113*, 202–209. [CrossRef] [PubMed]
19. Nascimento, L.B.; de Paiva, A.C.; Silva, A.C. Lung nodules classification in CT images using Shannon and Simpson diversity indices and SVM. In Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition, Berlin, Germany, 13–20 July 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 454–466.
20. Khehrah, N.; Farid, M.S.; Bilal, S.; Khan, M.H. Lung Nodule Detection in CT Images Using Statistical and Shape-Based Features. *J. Imaging* **2020**, *6*, 6. [CrossRef] [PubMed]
21. Wang, W.; Liu, F.; Zhi, X.; Zhang, T.; Huang, C. An Integrated Deep Learning Algorithm for Detecting Lung Nodules with Low-dose CT and Its Application in 6G-enabled Internet of Medical Things. *IEEE Internet Things J.* **2020**, *8*, 5274–5284. [CrossRef]
22. Li, W.; Cao, P.; Zhao, D.; Wang, J. Pulmonary nodule classification with deep convolutional neural networks on computed tomography images. *Comput. Math. Methods Med.* **2016**, *2016*, 6215085. [CrossRef] [PubMed]
23. Kaya, A.; Can, A.B. A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. *J. Biomed. Inform.* **2015**, *56*, 69–79. [CrossRef]
24. Song, Q.; Zhao, L.; Luo, X.; Dou, X. Using deep learning for classification of lung nodules on computed tomography images. *J. Healthc. Eng.* **2017**, *2017*, 8314740. [CrossRef] [PubMed]
25. Shaukat, F.; Raja, G.; Frangi, A.F. Computer-aided detection of lung nodules: A review. *J. Med. Imaging* **2019**, *6*, 020901. [CrossRef]
26. Jia, T.; Zhang, H.; Bai, Y.K. Benign and malignant lung nodule classification based on deep learning feature. *J. Med. Imaging Health Inform.* **2015**, *5*, 1936–1940. [CrossRef]
27. Wang, X.; Mao, K.; Wang, L.; Yang, P.; Lu, D.; He, P. An appraisal of lung nodules automatic classification algorithms for CT images. *Sensors* **2019**, *19*, 194. [CrossRef]
28. Yushkevich, P.A.; Piven, J.; Hazlett, H.C.; Smith, R.G.; Ho, S.; Gee, J.C.; Gerig, G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef]
29. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
30. Kendall, A.; Badrinarayanan, V.; Cipolla, R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv* **2015**, arXiv:1511.02680.
31. Rajaraman, S.; Antani, S. Weakly labeled data augmentation for deep learning: A study on COVID-19 detection in chest X-rays. *Diagnostics* **2020**, *10*, 358. [CrossRef] [PubMed]
32. El Adoui, M.; Mahmoudi, S.A.; Larhmam, M.A.; Benjelloun, M. MRI breast tumor segmentation using different encoder and decoder CNN architectures. *Computers* **2019**, *8*, 52. [CrossRef]

33. Khan, S.A.; Khan, M.A.; Song, O.Y.; Nazir, M. Medical Imaging Fusion Techniques: A Survey Benchmark Analysis, Open Challenges and Recommendations. *J. Med. Imaging Health Inform.* **2020**, *10*, 2523–2531. [CrossRef]
34. Khan, M.A.; Sarfraz, M.S.; Alhaisoni, M.; Albesher, A.A.; Wang, S.; Ashraf, I. StomachNet: Optimal Deep Learning Features Fusion for Stomach Abnormalities Classification. *IEEE Access* **2020**, *8*, 197969–197981. [CrossRef]
35. Khan, M.A.; Ashraf, I.; Alhaisoni, M.; Damaševičius, R.; Scherer, R.; Rehman, A.; Bukhari, S.A.C. Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists. *Diagnostics* **2020**, *10*, 565. [CrossRef]
36. Dey, N.; Rajinikanth, V.; Shi, F.; Tavares, J.M.R.; Moraru, L.; Karthik, K.A.; Lin, H.; Kamalanand, K.; Emmanuel, C. Social-Group-Optimization based tumor evaluation tool for clinical brain MRI of Flair/diffusion-weighted modality. *Biocybern. Biomed. Eng.* **2019**, *39*, 843–856. [CrossRef]
37. Zhang, Y.D.; Satapathy, S.C.; Liu, S.; Li, G.R. A five-layer deep convolutional neural network with stochastic pooling for chest CT-based COVID-19 diagnosis. *Mach. Vis. Appl.* **2020**, *32*, 14. [CrossRef]
38. Zhang, Y.D.; Satapathy, S.C.; Zhu, L.Y.; Górriz, J.M.; Wang, S.H. A seven-layer convolutional neural network for chest CT based COVID-19 diagnosis using stochastic pooling. *IEEE Sens. J.* **2020**. [CrossRef]
39. Kumar, D.; Jain, N.; Khurana, A.; Mittal, S.; Satapathy, S.C.; Senkerik, R.; Hemanth, J.D. Automatic Detection of White Blood Cancer From Bone Marrow Microscopic Images Using Convolutional Neural Networks. *IEEE Access* **2020**, *8*, 142521–142531. [CrossRef]
40. Rodrigues, D.D.A.; Ivo, R.F.; Satapathy, S.C.; Wang, S.; Hemanth, J.; Rebouças Filho, P.P. A new approach for classification skin lesion based on transfer learning, deep learning, and IoT system. *Pattern Recognit. Lett.* **2020**, *136*, 8–15. [CrossRef]
41. Rajinikanth, V.; Joseph Raj, A.N.; Thanaraj, K.P.; Naik, G.R. A Customized VGG19 Network with Concatenation of Deep and Handcrafted Features for Brain Tumor Detection. *Appl. Sci.* **2020**, *10*, 3429. [CrossRef]
42. Arshad, H.; Khan, M.A.; Sharif, M.I.; Yasmin, M.; Tavares, J.M.R.; Zhang, Y.D.; Satapathy, S.C. A multilevel paradigm for deep convolutional neural network features selection with an application to human gait recognition. *Expert Syst.* **2020**, e12541. [CrossRef]
43. Khan, M.A.; Kadry, S.; Alhaisoni, M.; Nam, Y.; Zhang, Y.; Rajinikanth, V.; Sarfraz, M.S. Computer-Aided Gastrointestinal Diseases Analysis from Wireless Capsule Endoscopy: A Framework of Best Features Selection. *IEEE Access* **2020**, *8*, 132850–132859. [CrossRef]
44. Akram, T.; Khan, M.A.; Sharif, M.; Yasmin, M. Skin lesion segmentation and recognition using multichannel saliency estimation and M-SVM on selected serially fused features. *J. Ambient Intell. Humaniz. Comput.* **2018**, 1–20. [CrossRef]
45. Khan, M.A.; Khan, M.A.; Ahmed, F.; Mittal, M.; Goyal, L.M.; Hemanth, D.J.; Satapathy, S.C. Gastrointestinal diseases segmentation and classification based on duo-deep architectures. *Pattern Recognit. Lett.* **2020**, *131*, 193–204. [CrossRef]
46. Saba, T.; Mohamed, A.S.; El-Affendi, M.; Amin, J.; Sharif, M. Brain tumor detection using fusion of hand crafted and deep learning features. *Cogn. Syst. Res.* **2020**, *59*, 221–230. [CrossRef]
47. Batool, F.E.; Attique, M.; Sharif, M.; Javed, K.; Nazir, M.; Abbasi, A.A.; Iqbal, Z.; Riaz, N. Offline signature verification system: A novel technique of fusion of GLCM and geometric features using SVM. *Multimed. Tools Appl.* **2020**, 1–20. [CrossRef]
48. Murtza, I.; Khan, A.; Akhtar, N. Object detection using hybridization of static and dynamic feature spaces and its exploitation by ensemble classification. *Neural Comput. Appl.* **2019**, *31*, 347–361. [CrossRef]
49. Bakiya, A.; Kamalanand, K.; Rajinikanth, V.; Nayak, R.S.; Kadry, S. Deep neural network assisted diagnosis of time-frequency transformed electromyograms. *Multimed. Tools Appl.* **2020**, *79*, 11051–11067. [CrossRef]
50. Acharya, U.R.; Fernandes, S.L.; WeiKoh, J.E.; Ciaccio, E.J.; Fabell, M.K.M.; Tanik, U.J.; Rajinikanth, V.; Yeong, C.H. Automated detection of Alzheimer's disease using brain MRI images—A study with various feature extraction techniques. *J. Med. Syst.* **2019**, *43*, 302. [CrossRef] [PubMed]
51. Jahmunah, V.; Oh, S.L.; Rajinikanth, V.; Ciaccio, E.J.; Cheong, K.H.; Arunkumar, N.; Acharya, U.R. Automated detection of schizophrenia using nonlinear signal processing methods. *Artif. Intell. Med.* **2019**, *100*, 101698. [CrossRef] [PubMed]
52. Gudigar, A.; Raghavendra, U.; Devasia, T.; Nayak, K.; Danish, S.M.; Kamath, G.; Samanth, J.; Pai, U.M.; Nayak, V.; Tan, R.S.; et al. Global weighted LBP based entropy features for the assessment of pulmonary hypertension. *Pattern Recognit. Lett.* **2019**, *125*, 35–41. [CrossRef]
53. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
54. Kadry, S.; Rajinikanth, V.; Raja, N.S.M.; Hemanth, D.J.; Hannon, N.M.; Raj, A.N.J. Evaluation of brain tumor using brain MRI with modified-moth-flame algorithm and Kapur's thresholding: A study. *Evol. Intell.* **2021**, *14*, 1053–1063. [CrossRef]
55. Meraj, T.; Rauf, H.T.; Zahoor, S.; Hassan, A.; Lali, M.I.; Ali, L.; Bukhari, S.A.C.; Shoaib, U. Lung nodules detection using semantic segmentation and classification with optimal features. *Neural Comput. Appl.* **2021**, *33*, 10737–10750. [CrossRef]
56. Aziz, A.; Tariq, U.; Nam, Y.; Nazir, M.; Jeong, C.-W.; Mostafa, R.R.; Sakr, R.H. An Ensemble of Optimal Deep Learning Features for brain tumor classification. *Comput. Mater. Continua* **2021**, *69*, 2653–2670. [CrossRef]
57. Sharif, M.I.; Alhussein, M.; Aurangzeb, K.; Raza, M. A decision support system for multimodal brain tumor classification using deep learning. *Complex Intell. Syst.* **2021**, *3*, 1–14. [CrossRef]

58. Albahli, S.; Rauf, H.T.; Arif, M.; Nafis, M.T.; Algosaibi, A. Identification of thoracic diseases by exploiting deep neural networks. *Neural Netw.* **2021**, *5*, 6. [CrossRef]
59. Albahli, S.; Rauf, H.T.; Algosaibi, A.; Balas, V.E. AI-driven deep CNN approach for multi-label pathology classification using chest X-Rays. *PeerJ Comput. Sci.* **2021**, *7*, e495. [CrossRef] [PubMed]

diagnostics

MDPI

Article

Inter-Variability Study of COVLIAS 1.0: Hybrid Deep Learning Models for COVID-19 Lung Segmentation in Computed Tomography

Jasjit S. Suri [1,2,*], Sushant Agarwal [2,3], Pranav Elavarthi [2,4], Rajesh Pathak [5], Vedmanvitha Ketireddy [6], Marta Columbu [7], Luca Saba [7], Suneet K. Gupta [8], Gavino Faa [9], Inder M. Singh [1], Monika Turk [10], Paramjit S. Chadha [1], Amer M. Johri [11], Narendra N. Khanna [12], Klaudija Viskovic [13], Sophie Mavrogeni [14], John R. Laird [15], Gyan Pareek [16], Martin Miner [17], David W. Sobel [16], Antonella Balestrieri [7], Petros P. Sfikakis [18], George Tsoulfas [19], Athanasios Protogerou [20], Durga Prasanna Misra [21], Vikas Agarwal [21], George D. Kitas [22,23], Jagjit S. Teji [24], Mustafa Al-Maini [25], Surinder K. Dhanjil [26], Andrew Nicolaides [27], Aditya Sharma [28], Vijay Rathore [26], Mostafa Fatemi [29], Azra Alizad [30], Pudukode R. Krishnan [31], Ferenc Nagy [32], Zoltan Ruzsa [33], Archna Gupta [34], Subbaram Naidu [35] and Mannudeep K. Kalra [36]

1 Stroke Diagnostic and Monitoring Division, AtheroPoint™, Roseville, CA 95661, USA; drindersingh1@gmail.com (I.M.S.); pomchadha@gmail.com (P.S.C.)
2 Advanced Knowledge Engineering Centre, GBTI, Roseville, CA 95661, USA; sushant.ag09@gmail.com (S.A.); pmelavarthi@gmail.com (P.E.)
3 Department of Computer Science Engineering, PSIT, Kanpur 209305, India
4 Thomas Jefferson High School for Science and Technology, Alexandria, VA 22312, USA
5 Department of Computer Science Engineering, Rawatpura Sarkar University, Raipur 492001, India; drrkpathak20@gmail.com
6 Mira Loma High School, Sacramento, CA 95821, USA; manvi.ketireddy@gmail.com
7 Department of Radiology, Azienda Ospedaliero Universitaria (A.O.U.), 10015 Cagliari, Italy; martagiuliacol@gmail.com (M.C.); lucasabamd@gmail.com (L.S.); antonellabalestrieri@hotmail.com (A.B.)
8 Department of Computer Science, Bennett University, Noida 201310, India; suneet.gupta@bennett.edu.in
9 Department of Pathology, Azienda Ospedaliero Universitaria (A.O.U.), 10015 Cagliari, Italy; gavinofaa@gmail.com
10 The Hanse-Wissenschaftskolleg Institute for Advanced Study, 27753 Delmenhorst, Germany; monika.turk84@gmail.com
11 Department of Medicine, Division of Cardiology, Queen's University, Kingston, ON K7L 3N6, Canada; johria@queensu.ca
12 Department of Cardiology, Indraprastha APOLLO Hospitals, New Delhi 110076, India; drnnkhanna@gmail.com
13 University Hospital for Infectious Diseases, 10000 Zagreb, Croatia; klaudija.viskovic@bfm.hr
14 Cardiology Clinic, Onassis Cardiac Surgery Center, 10558 Athens, Greece; soma13@otenet.gr
15 Heart and Vascular Institute, Adventist Health St. Helena, St. Helena, CA 94574, USA; Lairdjr@ah.org
16 Minimally Invasive Urology Institute, Brown University, Providence, RI 02912, USA; gyan_pareek@brown.edu (G.P.); dwsobel@gmail.com (D.W.S.)
17 Men's Health Center, Miriam Hospital, Providence, RI 02906, USA; martin_miner@brown.edu
18 Rheumatology Unit, National & Kapodistrian University of Athens, 10679 Athens, Greece; psfikakis@med.uoa.gr
19 Aristoteleion University of Thessaloniki, 54636 Thessaloniki, Greece; tsoulfasg@gmail.com
20 National & Kapodistrian University of Athens, 10679 Athens, Greece; aprotog@med.uoa.gr
21 Department of Immunology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow 226014, India; durgapmisra@gmail.com (D.P.M.); vikasagr@yahoo.com (V.A.)
22 Academic Affairs, Dudley Group NHS Foundation Trust, Dudley DY1 2HQ, UK; george.kitas@nhs.net
23 Arthritis Research UK Epidemiology Unit, Manchester University, Manchester M13 9PT, UK
24 Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, IL 60611, USA; jteji@mercy-chicago.org
25 Allergy, Clinical Immunology and Rheumatology Institute, Toronto, ON L4Z 4C4, Canada; almaini@hotmail.com
26 AtheroPoint LLC, Roseville, CA 95611, USA; surinderdhanjil@gmail.com (S.K.D.); Vijay.s.rathore@kp.org (V.R.)
27 Vascular Screening and Diagnostic Centre, University of Nicosia Medical School, Nicosia 2368, Cyprus; anicolaides1@gmail.com

Citation: Suri, J.S.; Agarwal, S.; Elavarthi, P.; Pathak, R.; Ketireddy, V.; Columbu, M.; Saba, L.; Gupta, S.K.; Faa, G.; Singh, I.M.; et al. Inter-Variability Study of COVLIAS 1.0: Hybrid Deep Learning Models for COVID-19 Lung Segmentation in Computed Tomography. *Diagnostics* **2021**, *11*, 2025. https://doi.org/10.3390/diagnostics11112025

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 30 September 2021
Accepted: 27 October 2021
Published: 1 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[28] Division of Cardiovascular Medicine, University of Virginia, Charlottesville, VA 22904, USA; AS8AH@hscmail.mcc.virginia.edu
[29] Department of Physiology & Biomedical Engineering, Mayo Clinic College of Medicine and Science, Rochester, MN 55905, USA; fatemi.mostafa@mayo.edu
[30] Department of Radiology, Mayo Clinic College of Medicine and Science, Rochester, MN 55905, USA; Alizad.Azra@mayo.edu
[31] Neurology Department, Fortis Hospital, Bangalore 560076, India; prkrish12@rediffmail.com
[32] Internal Medicine Department, University of Szeged, 6725 Szeged, Hungary; drnagytfer@hotmail.com
[33] Zoltan Invasive Cardiology Division, University of Szeged, 6725 Szeged, Hungary; zruzsa@icloud.com
[34] Radiology Department, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow 226014, India; garchna@gmail.com
[35] Electrical Engineering Department, University of Minnesota, Duluth, MN 55812, USA; dsnaidu@d.umn.edu
[36] Department of Radiology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA; mkalra@mgh.harvard.edu
* Correspondence: jasjit.suri@atheropoint.com; Tel.: +1-(916)-749-5628

Abstract: *Background*: For COVID-19 lung severity, segmentation of lungs on computed tomography (CT) is the first crucial step. Current deep learning (DL)-based Artificial Intelligence (AI) models have a bias in the training stage of segmentation because only one set of ground truth (GT) annotations are evaluated. We propose a robust and stable inter-variability analysis of CT lung segmentation in COVID-19 to avoid the effect of bias. *Methodology*: The proposed inter-variability study consists of two GT tracers for lung segmentation on chest CT. Three AI models, PSP Net, VGG-SegNet, and ResNet-SegNet, were trained using GT annotations. We hypothesized that if AI models are trained on the GT tracings from multiple experience levels, and if the AI performance on the test data between these AI models is within the 5% range, one can consider such an AI model robust and unbiased. The K5 protocol (training to testing: 80%:20%) was adapted. Ten kinds of metrics were used for performance evaluation. *Results*: The database consisted of 5000 CT chest images from 72 COVID-19-infected patients. By computing the coefficient of correlations (CC) between the output of the two AI models trained corresponding to the two GT tracers, computing their differences in their CC, and repeating the process for all three AI-models, we show the differences as 0%, 0.51%, and 2.04% (all < 5%), thereby validating the hypothesis. The performance was comparable; however, it had the following order: ResNet-SegNet > PSP Net > VGG-SegNet. *Conclusions*: The AI models were clinically robust and stable during the inter-variability analysis on the CT lung segmentation on COVID-19 patients.

Keywords: COVID-19; computed tomography; lungs; variability; segmentation; hybrid deep learning

1. Introduction

The WHO's International Health Regulations and Emergency Committee (IHREC) proclaimed COVID-19 a "public health emergency of international significance" or "pandemic" on 30 January 2020. More than 231 million people have been infected worldwide, and nearly 4.7 million people have died due to COVID-19 [1]. Although this "severe acute respiratory syndrome coronavirus 2" (SARS-CoV-2) virus specifically targets the pulmonary and vascular system, it has the potential to travel through the body and lead to complications such as pulmonary embolism [2], myocardial infarction, stroke, or mesenteric ischemia [3–5]. Comorbidities such as diabetes mellitus, hypertension, and obesity substantially increase the severity and mortality of COVID-19 [6,7]. A real-time reverse transcription-polymerase chain reaction (RT-PCR) is the recommended method for diagnosis [8]. Chest radiographs and computed tomography (CT) [9–11] are used to determine disease severity in patients with moderate to severe disease or underlying comorbidities based on the extent of pulmonary opacities such as ground-glass (GGO), consolidation, and mixed opacities in CT scans [7,12–14].

Most radiologists provide a semantic description of the extent and type of opacities to describe the severity of COVID-19 pneumonia. The semiquantitative evaluation of pulmonary opacities is time-consuming, subjective, and tedious [15–18]. Thus, there is a need for a fast and error-free early COVID-19 disease diagnosis and real-time prognosis solutions. Machine learning (ML) offers a solution to this problem by providing a rich set of algorithms [19]. Previously, ML has been used for detection of cancers in breast [20], liver [21,22], thyroid [23–25], skin [26,27], prostate [28,29], ovary [30], and lung [31]. There are two main components in disease detection, i.e., segmentation [32–35] and classification [36,37], where segmentation plays a crucial step. An extension of ML called deep learning (DL) employs dense layers to automatically extract and classify all relevant imaging features [38–43]. Hybrid DL (HDL), a method that combines two DL systems, helps address some of the challenges in solo DL models [44,45]. This includes overfitting and optimization of hyperparameters, thereby removing the bias [45].

During the AI model training, the most crucial stage is the ground truth (GT) annotation of organs that need to be segmented. It is a time-consuming operation with monetary constraints since skilled personnel such as radiologists are expensive to recruit and difficult to find. These annotations, if conducted by one tracer, make the AI system biased. A plurality of tracers being used to produce the GT annotated dataset makes the system more resilient and lowers the AI bias [46–49]. This is because the AI model can grasp and adjust to the sensitivity of the difference in the tracings of the tracers. Thus, to avoid AI bias, one needs to have an automated AI-based system with multiple tracers. To establish the validity of such automated AI systems, one must undergo inter-variability analysis with two or more observers.

To validate the AI systems, we hypothesize that two conditions must be met: (a) the two observers should perform within 5% range of each other and (b) the performance of the AI system using the ground truth tracings from these two observers should also be within the 5% threshold [48]. The AI performance is computed between the GT-area and the AI model-estimated area. The focus of the proposed research is to design a reliable AI system based on the inter-observer paradigm. Figure 1 depicts a COVID-19 CT lung segmentation system in which the CT machine is used to acquire CT volumes. This volume is then annotated by multiple observers (Figure 1, n denotes the number of observers), and multiple AI models are generated, which is then used for lung segmentation. The segmentation output is the binary mask of the lung, its boundary, and the corresponding boundary overlays. This output can be used for evaluating the performance, analysis, and quantification of the results.

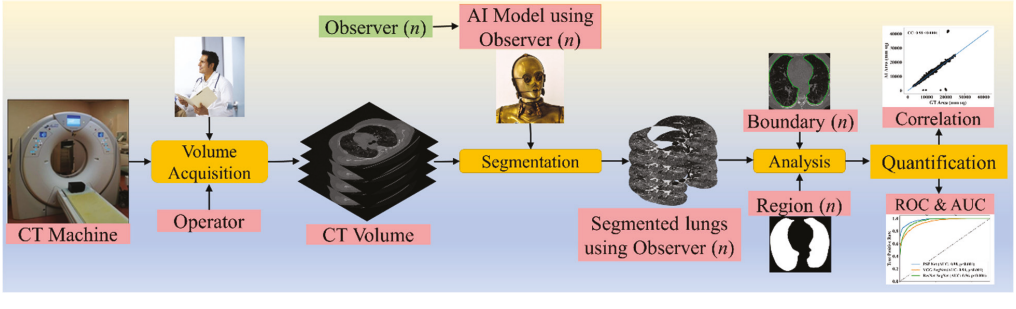


Figure 1. COVLIAS 1.0: Inter-variability analysis of CT-based lung segmentation and quantification system for COVID-19 patients. ROC: Receiver operating characteristic; AUC: Area-under-the-curve.

The layout of this inter-variability study is as follows: Section 2 presents the methodology with the demographics, COVLIAS 1.0 pipeline, AI architectures, and loss functions. The experi-

mental protocol is shown in Section 3, while results and performance evaluation are presented in Section 4. The discussions and conclusions are presented in Sections 5 and 6, respectively.

2. Methodology

2.1. Patient Demographics, Image Acquisition, and Data Preparation

2.1.1. Demographics

The dataset consists of 72 adult Italian patients with 46 being male and the remaining being female. The mean height and weight were 173 cm and 79 kg, respectively. A total of 60 patients tested positive for RT-PCR, while 12 patients were confirmed using broncho-alveolar lavage [50]. Overall, the cohort had an average of 4.1 GGO, which was considered low.

2.1.2. Image Acquisition

All chest CT scans were performed in a supine posture during a single full inspiratory breath-hold using a 128-slice multidetector-row Philips Healthcare's "Philips Ingenuity Core" CT scanner. There were no intravenous or oral contrast media administrations. The CT exams were performed using a 120 kV, 226 mAs/slice (utilizing an automatic tube current modulation—Z-DOM by Philips), a 1.08 spiral pitch factor, 0.5-s gantry rotation time, and 64 × 0.625 detector setup. Soft tissue kernel with 512 × 512 matrix (mediastinal window) and lung kernel with 768 × 768 matrix (lung window) was used to reconstruct 1 mm-thick images. The Picture Archiving and Communication System (PACS) workstation that was utilized to review the CT images was outfitted with two Eizo 35 × 43 cm displays with a 2048 × 1536 matrix. Figure 2 shows the raw sample CT scans of COVID-19 patients with varying lung sizes and variable intensity patterns, posing a challenge.

2.1.3. Data Preparation

The proposed study makes use of the CT data of 72 COVID-positive individuals. Each patient had 200 slices, out of which the radiologist [LS] chose 65–70 slices from the visible lung region, resulting in 5000 images in total. The AI-based segmentation models were trained and tested using these 5000 images. To prepare the data for segmentation, a binary mask was created manually in a selected slice with the help of ImgTracer™ under the supervision of a qualified radiologist [LS] (Global Biomedical Technologies, Inc., Roseville, CA, USA) [47,48,51]. Figure 3 shows the white binary mask of the lung region computed using ImgTracer™ during manual tracings by Observer 1 and 2 (both were postgraduate researchers trained by our radiological team).

2.2. Architecture

COVLIAS 1.0 system incorporates three models: one solo DL (SDL) and two hybrid DL (HDL). The proposed study incorporates three AI models: (a) PSP Net, (b) VGG-SegNet, and (c) ResNet-SegNet.

2.2.1. Three AI Models: PSP Net, VGG-SegNet, and ResNet-SegNet

The Pyramid Scene Parsing Network (PSP Net) [52] is a semantic segmentation network with the ability to consider the global context of the image. The architecture of PSP Net (Figure 4) has four parts: (i) input, (ii) feature map, (iii) pyramid pooling module, and (iv) output. The input to the network is the image to be segmented, which undergoes extraction of the feature map using a set of dilated convolution and pooling blocks. The dilated convolution layer is added at the last two blocks of the network to keep more prominent features at the end. The next stage is the pyramid pooling module; it is the heart of the network, as it helps capture the global context of the image/feature map generated in the previous step. This section consists of four parts, each with a different scaling ability. The scaling of this module includes 1, 2, 3, and 6, where 1 × 1 scaling helps capture the spatial features and thereby increases the resolution of the features captured.

The 6×6 scaling captures the higher-resolution features. At the end of this module, all the output from these four parts is pooled using global average pooling. For the last part, the global average pooling output is fed to a set of convolutional layers. Finally, the set of prediction classes are generated as the output binary mask.

Figure 2. Raw lung COVID-19 CT scans taken from different patients in the database.

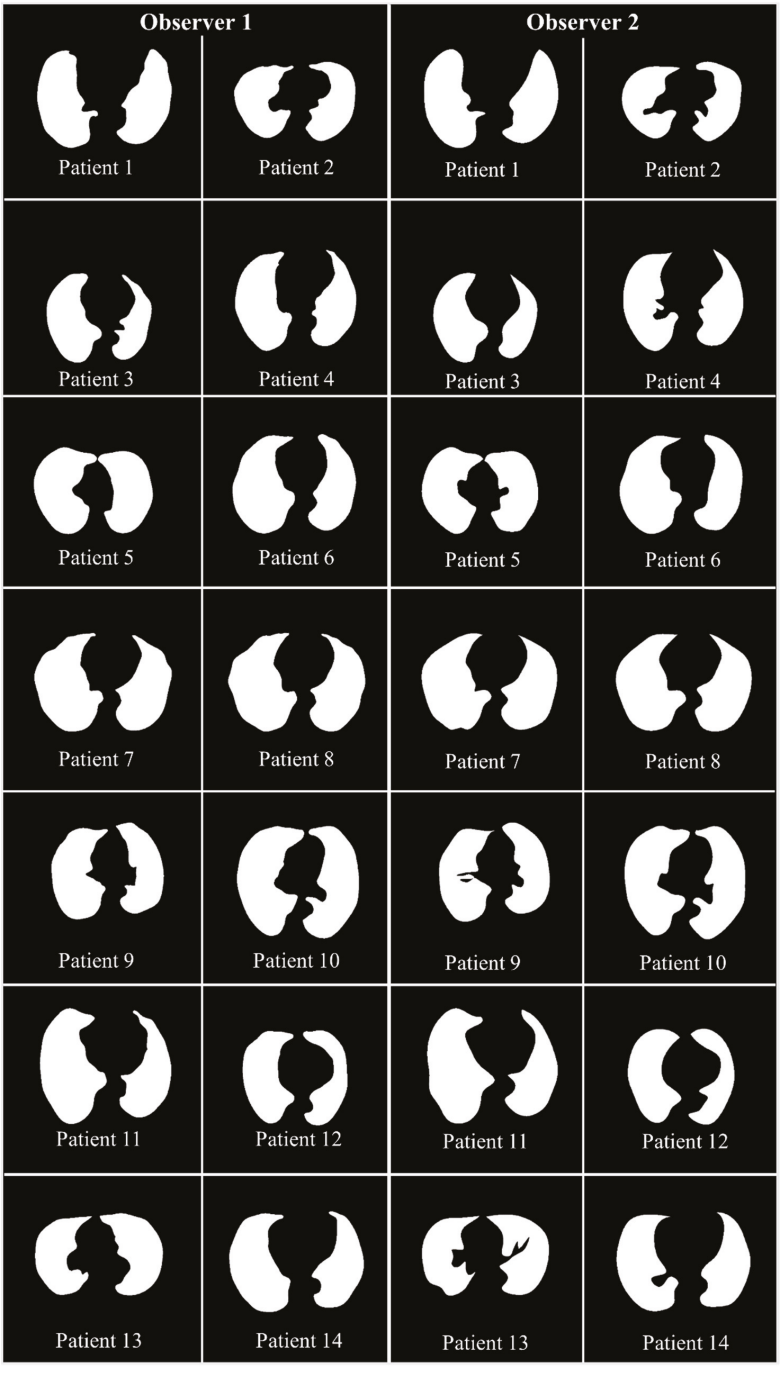


Figure 3. GT white binary mask for AI model training for Observer 1 vs. Observer 2.

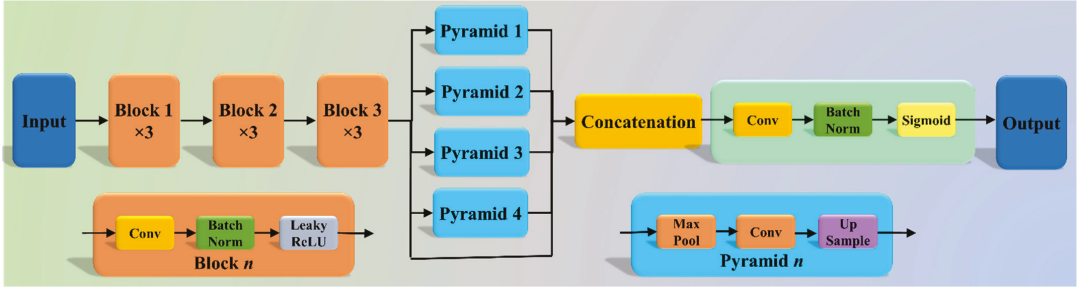


Figure 4. PSP Net architecture.

The VGGNet architecture (Figure 5) was designed to reduce the training time by replacing the kernel filter in the initial layer with an 11 and 5 sized filter, thereby reducing the # of parameters in the two-dimension convolution (Conv) layers [53]. The VGG-SegNet architecture used in this study is composed of three parts (i) encoder, (ii) decoder part, and (iii) a pixel-wise SoftMax classifier at the end. It consists of 16 Conv layers compared to the SegNet architecture, where only 13 Conv layers are used [54] in the encoder part. This increase in #layers helps the model extract more features from the image. The final output of the model is a binary mask with the lung region annotated as 1 (white) and the rest of the image as 0 (black).

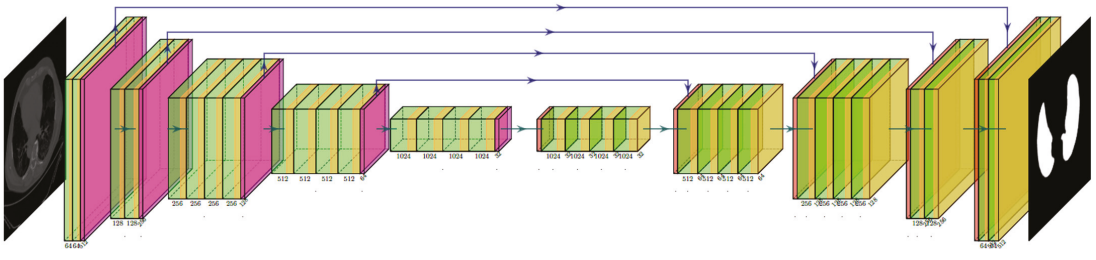


Figure 5. VGG-SegNet architecture.

Although VGGNet was very efficient and fast, it suffered from the problem of vanishing gradients. It results in significantly less or no weight training during backpropagation; at each epoch, it keeps getting multiplied with the gradient, and the update to the initial layers is very small. To overcome this problem, Residual Network or ResNet [55] came into existence (Figure 6). In this architecture, a new connection was introduced known as skip connection which allowed the gradients to bypass a certain number of layers, solving the vanishing gradient problem. Moreover, with the help of one more additions to the network, i.e., an identity function, the local gradient value was kept to one during the backpropagation step.

2.2.2. Loss Functions for AI Models

The proposed system uses cross-entropy (CE)-loss during the training of the AI models. Equation (1) below represents the CE-loss, symbolized as l_{CE}, for the three AI models:

$$l_{CE} = -[(x_i \times \log p_i) + (1 - x_i) \times \log(1 - p_i)] \quad (1)$$

where x_i represents the input GT label 1, $(1 - x_i)$ represents the GT label 0, p_i represents the probability of the classifier (SoftMax) used at the last layer of the AI model, and $\times$ represents the product of the two terms. Figures 4–6 presents the three AI architectures that have been trained using the CE-loss function.

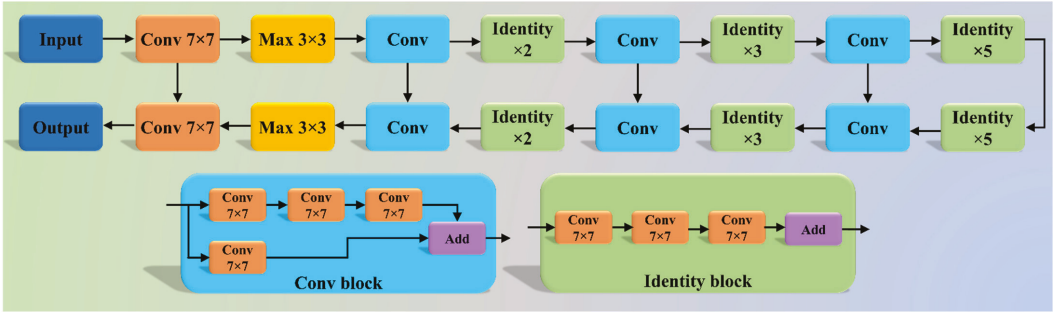


Figure 6. ResNet-SegNet architecture.

3. Experimental Protocol

3.1. Accuracy Estimation of AI Models Using Cross-Validation

A standardized cross-validation (CV) protocol was adapted for determining the accuracy of the AI models. Our group has published several CV-based protocols of different kinds using AI framework [27,30,37,56,57]. Since the data were moderate, the K5 protocol was used, which consisted of 80% training data (4000 CT images) and 20% testing (1000 CT images). Five folds were designed in such a way that each fold got a chance to have a unique test set. An internal validation mechanism was part of the K5 protocol where 10% data was considered for validation.

3.2. Lung Quantification

There were two methods used for quantification of the segmented lungs using AI models. The spirit of these two methods originates from the shape analysis concept. In the first method, lung area (LA) is computed since the region is balloon-shaped, thus the area parameter is well suited for the measurement [58,59]. In the second method, we compute the long-axis of the lung (LLA) since the shape of the lung is more longitudinal than circular. A similar approach was taken for the long-axis view in heart computation [60]. The lung area (LA) was calculated by counting the number of white pixels in the binary mask segmented lungs, and the lung long axis (LLA) was calculated by the most distant distance segment joining anterior to posterior of the lungs. A resolution factor of 0.52 was used to convert (i) pixel to mm^2 for the LA and (ii) pixel to mm for the LLA computation and quantification.

If the total number of the image is represented by N in the database, $A_{ai}(m,n)$ represents lung area for in the image "n" using the AI model "m", $\overline{A}_{ai}(m)$ represents the mean lung area corresponding to the AI model "m," and mean area of the GT binary mask is represented by $\overline{A}_{gt}$, then mathematically $\overline{A}_{ai}(m)$ and $\overline{A}_{gt}$ can be computed as shown in Equation (2).

$$\left.\begin{aligned} \overline{A}_{ai}(m) &= \frac{\sum_{n=1}^{N} A_{ai}(m,n)}{N} \\ \overline{A}_{gt} &= \frac{\sum_{n=1}^{N} A_{gt}(n)}{N} \end{aligned}\right\} \quad (2)$$

Similarly, $LA_{ai}(m,n)$ represents LLA for in the image "n" using the AI model "m", $\overline{LA}_{ai}(m)$ represents the mean LLA corresponding to the AI model "m,"$\overline{LA}_{gt}$ represents the

corresponding mean LLA of the GT binary lung mask, then mathematically $\overline{LA}_{ai}(m)$ and $\overline{LA}_{gt}$ can be computed as shown in Equation (3).

$$\left.\begin{aligned} \overline{LA}_{ai}(m) &= \frac{\sum_{n=1}^{N} LA_{ai}(m,n)}{N} \\ \overline{LA}_{gt} &= \frac{\sum_{n=1}^{N} LA_{gt}(n)}{N} \end{aligned}\right\} \tag{3}$$

3.3. AI Model Accuracy Computation

The accuracy of the AI system was measured by comparing the predicted output and the ground truth pixel values. These values were interpreted as binary (0 or 1) numbers as the output lung mask was only black and white, respectively. Finally, these binary numbers were summed up and divided by the total number of pixels in the image. If TP, TN, FN, and FP represent true positive, true negative, false negative, and false positive, then the accuracy of the AI system can be computed as shown in Equation (4) [61].

$$\text{ACC (ai) (\%)} = \left(\frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}\right) \times 100 \tag{4}$$

4. Results and Performance Evaluation

4.1. Results

Previously, COVLIAS 1.0 [54] was designed to run on a training: testing ratio of 2:3 dataset from 5000 images. However, this study proposes an inter-observer variability study with K5 in a CV framework. The training was performed on two sets of annotations, i.e., Observer 1 and Observer 2. The output results are similar to the previously published study, i.e., a binary mask of the segmented lungs. Figures 7–9 show the AI-generated binary mask, segmented lung, and color segmented lung with grayscale background as an overlay for the three AI models.

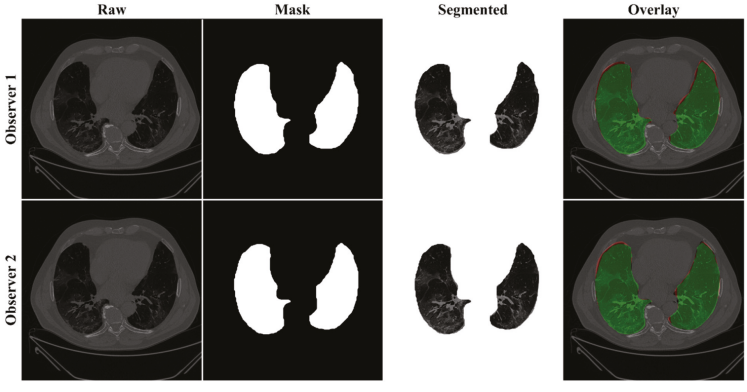


Figure 7. Results from PSP Net while using Observers 1 and 2. Columns are the raw, binary mask output, segmented lung region, and overlay of the estimated lung region vs. ground truth region.

4.2. Performance Evaluation

This section deals with the performance evaluation (PE) of the three AI models for Observer 1 vs. Observer 2. Section 4.2.1 presents the visual comparison of the results, which includes (i) boundary overlays against the ground truth boundary and (ii) lung long axis against the ground truth axis. Section 4.2.2 shows the PE for lung area error, which consists of (i) cumulative frequency (CF) plot, (ii) Bland-Altman plot, (iii) Jaccard Index (JI) and Dice Similarity (DS), and (iv) ROC and AUC curves for the three AI-based models' performance for Observer 1 vs. Observer 2. Similarly, lung long axis error (LLAE) presents PE using

(i) cumulative plot, (ii) correlation coefficient (CC), and (iii) Bland-Altman plot. Finally, statistical analyses of the LA and LLA are presented using paired *t*-test, Wilcoxon, Mann-Whitney, and CC values for all 12 possible combinations for three AI models between Observer 1 and Observer 2.

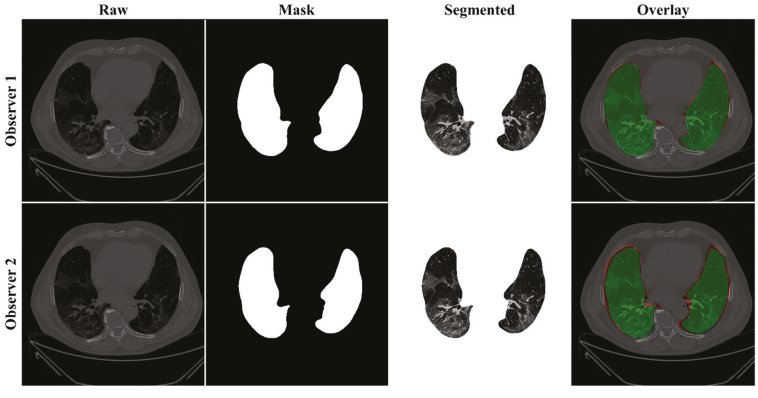


Figure 8. Results from VGG-SegNet while using Observers 1 and 2. Columns are the raw, binary mask output, segmented lung region, and overlay of the estimated lung region vs. ground truth region.

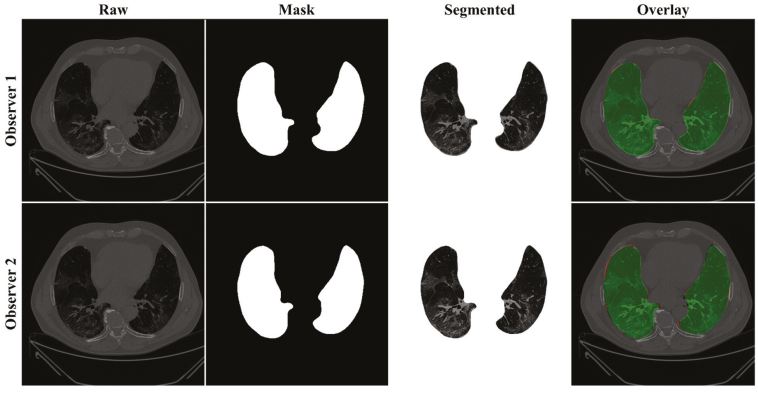


Figure 9. Results from ResNet-SegNet while using Observers 1 and 2. Columns are the raw, binary mask output, segmented lung region, and overlay of the estimated lung region vs. ground truth region.

4.2.1. Lung Boundary and Long Axis Visualization

The overlay for the three AI model boundaries (green) and GT-boundary (red) corresponding to Observer 1 (left) and Observer 2 (right) with a grayscale COVID-19 CT slice in the background is shown in Figure 10, while Figure 11 shows the AI-long axis (green) and GT-long axis (red) between Observer 1 and Observer 2 for three AI models. It shows the reach of anterior to posterior of the left and right lungs, with the GT boundary (white) corresponding to Observer 1 (left) and Observer 2 (right) of the lungs by the tracer using ImgTracer™. The three AI models follow the order: PSP Net, VGG-SegNet, and ResNet-SegNet.

4.2.2. Performance Metrics for the Lung Area Error

Cumulative Frequency Plot for Lung Area Error

The frequency of occurrence of the LAE is compared to a reference value in the cumulative frequency analysis and shown in Figure 12 (left lung) and Figure 13 (right lung) for three AI models between Observer 1 and Observer 2. A cutoff-score of 80% was chosen to show the difference between the three AI models. The LAE with the selected cutoff for the left lung was 1123.36 mm^2, 725.90 mm^2, and 571.65 mm^2 for the three AI models using Observer 1, and 834.08 mm^2, 1730.58 mm^2, and 683.42 mm^2, respectively, for the three AI models using Observer 2. A similar trend was followed by the right lung with 1158.93 mm^2, 612.47 mm^2, and 532.44 mm^2 for the three AI models using Observer 1, and 809.77 mm^2, 1610.15 mm^2, and 572.56 mm^2, respectively, for the three AI models using Observer 2. The three AI models follow the order: PSP Net, VGG-SegNet, and ResNet-SegNet.

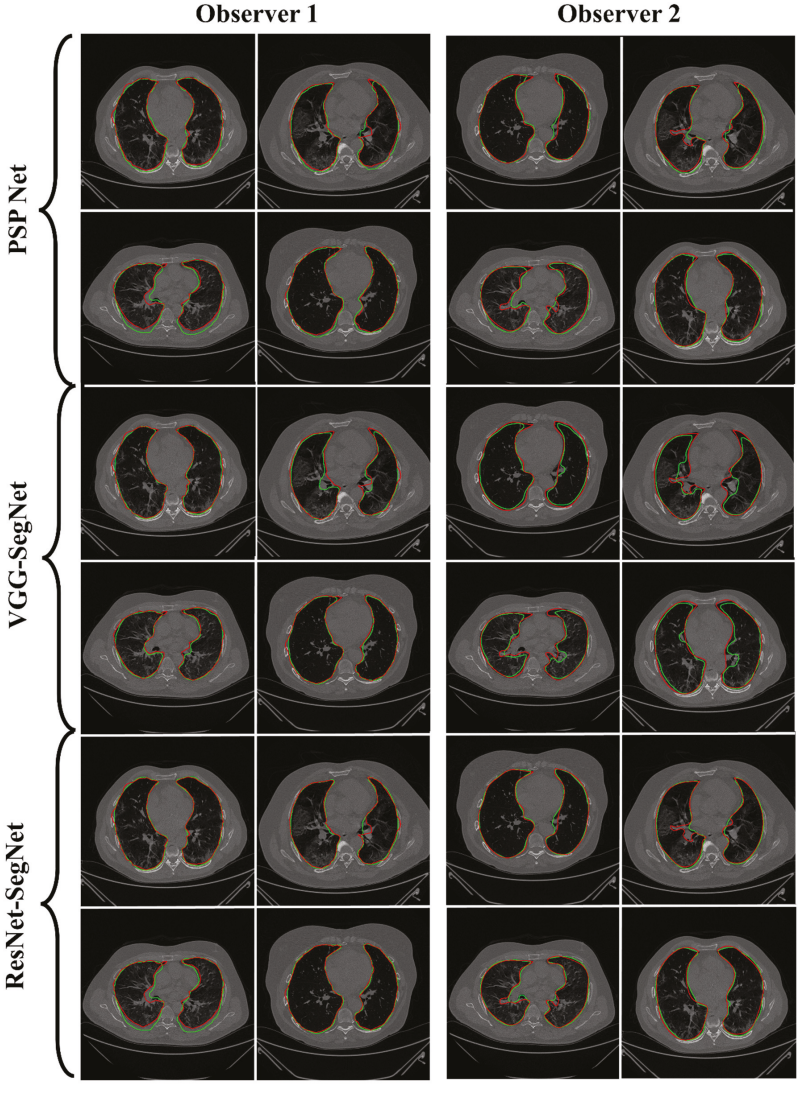


Figure 10. AI-model segmented boundary (green) vs. GT boundary (red) for Observer 1 and Observer 2.

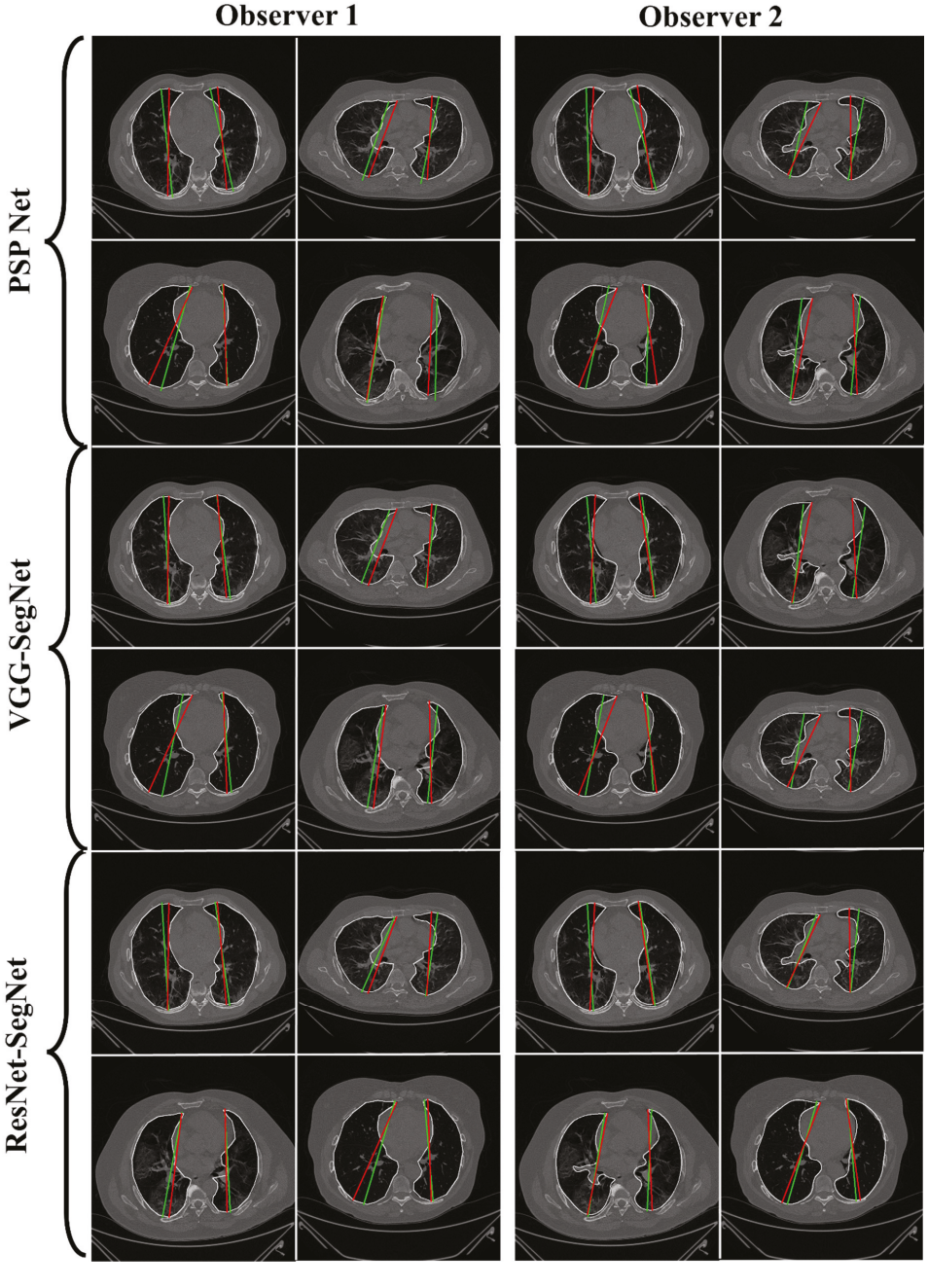


Figure 11. AI-model long axis (green) vs. GT long axis (red) for Observer 1 and Observer 2.

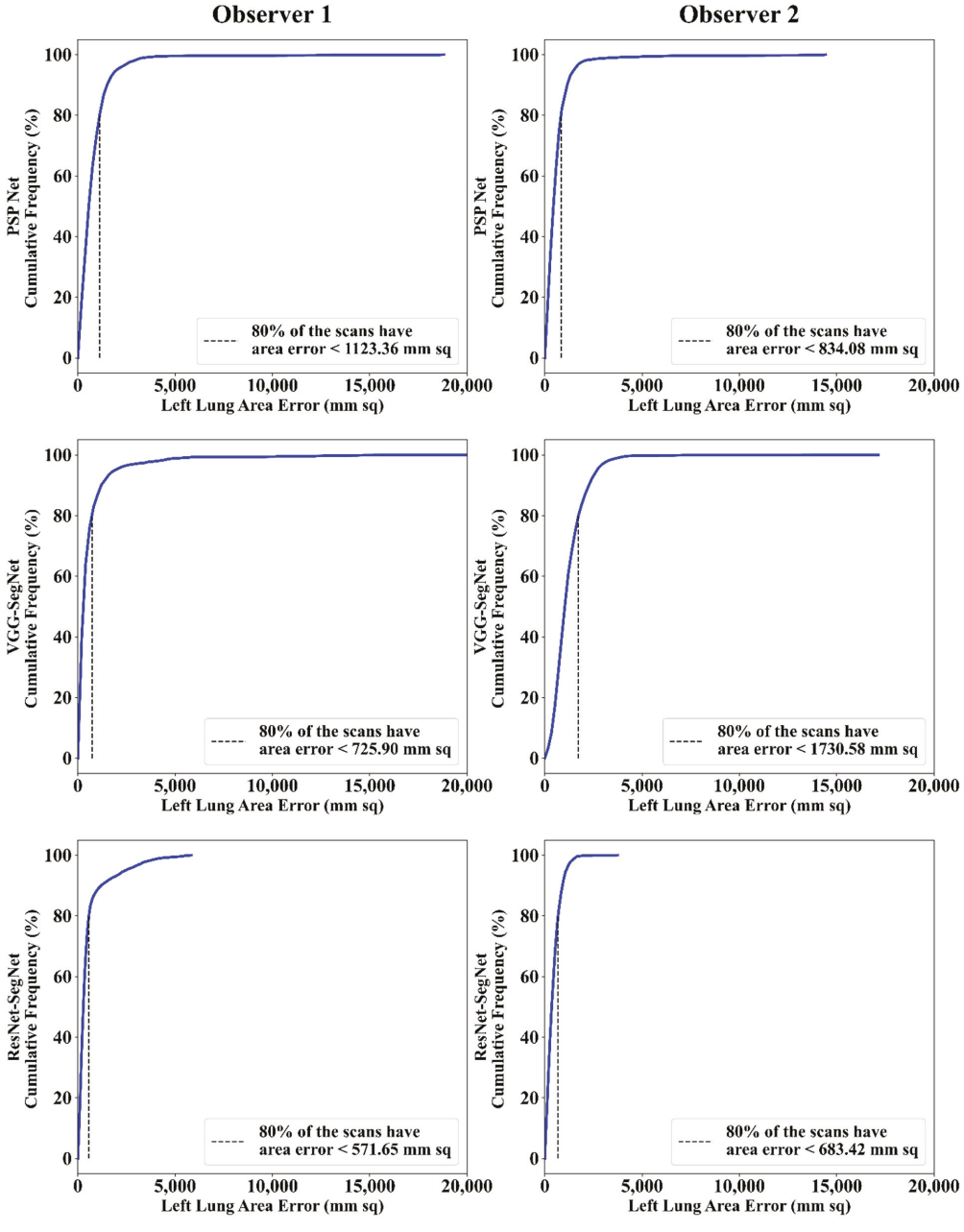


Figure 12. Cumulative frequency plot of left LAE using three AI models: Observer 1 vs. Observer 2.

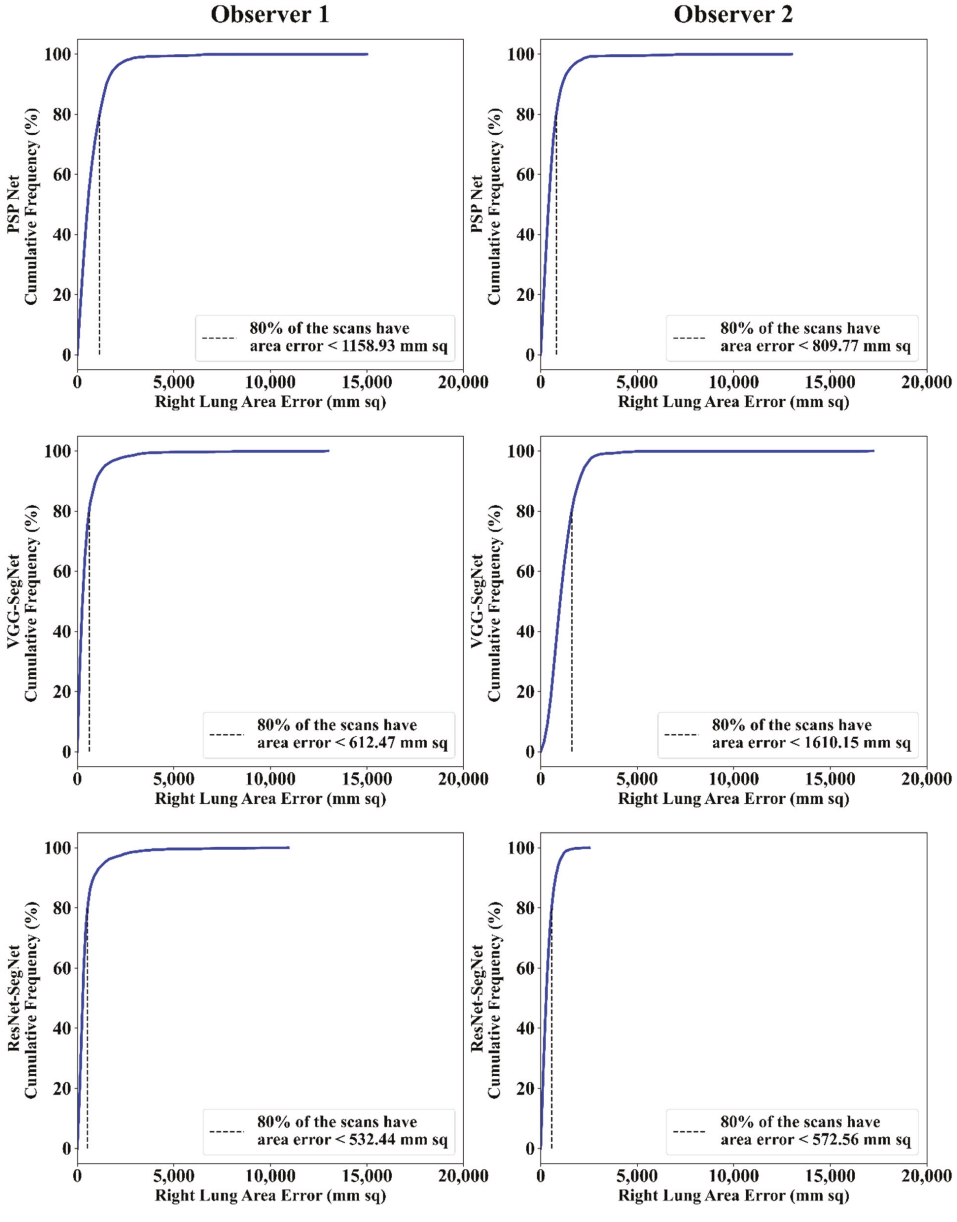


Figure 13. Cumulative frequency plot of right LAE using three AI models: Observer 1 vs. Observer 2.

Correlation Plot for Lung Area Error

Coefficient of correlations (CC) plots for the three AI models' LA vs. GT, area corresponding to the left and right between Observers 1 and 2, are shown in Figures 14 and 15. The CC values are summarized in Table 1 with a percentage difference between Observers 1 and 2. The percentage difference for the CC value ($p < 0.001$) ranges from 0% to 2.04%, which is <5% as part of the error threshold chosen as the hypothesis. This clearly shows that the AI models are clinically valid for the proposed setting of the inter-observer variability study.

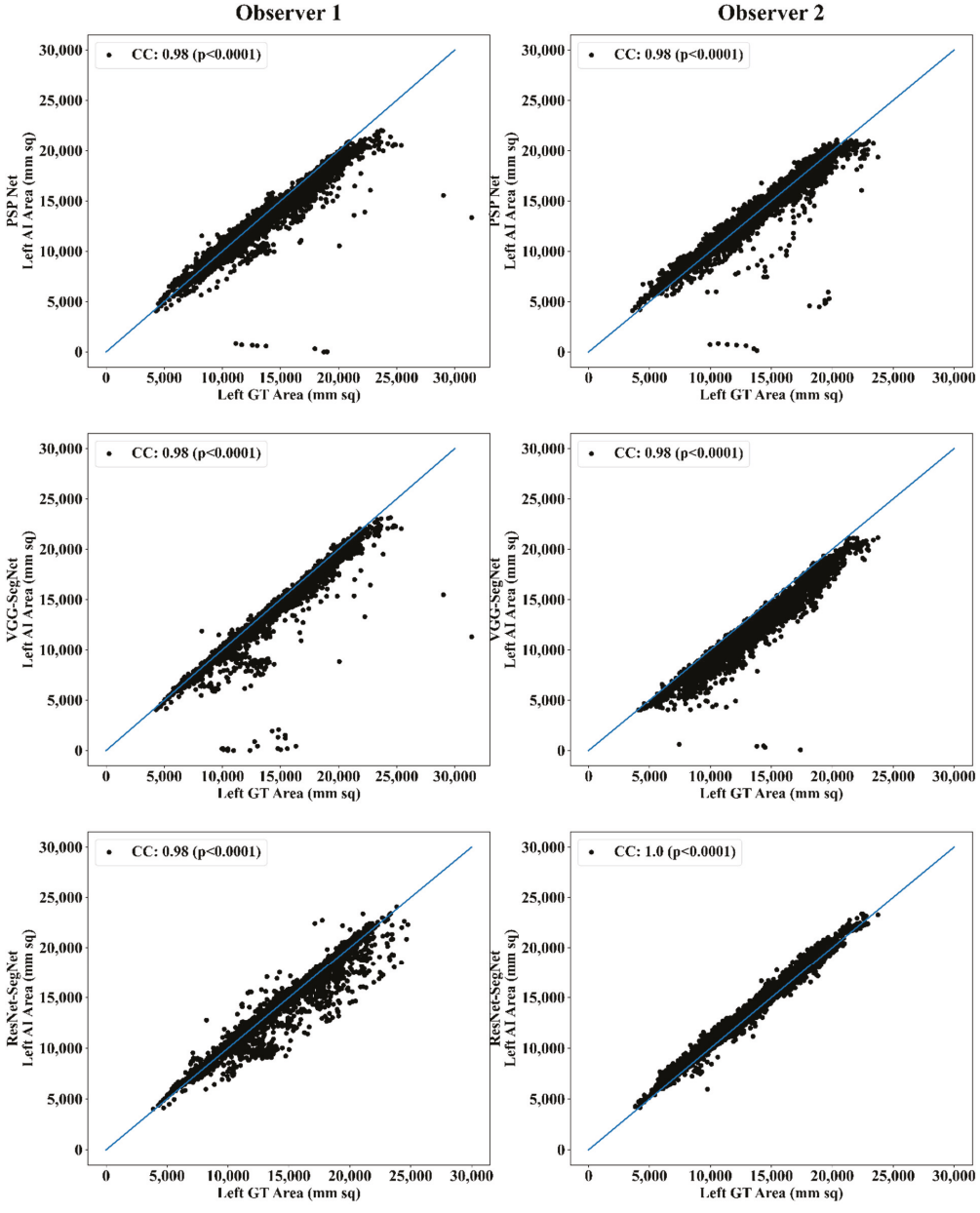


Figure 14. CC of left lung area using three AI models: Observer 1 vs. Observer 2.

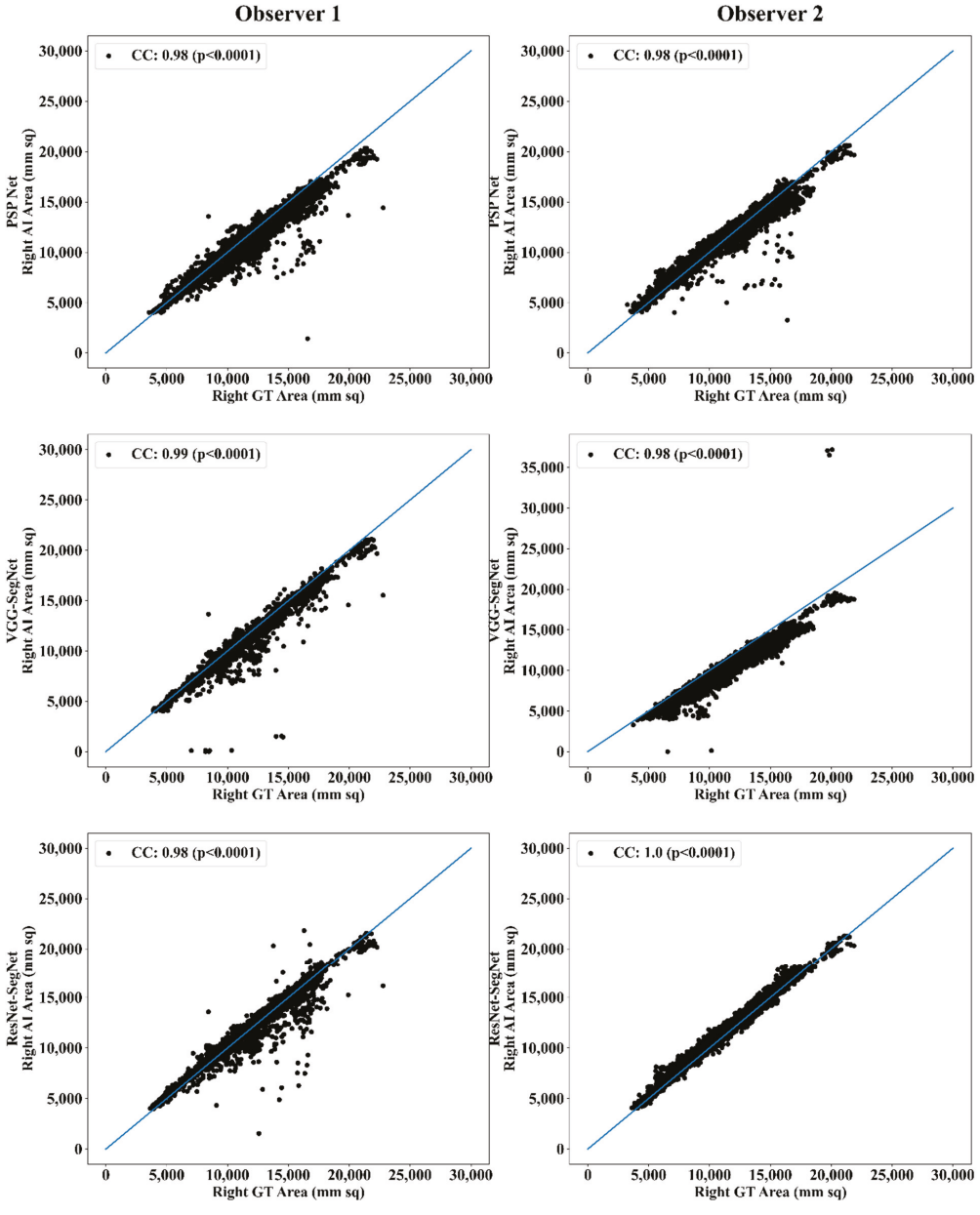


Figure 15. CC of right lung area using three AI models: Observer 1 vs. Observer 2.

Jaccard Index and Dice Similarity

Figure 16 depicts a cumulative frequency plot for dice similarity (DS) for three AI models between Observers 1 and Observer 2. It shows that 80% of the CT images had a DS > 0.95. A cumulative frequency plot for the Jaccard Index (JI) is presented in Figure 17

and shows that 80% of the CT scans had a JI > 0.90 between Observer 1 and Observer 2. The three AI models follow the order: PSP Net, VGG-SegNet, and ResNet-SegNet.

Table 1. Comparison of the CC values obtained between AI model area and the GT area corresponding to Observer 1 and Observer 2.

	PSP Net			VGG-SegNet			ResNet-SegNet		
	Left	**Right**	**Mean**	**Left**	**Right**	**Mean**	**Left**	**Right**	**Mean**
Observer 1	0.98	0.98	0.98	0.98	0.99	0.99	0.98	0.98	0.98
Observer 2	0.98	0.98	0.98	0.98	0.98	0.98	1.00	1.00	1.00
% Difference	0.00	0.00	0.00	0.00	1.01	0.51	2.04	2.04	2.04

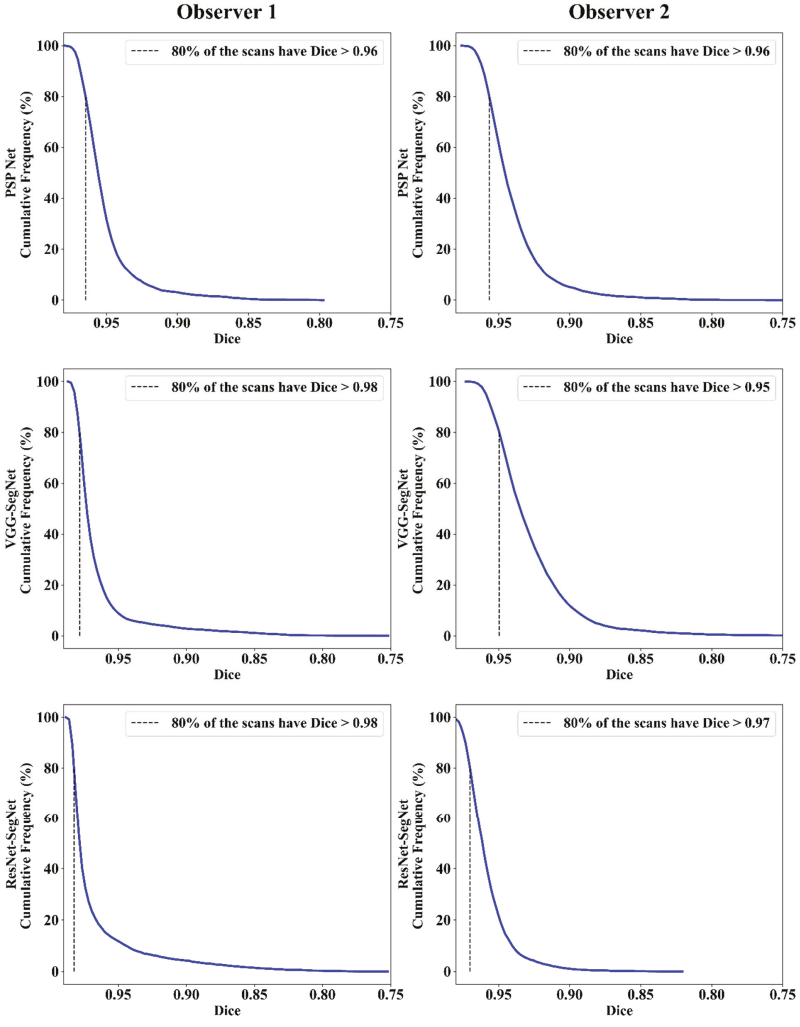


Figure 16. DS for combined lung using the three AI models: Observer 1 vs. Observer 2.

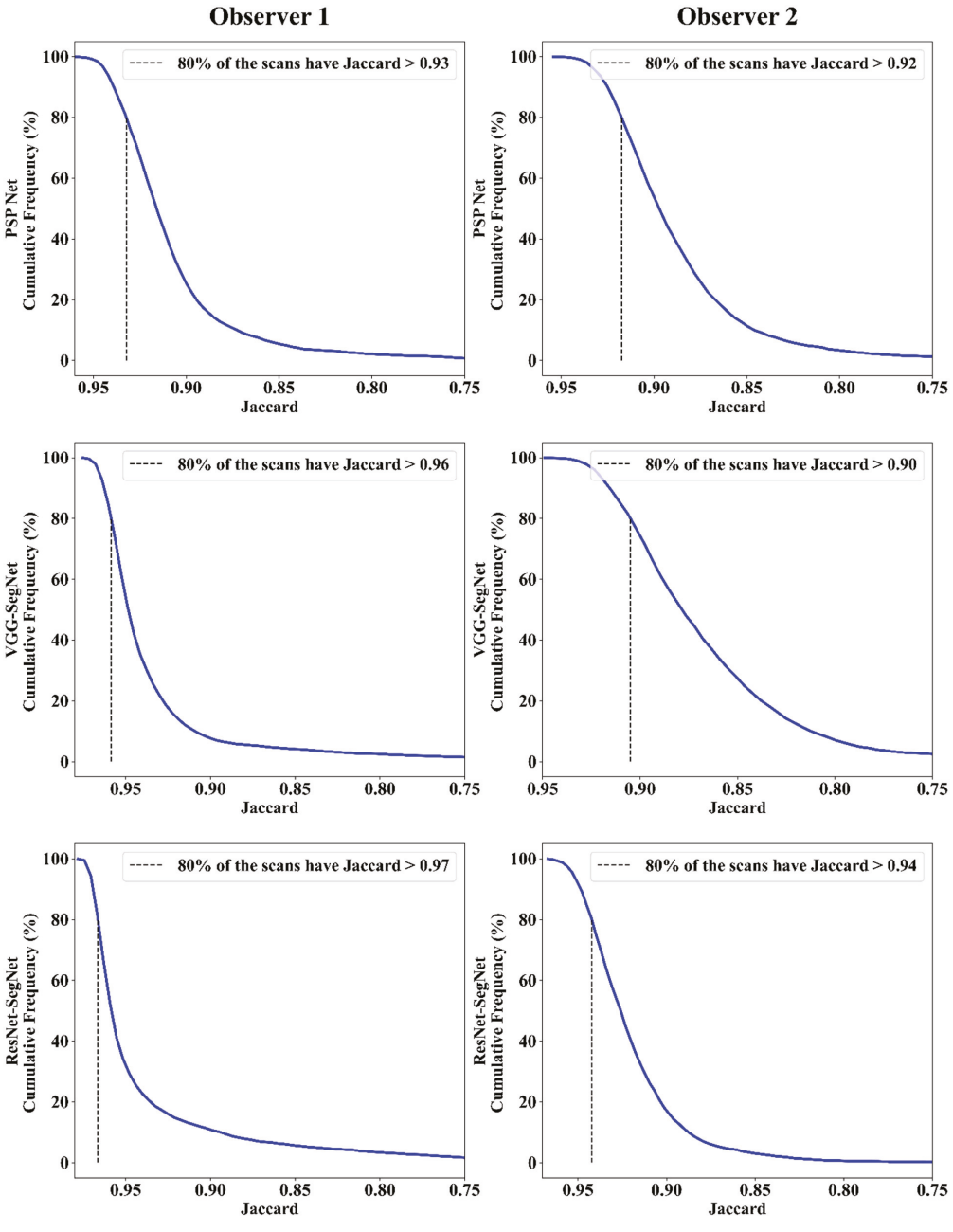


Figure 17. JI for combined lung using three AI models: Observer 1 vs. Observer 2.

Bland-Altman Plot for Lung Area

A Bland-Altman plot is used to demonstrate the consistency of two methods that employ the same variable. Based on our prior paradigms [48,62], we follow the Bland-Altman

computing procedure. Figures 18 and 19 show the (i) mean and (ii) standard deviation of the lung area between the AI model and GT area corresponding to Observers 1 and Observer 2.

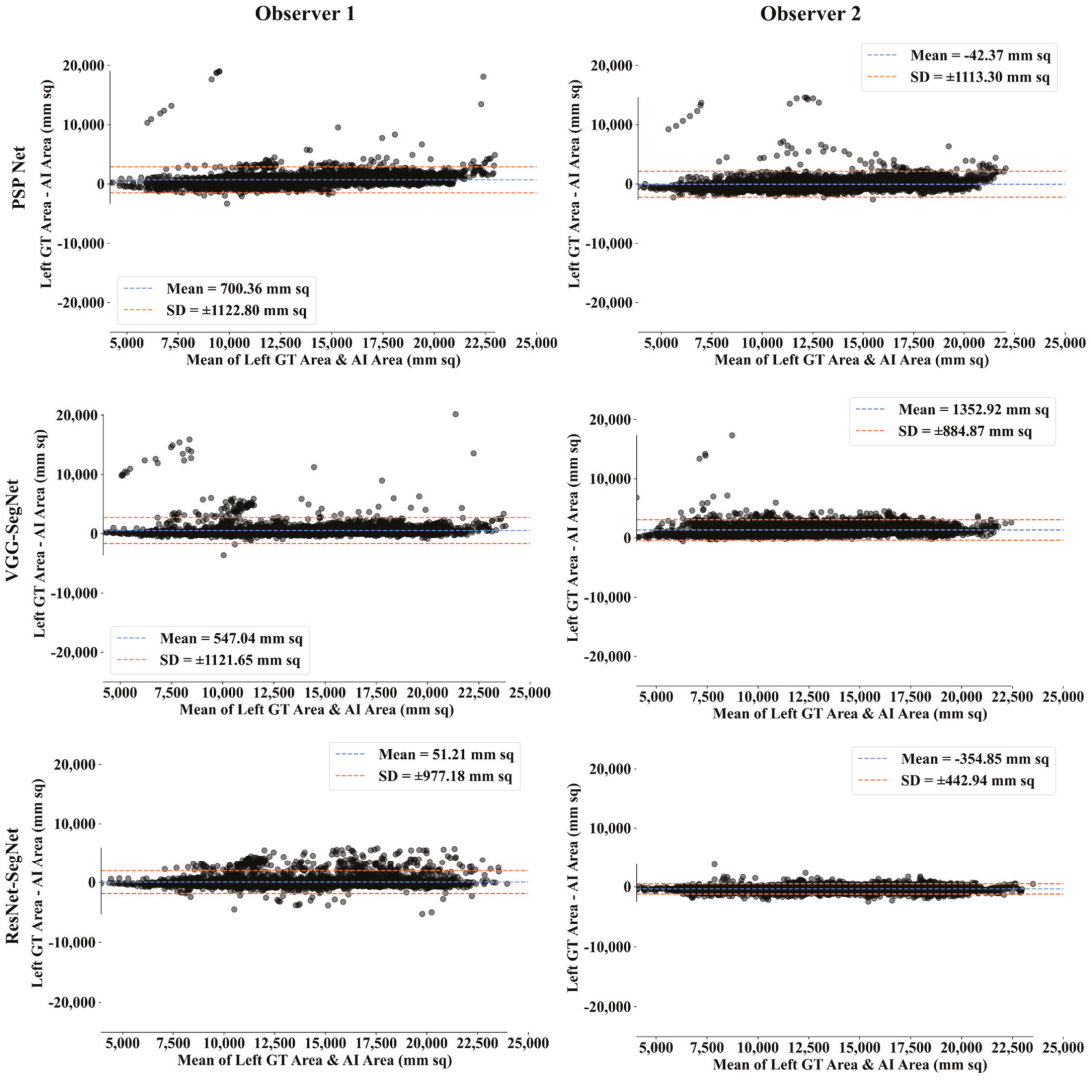


Figure 18. BA for left LA for three AI models: Observer 1 vs. Observer 2.

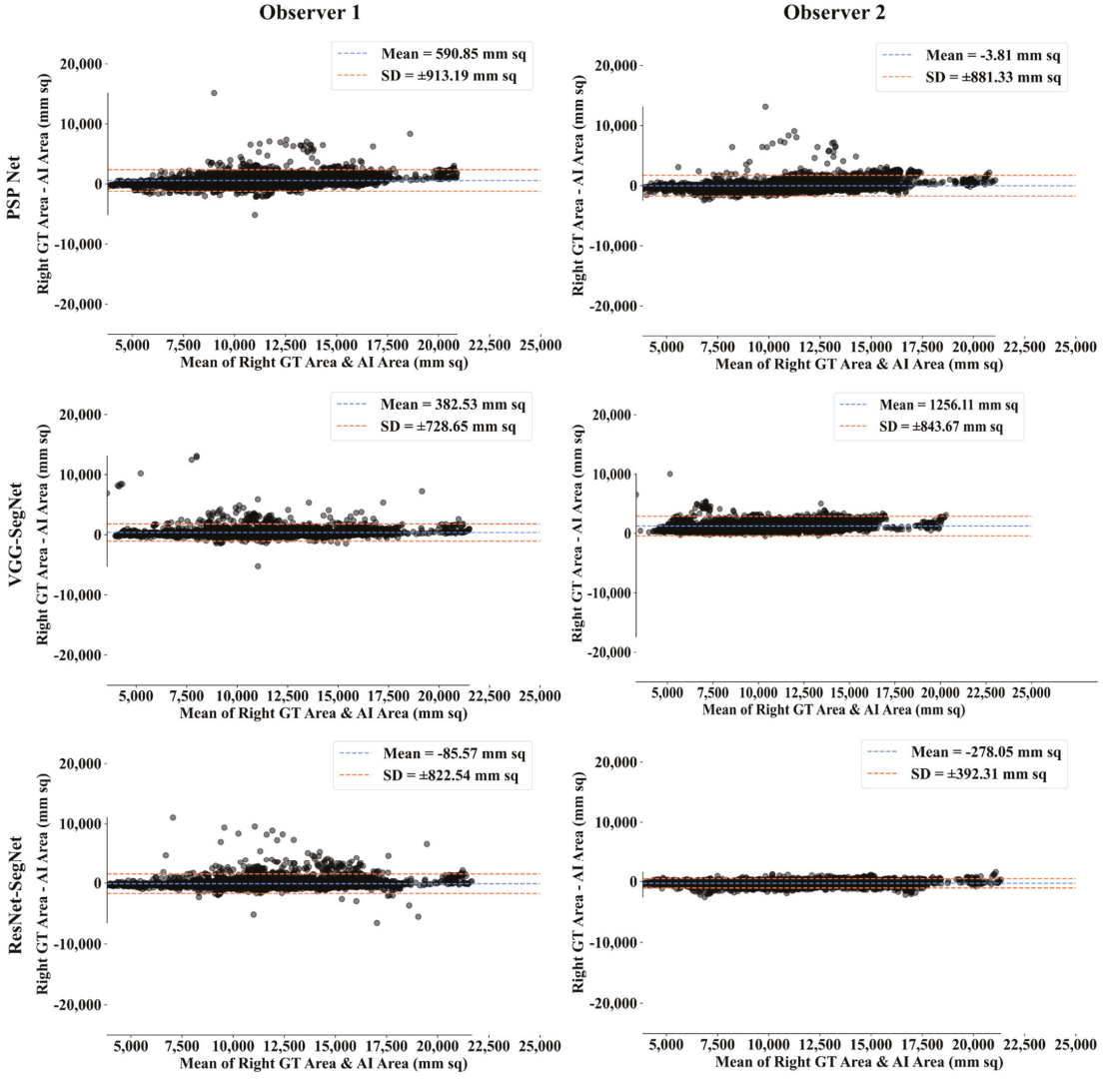


Figure 19. BA for right LA using three AI models: Observer 1 vs. Observer 2.

ROC Plots for Lung Area

An ROC curve represents how an AI system's diagnostic performance changes as the discrimination threshold changes. Figure 20 shows the ROC curve and corresponding AUC value for the three AI models between Observer 1 and Observer 2. The three AI models follow the order: PSP Net, VGG-SegNet, and ResNet-SegNet.

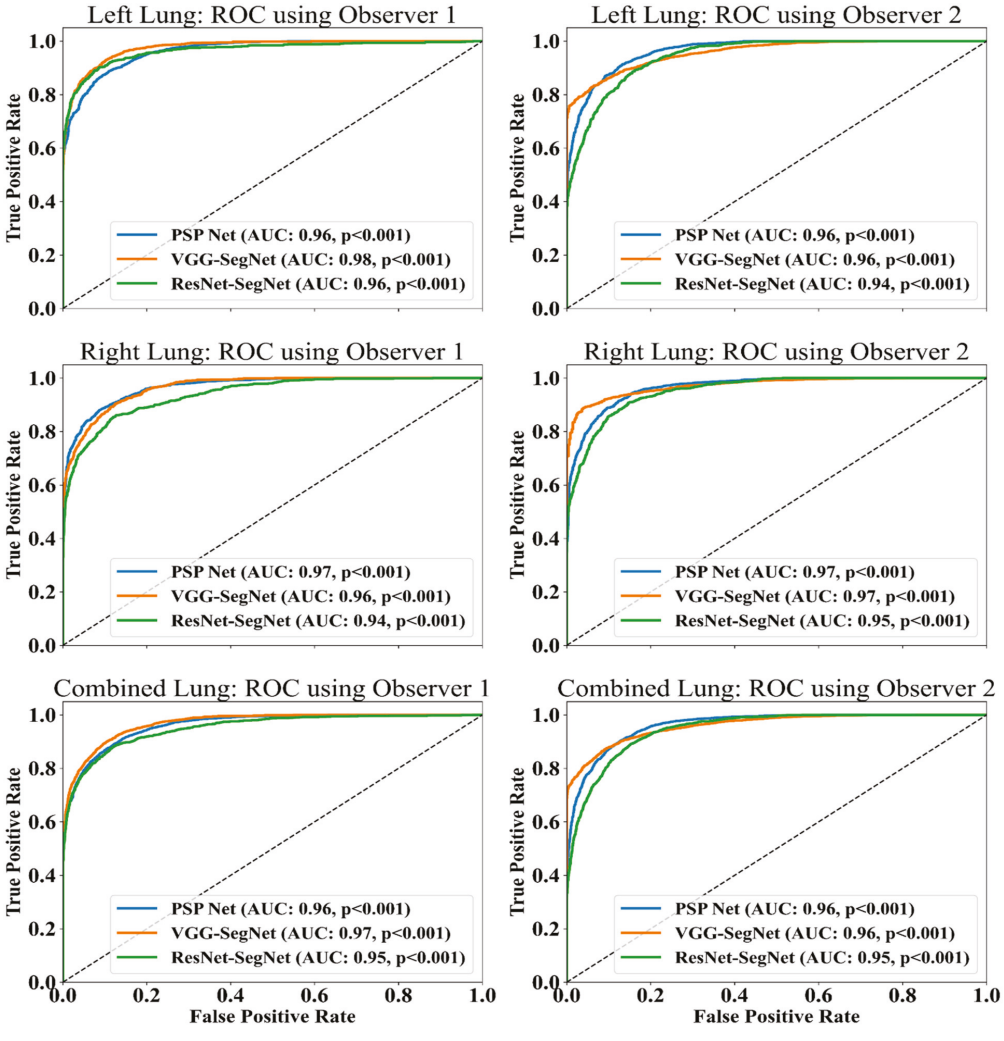


Figure 20. ROC and AUC curve for the three AI models: Observer 1 vs. Observer 2.

4.2.3. Performance Evaluation Using Lung Long Axis Error

Cumulative Frequency Plot for Lung Long Axis Error

Figures 21 and 22 show the cumulative frequency plot LLAE for left and right lung, respectively, corresponding to Observer 1 and Observer 2 for the three AI models. Based on the 80% threshold, the LLAE for the left lung (Figure 21) using the three AI models for Observer 1 and Observer 2 were 6.12 mm (for PSP Net), 4.77 mm (for VGG-SegNet), and 5.01 mm (for ResNet-SegNet) and 10.88 mm (for PSP Net), 13.30 mm (for VGG-SegNet), and 9.18 mm (for ResNet-SegNet), respectively. Similarly, for the right lung (Figure 22), the error was 7.81 mm (for PSP Net), 5.47 mm (for VGG-SegNet), and 3.10 mm (for ResNet-SegNet) and 9.14 mm (for PSP Net), 11.33 mm (for VGG-SegNet), and 6.88 mm (for ResNet-SegNet), respectively, for Observer 1 and Observer 2. The three AI models follow the order: PSP Net, VGG-SegNet, and ResNet-SegNet.

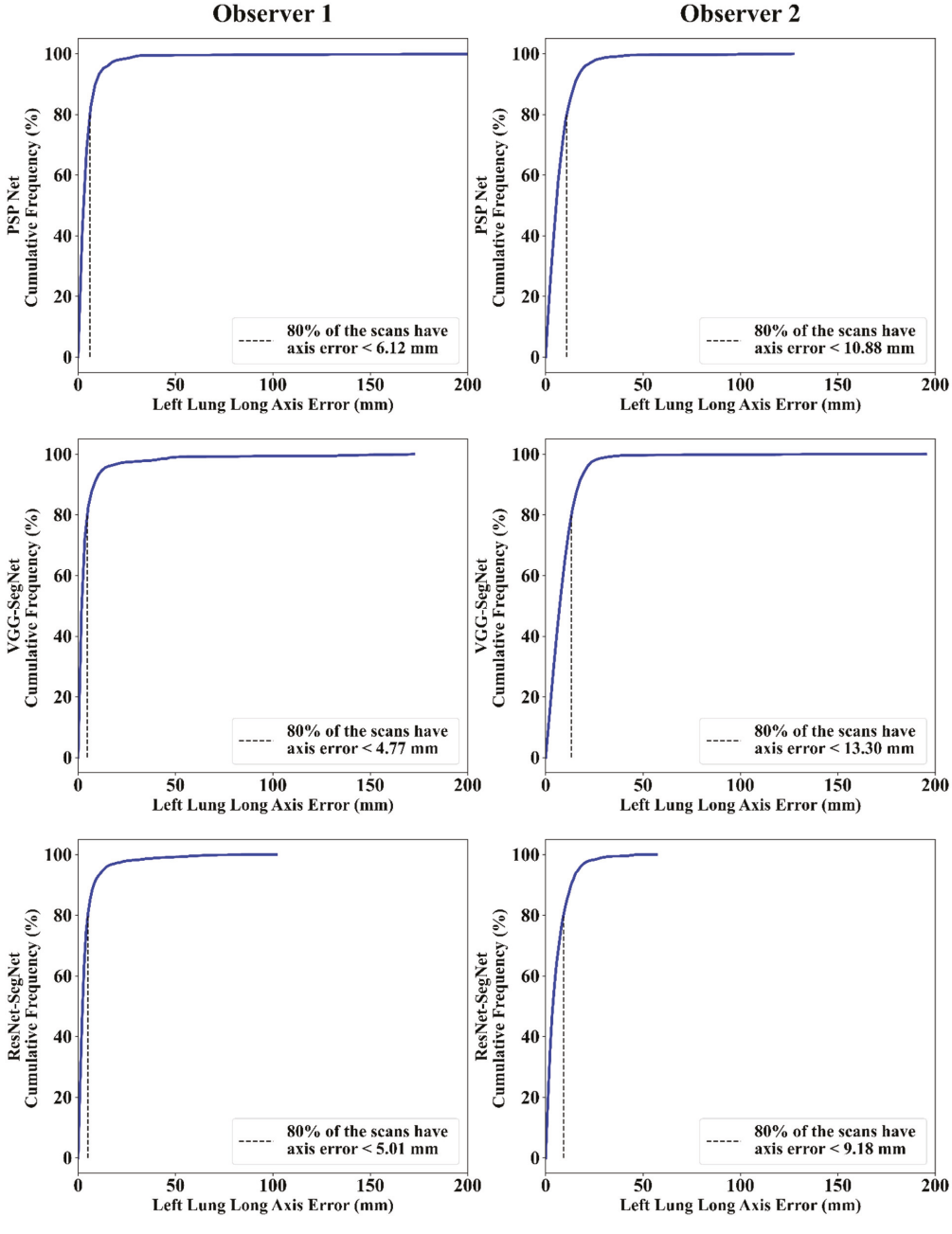


Figure 21. Cumulative frequency plot for left LLAE using three AI models: Observer 1 vs. Observer 2.

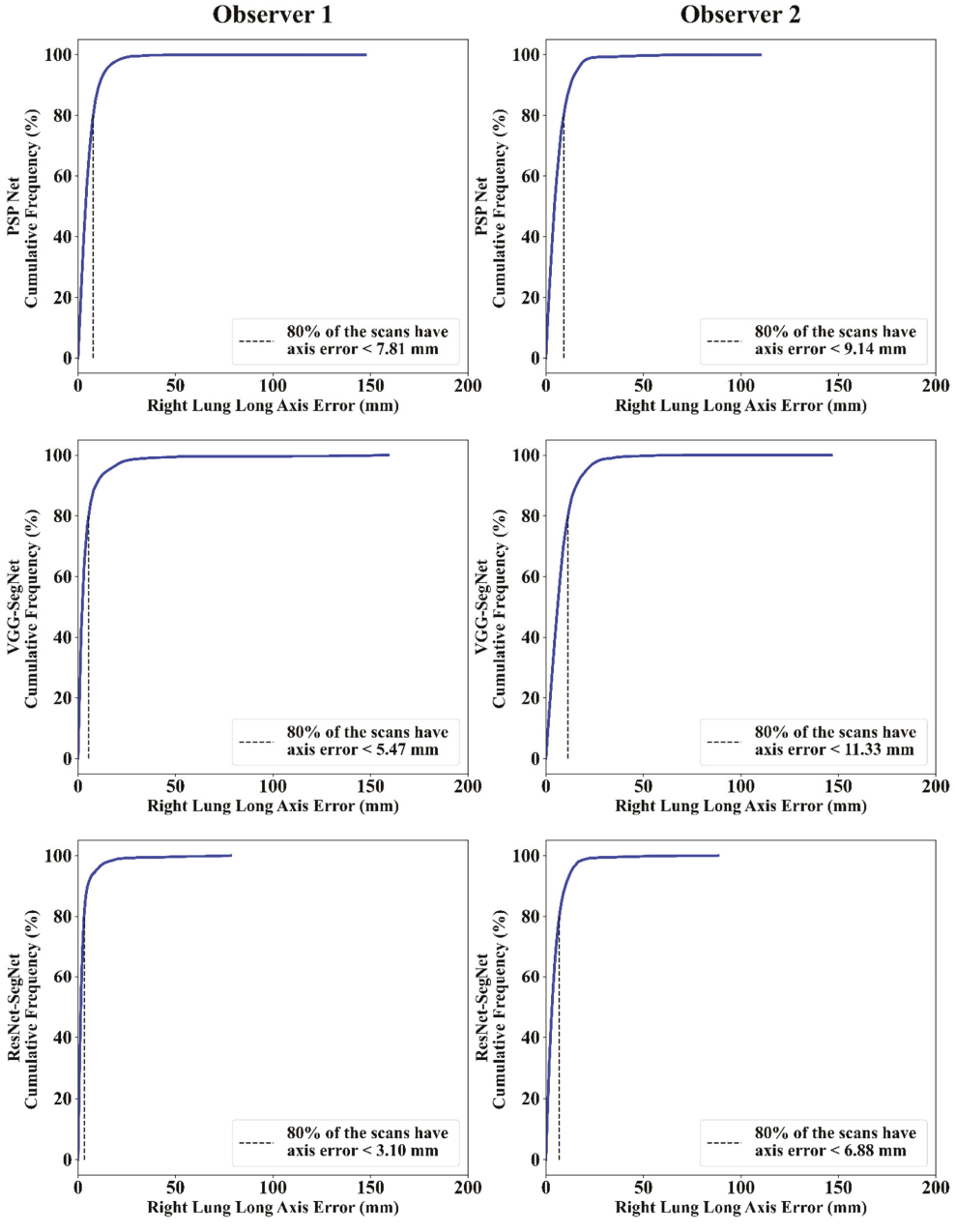


Figure 22. Cumulative frequency plot for right LLAE using three AI models: Observer 1 vs. Observer 2.

Correlation Plot for Lung Long Axis Error

Figures 23 and 24 show the CC plot for the three AI models considered in the proposed inter-observer variability study for Observers 1 and 2. Table 2 summarizes the CC values for the left, right, and mean errors of the LLA. It proves the hypothesis that the percentage

difference between the results using the two observers has a difference of <5%. This demonstrates that the proposed system is clinically valid in the suggested inter-observer variability study context.

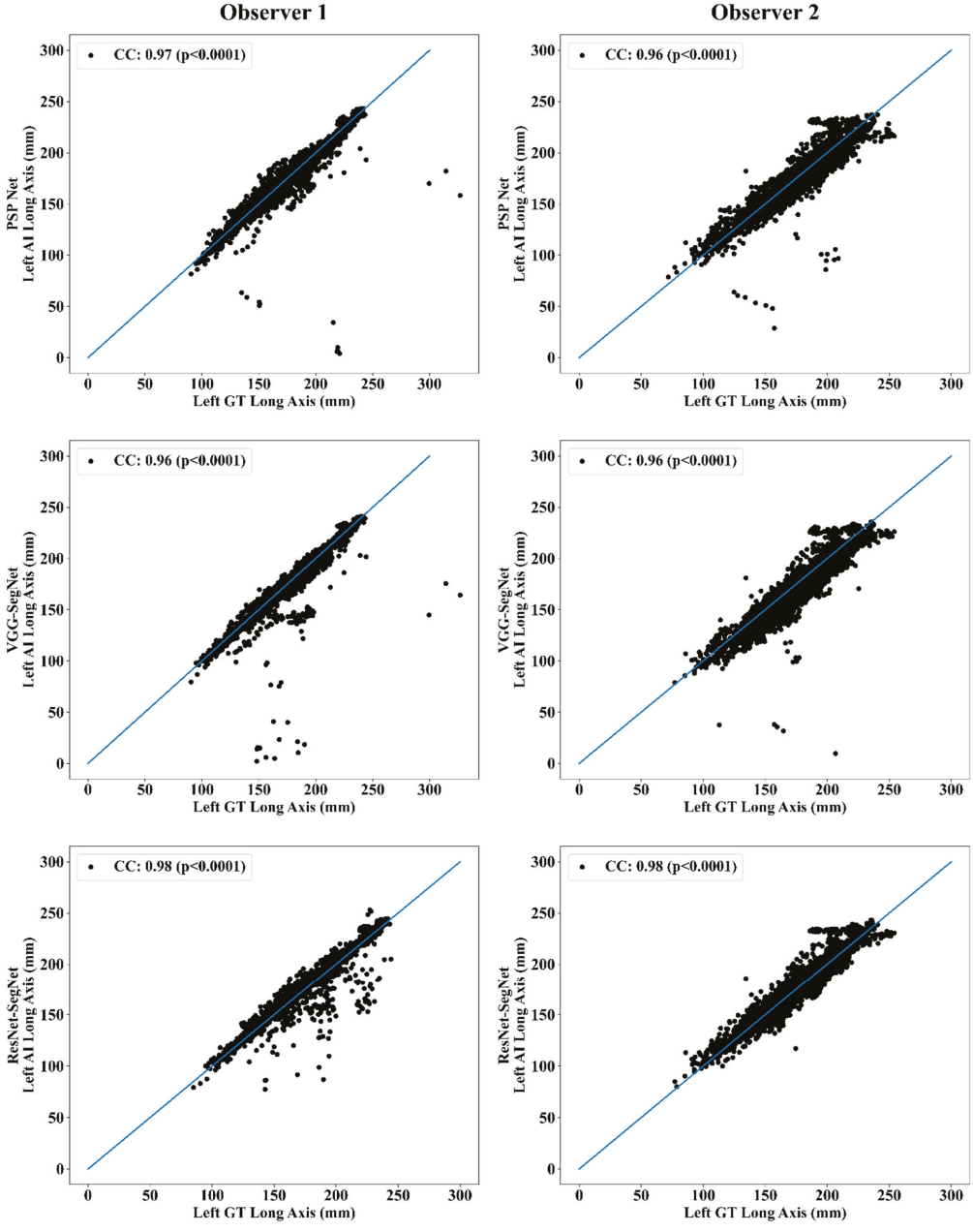


Figure 23. CC of left LLA for three AI models: Observer 1 vs. Observer 2.

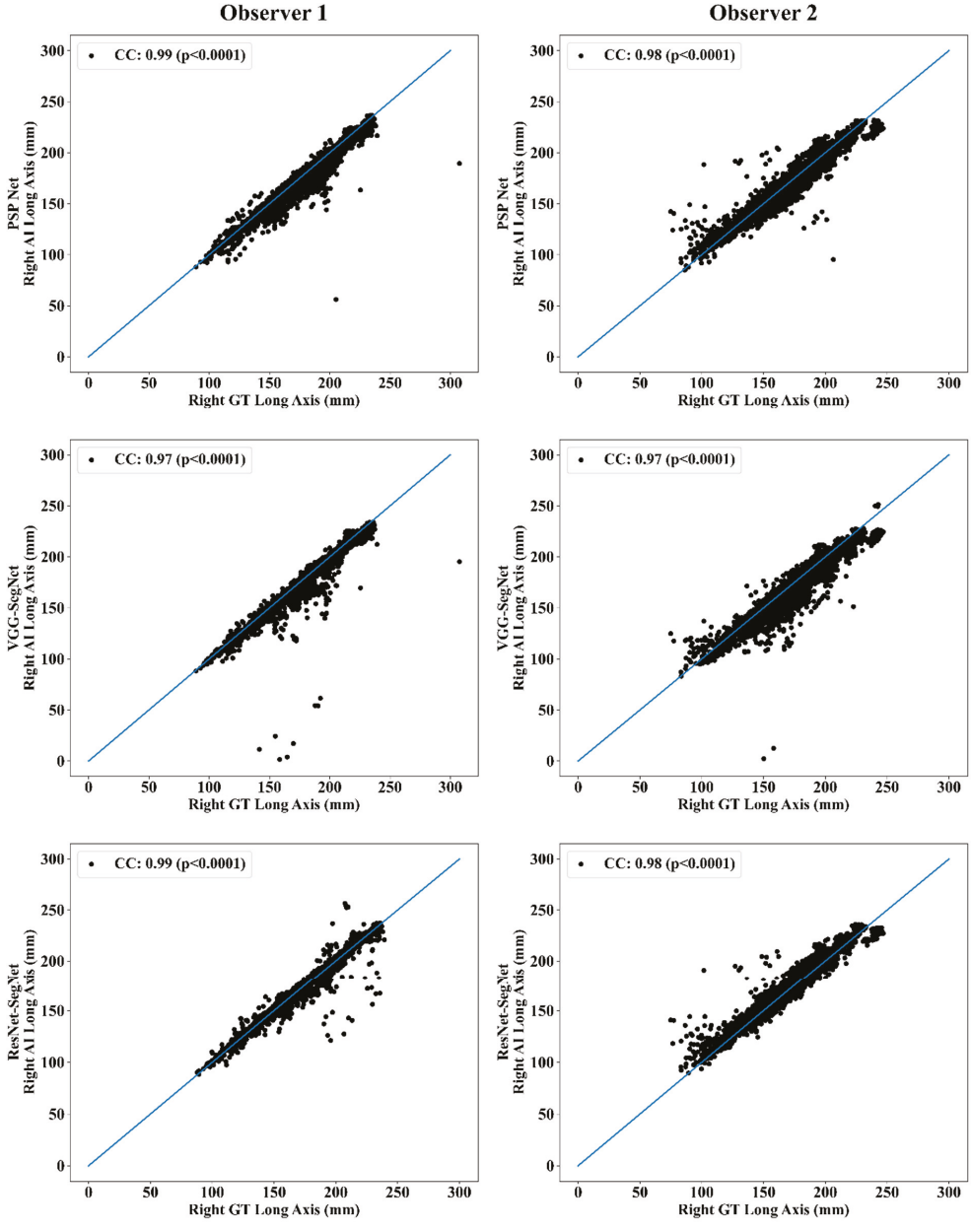


Figure 24. CC of right LLA using three AI models: Observer 1 vs. Observer 2.

Bland-Altman Plots for Lung Long Axis Error

The (i) mean and (ii) standard deviation of the lung long axis corresponding to Observer 1 and Observer 2 for the three AI models is shown in Figure 25 for the left lung and Figure 26 for the right lung.

Table 2. Comparison of the CC values obtained between AI model lung long axis and the GT lung long axis corresponding to Observer 1 and Observer 2.

	PSP Net			VGG-SegNet			ResNet-SegNet		
	Left	**Right**	**Mean**	**Left**	**Right**	**Mean**	**Left**	**Right**	**Mean**
Observer 1	0.97	0.99	0.98	0.96	0.97	0.97	0.98	0.99	0.99
Observer 2	0.96	0.98	0.97	0.96	0.97	0.97	0.98	0.98	0.98
% Difference	1.03	1.01	1.02	0.00	0.00	0.00	0.00	1.01	0.51

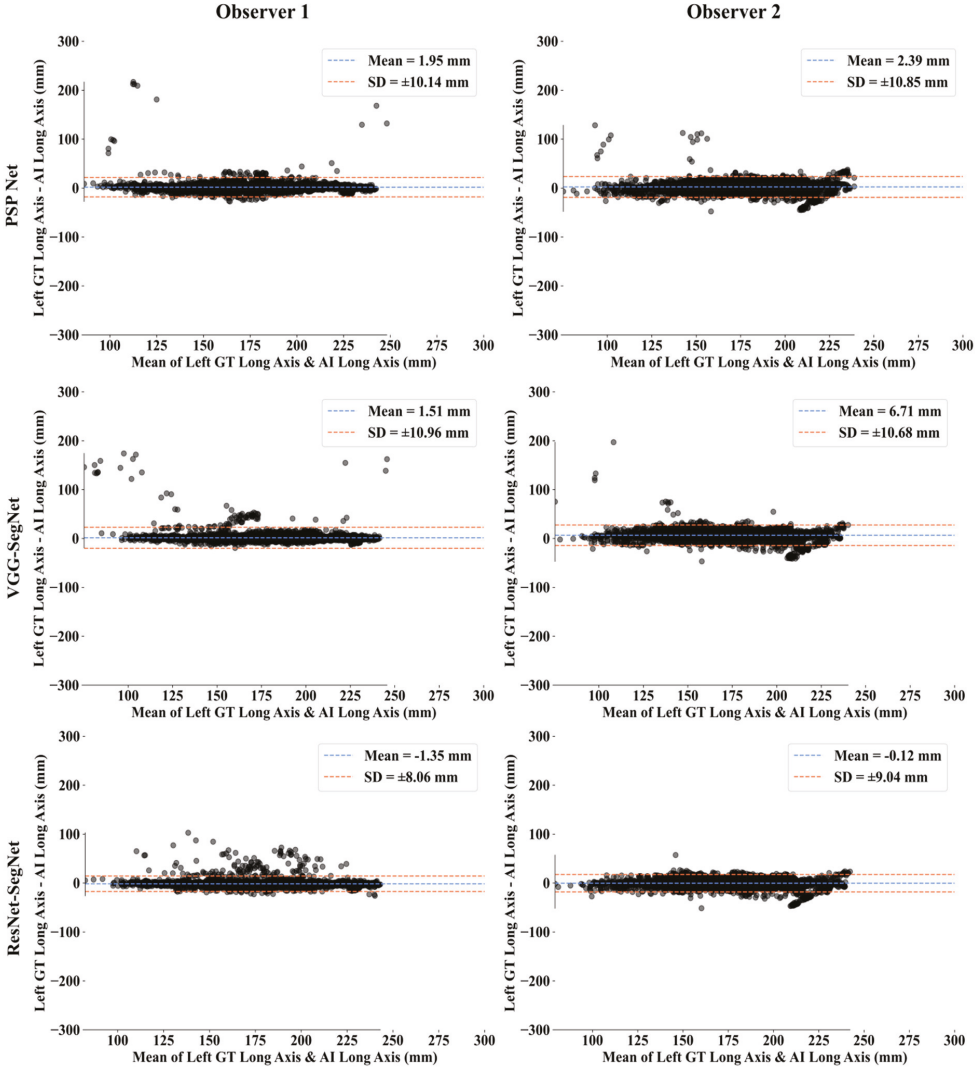


Figure 25. BA for the left LLA using the three: Observer 1 vs. Observer 2.

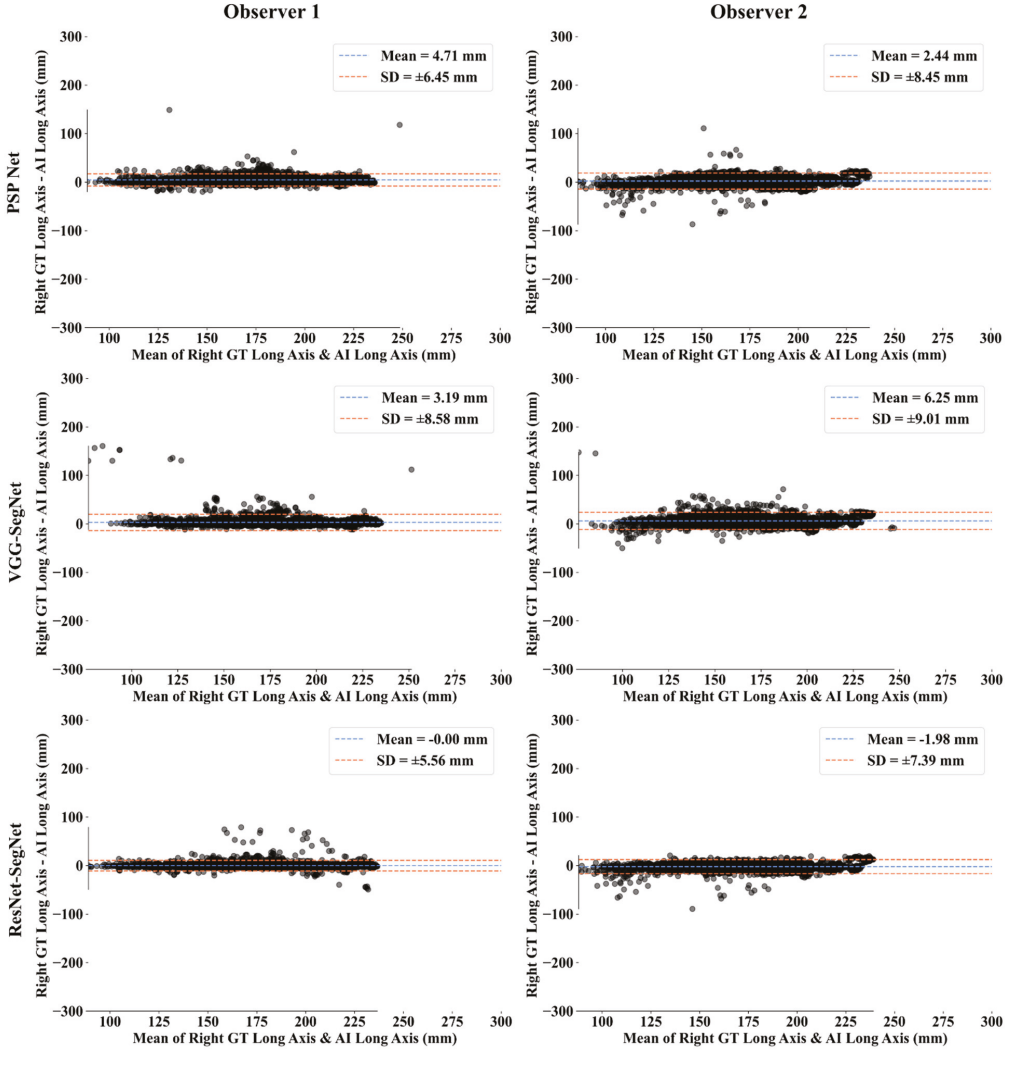


Figure 26. BA for the right LLA using the three AI models: Observer 1 vs. Observer 2.

Statistical Tests

The system's dependability and stability were assessed using a standard paired *t*-test, ANOVA, and Wilcoxon test. The paired *t*-test can be used to see if there is enough data to support a hypothesis; the Wilcoxon test is its alternative when the distribution is not normal. ANOVA helps in the analysis of the difference between the means of groups of the input data. MedCalc software (Osteen, Belgium) was used to perform the statistical analysis. To validate the system presented in this study, we have presented all the possible combinations (twelve in total) for the three AI models between Observer 1 and Observer 2. Table 3 shows the paired *t*-test, ANOVA, and Wilcoxon test results for the 12 combinations.

Table 3. Paired *t*-test, Wilcoxon, ANOVA, and CC for LA and LLA for the 12 combinations.

		Lung Area				Lung Long Axis			
SN	**Combinations**	**Paired *t*-Test (*p*-Value)**	**Wilcoxon (*p*-Value)**	**ANOVA (*p*-Value)**	**CC [0–1]**	**Paired *t*-Test (*p*-Value)**	**Wilcoxon (*p*-Value)**	**ANOVA (*p*-Value)**	**CC [0–1]**
1	P1 vs. V1	<0.0001	<0.0001	<0.001	0.9726	<0.0001	<0.0001	<0.001	0.9509
2	P1 vs. R1	<0.0001	<0.0001	<0.001	0.9514	<0.0001	<0.0001	<0.001	0.9506
3	P1 vs. P2	<0.0001	<0.0001	<0.001	0.9703	<0.0001	<0.0001	<0.001	0.9686
4	P1 vs. V2	<0.0001	<0.0001	<0.001	0.9446	<0.0001	<0.0001	<0.001	0.9445
5	P1 vs. R2	<0.0001	<0.0001	<0.001	0.9764	<0.0001	<0.0001	<0.001	0.9661
6	V1 vs. R1	<0.0001	<0.0001	<0.001	0.9663	<0.0001	<0.0001	<0.001	0.9561
7	V1 vs. P2	<0.0001	<0.0001	<0.001	0.9726	<0.0001	<0.0001	<0.001	0.9671
8	V1 vs. V2	<0.0001	<0.0001	<0.001	0.9766	<0.0001	<0.0001	<0.001	0.9638
9	V1 vs. R2	<0.0001	<0.0001	<0.001	0.9943	<0.0001	<0.0001	<0.001	0.9796
10	R1 vs. P2	<0.0001	<0.0001	<0.001	0.9549	<0.0001	<0.0001	<0.001	0.9617
11	R1 vs. V2	<0.0001	<0.0001	<0.001	0.9513	<0.0001	<0.0001	<0.001	0.9499
12	R1 vs. R2	<0.0001	<0.0001	<0.001	0.9690	<0.0001	<0.0001	<0.001	0.9726

CC: Correlation coefficient; P1: PSP Net for Observer 1; V1: VGG-SegNet for Observer 1; R1: ResNet-SegNet for Observer 1; P2: PSP Net for Observer 2; V2: VGG-SegNet for Observer 2; R2: ResNet-SegNet for Observer 2.

Figure of Merit

The likelihood of the error in the system is known as the figure of merit (FoM). We have calculated FoM for (i) lung area and (ii) lung long axis to show the acceptability of the hypothesis if the % difference between the two observers is <5%. Table 4 shows the values for FoM using Equation (5) and the % difference for the three AI models against the two observers. Similarly, Table 5 shows the values for FoM using Equation (6) and the % difference for the three AI models against the two observers.

$$FoM_A(m) = 100 - \left[\left(\frac{|\overline{A}_{ai}(m) - \overline{A}_{gt}|}{\overline{A}_{gt}}\right) \times 100\right], \tag{5}$$

$$FoM_{LA}(m) = 100 - \left[\left(\frac{|\overline{L}_{ai}(m) - \overline{L}_{gt}|}{\overline{L}_{gt}}\right) \times 100\right]$$

$$\text{where } \overline{A}_{ai}(m) = \frac{\sum_{n=1}^{N} A_{ai}(m,n)}{N},\ \overline{A}_{gt} = \frac{\sum_{n=1}^{N} A_{gt}(n)}{N}, \tag{6}$$

$$\overline{LA}_{ai}(m) = \frac{\sum_{n=1}^{N} LA_{ai}(m,n)}{N}\ \&\ \overline{LA}_{gt} = \frac{\sum_{n=1}^{N} LA_{gt}(n)}{N}$$

Table 4. FoM for lung area.

	Observer 1			Observer 2			% Difference			Hypothesis (<5%)		
	Left	**Right**	**Mean**	**Left**	**Right**	**Mean**	**Left**	**Right**	**Mean**	**Left**	**Right**	**Mean**
PSP Net	95.07	95.11	95.09	97.37	97.49	97.43	2%	3%	2%	✔	✔	✔
VGG-SegNet	96.73	97.40	97.04	97.74	97.27	97.52	1%	0%	0%	✔	✔	✔
ResNet-SegNet	98.33	99.98	99.11	97.88	99.20	98.50	0%	1%	1%	✔	✔	✔

Table 5. FoM for lung long axis.

	Observer 1			Observer 2			% Difference			Hypothesis (<5%)		
	Left	**Right**	**Mean**	**Left**	**Right**	**Mean**	**Left**	**Right**	**Mean**	**Left**	**Right**	**Mean**
PSP Net	98.91	97.34	98.13	98.65	98.60	98.62	0%	1%	1%	✔	✔	✔
VGG-SegNet	99.41	98.50	98.95	97.07	97.27	97.17	2%	1%	2%	✔	✔	✔
ResNet-SegNet	99.73	99.37	99.83	99.51	98.75	99.13	0%	1%	1%	✔	✔	✔

5. Discussion

The study presented the inter-observer variability analysis for the COVLIAS 1.0 using three AI models, PSP Net, VGG-SegNet, and ResNet-SegNet. These models have considered tissue characterization approaches since they analyze the tissue data for better feature extraction to evaluate for ground vs. background, thus are more akin to a tissue characterization in classification framework [30,37]. Our group has strong experience in tissue characterization approaches with different AI models and applications for classification using ML frameworks such as plaque, liver, thyroid, breast [21,28,30,63–68], and DL framework [1,36,69,70]. These three AI models were trained using the GT annotated data from the two observers. The percentage difference between the outputs of the two AI model results was less than 5%, and thus the hypothesis was confirmed. During the training, the K5 cross-validation protocol was adapted on a set of 5000 CT images. For the PE of the proposed inter-observer variability system, the following ten metrics were considered: (i) visualization of the lung boundary, (ii) visualization of the lung long axis, cumulative frequency plots for (iii) LAE, (iv) LLAE, CC plots for (v) lung area, (vi) lung long axis, BA plots for (vii) lung area, (viii) lung long axis, (ix) ROC and AUC curve, and (x) JI and DS for estimated AI model lung regions. These matrices showed consistent and stable results. The training, evaluation, and quantification were implemented on the GPU environment (DGX V100) using python. We adapted vectorization provided by python during the implementation of the Numba library.

5.1. A Special Note on Three Model Behaviors with Respect to the Two OBSERVERS

The proposed inter-observer variability study used three AI models for the analysis, where PSP Net was implemented for the first time for COVID-19 lung segmentation. The other models VGG-SegNet and ResNet-SegNet were used for benchmarking. The AUC for the mean lung region for the three AI models was >0.95 for both Observer 1 and Observer 2.

Our results, shown below in Table 6, compared various metrics that included the inter-observer variability study for the three AI models. All the models behaved consistently while using the two different observers. Our results showed that ResNet-SegNet was the best performing model for all the PE metrics. The percentage difference between the two observers was 0.4%, 3.7%, and 0.4%, respectively, for the three models PSP Net, VGG-SegNet, and ResNet-SegNet, respectively. This further validated our hypothesis for every AI model, keeping the error threshold less than 5%. Even though all three AI models passed the hypothesis, VGG-SegNet is the least superior. This is because the number of the layers in the VGG-SegNet architecture (Figure 5) is 19, compared to ~50 in PSP Net (Figure 4) and 51 (encoder part) in the ResNet-SegNet model (Figure 6). By taking the results from both the observers into account, the order of the performance of the models is ResNet-SegNet > PSP Net > VGG-SegNet. Further, we also conclude that HDL models are superior to SDL (PSP Net). The aggregate score was computed as the mean for all the models for Observer 1, Observer 2, and the mean of the two Observers. Even though the performance of all the models was comparable, when carefully looking at the performance of Observer 1 the order of performance was ResNet-SegNet > VGG-SegNet > PSP Net. For Observer 2, the order of performance was ResNet-SegNet > PSP Net > VGG-SegNet.

Further, the performance of the left lung was better than the right lung for the reasons unclear at this point, and more investigations would be needed to evaluate this.

Table 6. Comparison of PE metrics for Observer 1 and Observer 2 and their mean.

Attributes	Observer 1			Observer 2			Mean Obs. 1 & Obs. 2		
	PSP Net	VGG-SegNet	ResNet-SegNet	PSP Net	VGG-SegNet	ResNet-SegNet	PSP Net	VGG-SegNet	ResNet-SegNet
DS	0.96	0.98	0.98	0.96	0.95	0.97	0.96	0.97	0.98
JI	0.93	0.96	0.97	0.92	0.9	0.94	0.93	0.93	0.96
CC Left LA	0.98	0.98	0.98	0.98	0.98	1	0.98	0.98	0.99
CC Right LA	0.98	0.99	0.98	0.98	0.98	1	0.98	0.99	0.99
CC Left LLA	0.97	0.96	0.98	0.96	0.96	0.98	0.97	0.96	0.98
CC Right LLA	0.99	0.97	0.99	0.98	0.97	0.98	0.99	0.97	0.99
CF Left LA < 10%	0.83	0.85	0.90	0.81	0.75	0.89	0.82	0.80	0.89
CF Right LA < 10%	0.78	0.85	0.90	0.80	0.75	0.88	0.79	0.80	0.89
Aggregate Score	**7.42**	**7.54**	**7.67**	**7.39**	**7.24**	**7.64**	**7.40**	**7.39**	**7.66**

DS: Dice similarity; JI: Jaccard index; CC: Correlation coefficient; LA: Lung area; LLA: Lung long axis; CF: Cumulative frequency; Obs: Observer.

5.2. Benchmarking

There have been several studies in the area of DL for lung segmentation, but only a few in the region of COVID-19 [71–74], and even less that involved variability analysis. Table 7 shows the benchmarking table having three variability studies: Saba et al. [48], Jeremy et al. [75], and Joskowicz et al. [76], that are compared against Suri et al. in this proposed study. Saba et al. has used a dataset of 96 patients with three observers for tracings, and ROC curves were also not presented in the study. Jeremy et al. [60] have demonstrated the variability analysis using five different observers that used the area error as the metric. The boundary error, ROC, JI, and DS were not discussed. Finally, Joskowicz et al. [76] used 480 images and 11 observers to annotate the dataset, but no area and boundary errors were present. Moreover, they did not present the ROC curves, JI, and DS for the tracings. All three studies [48,75,76], only performed manual annotation of the non-COVID dataset, and there was no involvement of the AI techniques to generate the boundaries automatically. Comparatively, the proposed study provides a first-of-its-kind for inter-observer variability analysis alongside HDL and SDL solutions, supporting our hypothesis that the error between the AI models trained using the two observers involved is less than 5%.

5.3. Strengths, Weakness, and Extensions

The proposed study successfully validated the hypothesis for the inter-observer variability settings, demonstrating that the difference between the two AI models when trained by the two observers was less than 5%. It was the first-time inter-observer variability was presented for COVID-19 lung segmentation using HDL and SDL models.

In spite of encouraging results, the study could not include more than two observers due to reasons such as cost, time, and availability of the radiologists. The imaging analysis component could be extended to handle more dense pulmonary opacities such as consolidation or mixed opacities during lung segmentation.

As part of the extension, the HDL models can be extended, which combines DL with ML or two solo DL models for lung segmentation. Conventional methods [77,78] can be used for lung segmentation embedded with denoising methods [79] and benchmarked against the AI models. The system can be extended to unseen data where the training data is taken from one clinical site and testing data can be from the other clinical site. It would

also be interesting to explore the segmentation of lungs in the healthy patients using the AI model trained on COVID-19 patients. Other neural network techniques such as generative adversarial networks (GANs) [80] or transfer learning and loss schemes [38,44,81] can also be adapted. A big data framework can be used to integrate comorbidity factors [82] in the AI models.

Table 7. Benchmarking Table.

Attributes/Author	Saba et al. [49]	Jeremy et al. [77]	Joskowicz et al. [78]	Suri et al. (Proposed)
# of patients	96	33	18	72
# of Images	NA	NA	490	5000
# of Observers	3	5	11	2
Dataset	Non-COVID	Non-COVID	Non-COVID	COVID
Image Size	512	NA	512	768
# of tests/PE	5	0	2	13
CC	0.98	NA	NA	0.98
Boundary estimation	Manual	Manual	Manual	Manual & automatic
AI Models	NA	NA	NA	3
Modality	CT	CT	CT	CT
Area Error	✔	✔	✗	✔
Boundary Error	✔	✗	✗	✔
ROC	✗	✗	✗	✔
JI	✔	✗	✗	✔
DS	✔	✗	✗	✔

CC: Correlation coefficient; ROC: Receiver-Operating Characteristics; DS: Dice similarity; JI: Jaccard index.

6. Conclusions

The proposed study is the first of its kind to evaluate the effect of ground-truth tracings on the AI models for COVID-19 CT lung segmentation. Three kinds of AI models, PSP Net, VGG-SegNet, and ResNet-SegNet, were adapted for lung segmentation. Two different Observers were used to annotate 5000 CT lung slices taken from 72 COVID-19 patients. Thus, six AI training models (three AI models times two Observers) were generated and evaluated using the K5 cross-validation protocol. Ten different kinds of metrics were used for the evaluation of the six AI models. The two Observers' error metrics were compared to validate the hypothesis for every AI model, keeping below the error threshold of 5%. Our results showed that the difference in these errors were 0%, 0.51%, and 2.04% (all < 5%), respectively, for the three AI models, validating the hypothesis. Statistical analysis was conducted using a standard paired *t*-test, ANOVA, and Wilcoxon test to prove the system's hypothesis. The inter-variability COVLIAS 1.0 showed clinically robust and statistically stable outcomes for this pilot study and, thus, can be adapted in clinical settings.

Author Contributions: Conceptualization: J.S.S., M.K.K., N.N.K. and M.M.; Data Curation: M.C., L.S., G.F., M.T., K.V., P.R.K., F.N., Z.R. and A.G.; Formal Analysis: J.S.S., M.K.K.; Investigation: J.S.S., M.K.K., I.M.S., P.S.C., A.M.J., S.M., J.R.L., G.P., D.W.S., P.P.S., G.T., A.P., D.P.M., V.A., J.S.T., M.A.-M., S.K.D., A.N., A.S. and A.A.; Methodology: J.S.S., S.A. and A.B.; Project Administration: J.S.S., M.K.K.; Deep Learning Computing Resources: S.K.G.; Software Design and Usage: S.A., P.E. and V.K.; Software Verification: J.S.S., L.S. and M.K.K.; Supervision: J.S.S., M.K.K. and S.N.; Scientific Validation: J.S.S. and S.A.; Clinical Validation and Discussions: J.S.S., M.K.K. and L.S.; Visualization: S.A.; Writing—Original Draft: J.S.S. and S.A.; Writing—Review & Editing: J.S.S., M.K.K., L.S., R.P., S.K.G., I.M.S., M.T., P.S.C., A.M.J., N.N.K., K.V., S.M., J.R.L., G.P., M.M., D.W.S., A.B., P.P.S., G.T., A.P.,

D.P.M., V.A., G.D.K., M.A.-M., S.K.D., A.N., A.S., V.R., M.F., F.N., A.G. and S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. GBTI deals in lung image analysis and Jasjit S. Suri is affiliated with GBTI.

Abbreviations

SN	Symbol	Description of the Symbols
1	ACC (ai)	Accuracy
2	AE	Area Error
3	AI	Artificial Intelligence
4	ARDS	Acute Respiratory Distress Syndrome
5	AUC	Area Under the Curve
6	BA	Bland-Altman
7	BE	Boundary Error
8	CC	Correlation coefficient
9	CE	Cross Entropy
10	COVID	Coronavirus disease
11	COVLIAS	COVID Lung Image Analysis System
12	CT	Computed Tomography
13	DL	Deep Learning
14	DS	Dice Similarity
15	FoM	Figure of merit
16	GT	Ground Truth
17	HDL	Hybrid Deep Learning
18	IS	Image Size
19	JI	Jaccard Index
20	LAE	Lung Area Error
21	LLAE	Lung Long Axis Error
22	NIH	National Institute of Health
23	PC	Pixel Counting
24	RF	Resolution Factor
25	ROC	Receiver operating characteristic
26	SDL	Solo Deep Learning
27	VGG	Visual Geometric Group
28	VS	Variability studies
29	WHO	World Health Organization

Symbols

SN	Symbol	Description of the Symbols
1	l_{CE}	Cross Entropy-loss
2	m	Model used for segmentation in the total number of models M
3	n	Image scan number in total number N
4	$\overline{A}_{ai}(m)$	Mean estimated lung area for all images using AI model 'm'
5	$A_{ai}(m,n)$	Estimated Lung Area using AI model 'm' and image 'n'
6	$A_{gt}(n)$	GT lung area for image 'n'
7	$\overline{A}_{gt}$	Mean ground truth area for all images N in the database
8	$\overline{LA}_{ai}(m)$	Mean estimated lung long axis for all images using AI model 'm'
9	$LA_{ai}(m,n)$	Estimated lung long axis using AI model 'm' and image 'n'
10	$LA_{gt}(n)$	GT lung long axis for image 'n'
11	$\overline{LA}_{gt}$	Mean ground truth long axis for all images N in the database
12	$FoM_A(m)$	Figure-of-Merit for segmentation model 'm'
14	$FoM_{LA}(m)$	Figure-of-Merit for long axis for model 'm'
15	JI	Jaccard Index for a specific segmentation model
16	DSC	Dice Similarity Coefficient for a specific segmentation model
17	TP, TN	True Positive and True Negative
18	FP, FN	False Positive and False Negative
19	x_{i}	GT label
20	$\mathrm{p_i}$	SoftMax classifier probability
21	Y_p	Ground truth image
22	$\hat{Y}_p$	Estimated image
23	P	Total no of pixels in an image in x, y-direction
24	K5	Cross-validation protocol with 80% training and 20% testing (5 folds)
Deep Learning Segmentation Architectures		
25	PSP Net	SDL model for lung segmentation with pyramidal feature extraction
26	VGG-SegNet	HDL model designed by fusion of VGG-19 and SegNet architecture
27	ResNet-SegNet	HDL model designed by fusion of ResNet-50 and SegNet architecture

References

1. Agarwal, M.; Saba, L.; Gupta, S.K.; Johri, A.M.; Khanna, N.N.; Mavrogeni, S.; Laird, J.R.; Pareek, G.; Miner, M.; Sfikakis, P.P. Wilson disease tissue classification and characterization using seven artificial intelligence models embedded with 3D optimization paradigm on a weak training brain magnetic resonance imaging datasets: A supercomputer application. *Med. Biol. Eng. Comput.* **2021**, *59*, 511–533. [CrossRef]
2. Cau, R.; Pacielli, A.; Fatemeh, H.; Vaudano, P.; Arru, C.; Crivelli, P.; Stranieri, G.; Suri, J.S.; Mannelli, L.; Conti, M.; et al. Complications in COVID-19 patients: Characteristics of pulmonary embolism. *Clin. Imaging* **2021**, *77*, 244–249. [CrossRef] [PubMed]
3. Saba, L.; Gerosa, C.; Fanni, D.; Marongiu, F.; La Nasa, G.; Caocci, G.; Barcellona, D.; Balestrieri, A.; Coghe, F.; Orru, G.; et al. Molecular pathways triggered by COVID-19 in different organs: ACE2 receptor-expressing cells under attack? A review. *Eur. Rev. Med. Pharmacol. Sci.* **2020**, *24*, 12609–12622.
4. Cau, R.; Bassareo, P.P.; Mannelli, L.; Suri, J.S.; Saba, L. Imaging in COVID-19-related myocardial injury. *Int. J. Cardiovasc. Imaging* **2021**, *37*, 1349–1360. [CrossRef] [PubMed]
5. Viswanathan, V.; Viswanathan, V.; Puvvula, A.; Jamthikar, A.D.; Saba, L.; Johri, A.M.; Kotsis, V.; Khanna, N.N.; Dhanjil, S.K.; Majhail, M.; et al. Bidirectional link between diabetes mellitus and coronavirus disease 2019 leading to cardiovascular disease: A narrative review. *World J. Diabetes* **2021**, *12*, 215–237. [CrossRef]
6. Suri, J.S.; Agarwal, S.; Gupta, S.K.; Puvvula, A.; Biswas, M.; Saba, L.; Bit, A.; Tandel, G.S.; Agarwal, M.; Patrick, A.; et al. A narrative review on characterization of acute respiratory distress syndrome in COVID-19-infected lungs using artificial intelligence. *Comput. Biol. Med.* **2021**, *130*, 104210. [CrossRef]
7. Cau, R.; Falaschi, Z.; Paschè, A.; Danna, P.; Arioli, R.; Arru, C.D.; Zagaria, D.; Tricca, S.; Suri, J.S.; Karla, M.K.; et al. Computed tomography findings of COVID-19 pneumonia in Intensive Care Unit-patients. *J. Public Health Res.* **2021**, *10*, 2270. [CrossRef]
8. Emery, S.L.; Erdman, D.D.; Bowen, M.D.; Newton, B.R.; Winchell, J.M.; Meyer, R.F.; Tong, S.; Cook, B.T.; Holloway, B.P.; McCaustland, K.A.; et al. Real-time reverse transcription–polymerase chain reaction assay for SARS-associated coronavirus. *Emerg. Infect. Dis.* **2004**, *10*, 311–316. [CrossRef] [PubMed]

9. Wu, X.; Hui, H.; Niu, M.; Li, L.; Wang, L.; He, B.; Yang, X.; Li, L.; Li, H.; Tian, J.; et al. Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. *Eur. J. Radiol.* **2020**, *128*, 109041. [CrossRef] [PubMed]
10. Pathak, Y.; Shukla, P.K.; Tiwari, A.; Stalin, S.; Singh, S. Deep transfer learning based classification model for COVID-19 disease. *IRBM* **2020**, in press. [CrossRef]
11. Saba, L.; Suri, J.S. *Multi-Detector CT Imaging: Principles, Head, Neck, and Vascular Systems*; CRC Press: Boca Raton, FL, USA, 2013; Volume 1.
12. Gozes, O.; Frid-Adar, M.; Greenspan, H.; Browning, P.D.; Zhang, H.; Ji, W.; Bernheim, A.; Siegel, E. Rapid ai development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *arXiv* **2020**, arXiv:05037.
13. Shalbaf, A.; Vafaeezadeh, M. Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 115–123.
14. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. COVID-CT-dataset: A CT scan dataset about COVID-19. *arXiv* **2020**, arXiv:13865.
15. Alqudah, A.M.; Qazan, S.; Alquran, H.; Qasmieh, I.A.; Alqudah, A. COVID-2019 Detection Using X-ray Images and Artificial Intelligence Hybrid Systems. *Phys. Sci.* **2020**, *2*, 1.
16. Aslan, M.F.; Unlersen, M.F.; Sabanci, K.; Durdu, A. CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection. *Appl. Soft Comput.* **2021**, *98*, 10691. [CrossRef]
17. Wu, Y.H.; Gao, S.H.; Mei, J.; Xu, J.; Fan, D.P.; Zhang, R.G.; Cheng, M.M. Jcs: An explainable COVID-19 diagnosis system by joint classification and segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 3113–3126. [CrossRef] [PubMed]
18. Xu, X.; Jiang, X.; Ma, C.; Du, P.; Li, X.; Lv, S.; Yu, L.; Ni, Q.; Chen, Y.; Su, J.; et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering* **2020**, *6*, 1122–1129. [CrossRef] [PubMed]
19. El-Baz, A.; Suri, J. *Machine Learning in Medicine*; CRC Press: Boca Raton, FL, USA, 2021; ISBN 9781138106901.
20. Suri, J.S.; Rangayyan, R.M. *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer*; SPIE Publications: Bellingham, WA, USA, 2006.
21. Biswas, M.; Kuppili, V.; Edla, D.R.; Suri, H.S.; Saba, L.; Marinhoe, R.T.; Sanches, J.M.; Suri, J.S. Symtosis: A liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm. *Comput. Methods Programs Biomed.* **2018**, *155*, 165–177. [CrossRef] [PubMed]
22. Acharya, U.R.; Sree, S.V.; Ribeiro, R.; Krishnamurthi, G.; Marinho, R.T.; Sanches, J.; Suri, J.S. Data mining framework for fatty liver disease classification in ultrasound: A hybrid feature extraction paradigm. *Med. Phys.* **2012**, *39*, 4255–4264. [CrossRef] [PubMed]
23. Acharya, U.R.; Sree, S.V.; Krishnan, M.M.R.; Molinari, F.; Garberoglio, R.; Suri, J.S. Non-invasive automated 3D thyroid lesion classification in ultrasound: A class of ThyroScan™ systems. *Ultrasonics* **2012**, *52*, 508–520. [CrossRef]
24. Acharya, U.R.; Swapna, G.; Sree, S.V.; Molinari, F.; Gupta, S.; Bardales, R.H.; Witkowska, A.; Suri, J.S. A review on ultrasound-based thyroid cancer tissue characterization and automated classification. *Technol. Cancer Res. Treat.* **2014**, *13*, 289–301. [CrossRef]
25. Molinari, F.; Mantovani, A.; Deandrea, M.; Limone, P.; Garberoglio, R.; Suri, J.S. Characterization of single thyroid nodules by contrast-enhanced 3-D ultrasound. *Ultrasound Med. Biol.* **2010**, *36*, 1616–1625. [CrossRef]
26. Shrivastava, V.K.; Londhe, N.D.; Sonawane, R.S.; Suri, J.S. Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: A first comparative study of its kind. *Comput. Methods Programs Biomed.* **2016**, *126*, 98–109. [CrossRef] [PubMed]
27. Shrivastava, V.K.; Londhe, N.D.; Sonawane, R.S.; Suri, J.S. Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. *Expert Syst. Appl.* **2015**, *42*, 6184–6195. [CrossRef]
28. Pareek, G.; Acharya, U.R.; Sree, S.V.; Swapna, G.; Yantri, R.; Martis, R.J.; Saba, L.; Krishnamurthi, G.; Mallarini, G.; El-Baz, A.; et al. Prostate tissue characterization/classification in 144 patient population using wavelet and higher order spectra features from transrectal ultrasound images. *Technol. Cancer Res. Treat.* **2013**, *12*, 545–557. [CrossRef]
29. McClure, P.; Elnakib, A.; El-Ghar, M.A.; Khalifa, F.; Soliman, A.; El-Diasty, T.; Suri, J.S.; Elmaghraby, A.; El-Baz, A. In-vitro and in-vivo diagnostic techniques for prostate cancer: A review. *J. Biomed. Nanotechnol.* **2014**, *10*, 2747–2777. [CrossRef] [PubMed]
30. Mookiah, M.R.K.; Acharya, U.R.; Martis, R.J.; Chua, C.K.; Lim, C.M.; Ng, E.Y.K.; Laude, A. Evolutionary algorithm based classifier parameter tuning for automatic diabetic retinopathy grading: A hybrid feature extraction approach. *Knowl.-Based Syst.* **2013**, *39*, 9–22. [CrossRef]
31. Than, J.C.; Saba, L.; Noor, N.M.; Rijal, O.M.; Kassim, R.M.; Yunus, A.; Suri, H.S.; Porcu, M.; Suri, J.S. Lung disease stratification using amalgamation of Riesz and Gabor transforms in machine learning framework. *Comput. Biol. Med.* **2017**, *89*, 197–211. [CrossRef] [PubMed]
32. El-Baz, A.; Jiang, X.; Suri, J.S. *Biomedical Image Segmentation: Advances and Trends*; CRC Press: Boca Raton, FL, USA, 2016.
33. Than, J.C.; Saba, L.; Noor, N.M.; Rijal, O.M.; Kassim, R.M.; Yunus, A.; Suri, H.S.; Porcu, M.; Suri, J.S. Shape recovery algorithms using level sets in 2-D/3-D medical imagery: A state-of-the-art review. *IEEE Trans. Inf. Technol. Biomed.* **2002**, *6*, 8–28.
34. El-Baz, A.S.; Acharya, R.; Mirmehdi, M.; Suri, J.S. *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*; Springer Science & Business Media: Berlin, Germany, 2011; Volume 2.
35. El-Baz, A.; Suri, J.S. *Level Set Method in Medical Imaging Segmentation*; CRC Press: Boca Raton, FL, USA, 2019.

36. Saba, L.; Sanagala, S.S.; Gupta, S.K.; Koppula, V.K.; Johri, A.M.; Khanna, N.N.; Mavrogeni, S.; Laird, J.R.; Pareek, G.; Miner, M.; et al. Multimodality carotid plaque tissue characterization and classification in the artificial intelligence paradigm: A narrative review for stroke application. *Ann. Transl. Med.* **2021**, *9*, 1206. [CrossRef] [PubMed]
37. Acharya, U.R.; Sree, S.V.; Krishnan, M.M.R.; Krishnananda, N.; Ranjan, S.; Umesh, P.; Suri, J.S. Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. *Comput. Methods Programs Biomed.* **2013**, *112*, 624–632. [CrossRef]
38. Agarwal, M.; Saba, L.; Gupta, S.K.; Carriero, A.; Falaschi, Z.; Paschè, A.; Danna, P.; El-Baz, A.; Naidu, S.; Suri, J.S. A novel block imaging technique using nine artificial intelligence models for COVID-19 disease classification, characterization and severity measurement in lung computed tomography scans on an Italian cohort. *J. Med. Syst.* **2021**, *45*, 1–30. [CrossRef]
39. Saba, L.; Agarwal, M.; Patrick, A.; Puvvula, A.; Gupta, S.K.; Carriero, A.; Laird, J.R.; Kitas, G.D.; Johri, A.M.; Balestrieri, A.; et al. Six artificial intelligence paradigms for tissue characterisation and classification of non-COVID-19 pneumonia against COVID-19 pneumonia in computed tomography lungs. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 423–434. [CrossRef]
40. Skandha, S.S.; Gupta, S.K.; Saba, L.; Koppula, V.K.; Johri, A.M.; Khanna, N.N.; Mavrogeni, S.; Laird, J.R.; Pareek, G.; Miner, M.; et al. 3-D optimized classification and characterization artificial intelligence paradigm for cardiovascular/stroke risk stratification using carotid ultrasound-based delineated plaque: Atheromatic™ 2.0. *Comput. Biol. Med.* **2020**, *125*, 103958. [CrossRef]
41. 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]
42. Sarker, M.M.K.; Makhlouf, Y.; Banu, S.F.; Chambon, S.; Radeva, P.; Puig, D. Web-based efficient dual attention networks to detect COVID-19 from X-ray images. *Electron. Lett.* **2020**, *56*, 1298–1301. [CrossRef]
43. Sarker, M.M.K.; Makhlouf, Y.; Craig, S.G.; Humphries, M.P.; Loughrey, M.; James, J.A.; Salto-Tellez, M.; O'Reilly, P.; Maxwell, P. A Means of Assessing Deep Learning-Based Detection of ICOS Protein Expression in Colon Cancer. *Cancers* **2021**, *13*, 3825. [CrossRef] [PubMed]
44. Jain, P.K.; Sharma, N.; Giannopoulos, A.A.; Saba, L.; Nicolaides, A.; Suri, J.S. Hybrid deep learning segmentation models for atherosclerotic plaque in internal carotid artery B-mode ultrasound. *Comput. Biol. Med.* **2021**, *136*, 104721. [CrossRef]
45. Jena, B.; Saxena, S.; Nayak, G.K.; Saba, L.; Sharma, N.; Suri, J.S. Artificial Intelligence-based Hybrid Deep Learning Models for Image Classification: The First Narrative Review. *Comput. Biol. Med.* **2021**, *137*, 104803. [CrossRef] [PubMed]
46. Suri, J.; Agarwal, S.; Gupta, S.K.; Puvvula, A.; Viskovic, K.; Suri, N.; Alizad, A.; El-Baz, A.; Saba, L.; Fatemi, M.; et al. Systematic Review of Artificial Intelligence in Acute Respiratory Distress Syndrome for COVID-19 Lung Patients: A Biomedical Imaging Perspective. *IEEE J. Biomed. Health Inform.* **2021**, *25*.
47. Saba, L.; Banchhor, S.K.; Araki, T.; Viskovic, K.; Londhe, N.D.; Laird, J.R.; Suri, H.S.; Suri, J.S. Intra- and inter-operator reproducibility of automated cloud-based carotid lumen diameter ultrasound measurement. *Indian Heart J.* **2018**, *70*, 649–664. [CrossRef] [PubMed]
48. Saba, L.; Than, J.C.; Noor, N.M.; Rijal, O.M.; Kassim, R.M.; Yunus, A.; Ng, C.R.; Suri, J.S. Inter-observer Variability Analysis of Automatic Lung Delineation in Normal and Disease Patients. *J. Med. Syst.* **2016**, *40*, 142. [CrossRef]
49. Zhang, S.; Suri, J.S.; Salvado, O.; Chen, Y.; Wacker, F.K.; Wilson, D.L.; Duerk, J.L.; Lewin, J.S. Inter-and Intra-Observer Variability Assessment of in Vivo Carotid Plaque Burden Quantification Using Multi-Contrast Dark Blood MR Images. *Stud. Health Technol. Inform.* **2005**, *113*, 384–393. [PubMed]
50. Aggarwal, D.; Saini, V. Factors limiting the utility of bronchoalveolar lavage in the diagnosis of COVID-19. *Eur. Respir. J.* **2020**, *56*, 2003116. [CrossRef]
51. Saba, L.; Banchhor, S.K.; Suri, H.S.; Londhe, N.D.; Araki, T.; Ikeda, N.; Viskovic, K.; Shafique, S.; Laird, J.R.; Gupta, A.; et al. Accurate cloud-based smart IMT measurement, its validation and stroke risk stratification in carotid ultrasound: A web-based point-of-care tool for multicenter clinical trial. *Comput. Biol. Med.* **2016**, *75*, 217–234. [CrossRef]
52. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
53. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
54. Suri, J.S.; Agarwal, S.; Pathak, R.; Ketireddy, V.; Columbu, M.; Saba, L.; Gupta, S.K.; Faa, G.; Singh, I.M.; Turk, M.; et al. COVLIAS 1.0: Lung Segmentation in COVID-19 Computed Tomography Scans Using Hybrid Deep Learning Artificial Intelligence Models. *Diagnostics* **2021**, *11*, 1405. [CrossRef] [PubMed]
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
56. Acharya, U.R.; Faust, O.; Sree, S.V.; Molinari, F.; Saba, L.; Nicolaides, A.; Suri, J.S. An accurate and generalized approach to plaque characterization in 346 carotid ultrasound scans. *IEEE Trans. Instrum. Meas.* **2012**, *61*, 1045–1053. [CrossRef]
57. Acharya, U.R.; Saba, L.; Molinari, F.; Guerriero, S.; Suri, J.S. Ovarian tumor characterization and classification: A class of GyneScan™ systems. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; IEEE: Piscataway, NJ, USA, 2012.
58. Araki, T.; Ikeda, N.; Dey, N.; Acharjee, S.; Molinari, F.; Saba, L.; Godia, E.C.; Nicolaides, A.; Suri, J.S. Shape-based approach for coronary calcium lesion volume measurement on intravascular ultrasound imaging and its association with carotid intima-media thickness. *J. Ultrasound Med.* **2015**, *34*, 469–482. [CrossRef]

59. Barqawi, A.B.; Li, L.; Crawford, E.D.; Fenster, A.; Werahera, P.N.; Kumar, D.; Miller, S.; Suri, J.S. Three different strategies for real-time prostate capsule volume computation from 3-D end-fire transrectal ultrasound. In Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 22–26 August 2007; IEEE: Piscataway, NJ, USA, 2007.
60. Suri, J.S.; Haralick, R.M.; Sheehan, F.H. Left ventricle longitudinal axis fitting and its apex estimation using a robust algorithm and its performance: A parametric apex model. In Proceedings of the International Conference on Image Processing, Santa Barbara, CA, USA, 14–17 July 1997; IEEE: Piscataway, NJ, USA, 1997.
61. Singh, B.K.; Verma, K.; Thoke, A.S.; Suri, J.S. Risk stratification of 2D ultrasound-based breast lesions using hybrid feature selection in machine learning paradigm. *Measurement* **2017**, *105*, 146–157. [CrossRef]
62. Riffenburgh, R.H.; Gillen, D.L. Contents. In *Statistics in Medicine*; Academic Press: Cambridge, MA, USA, 2020; pp. ix–xvi.
63. Acharya, R.U.; Faust, O.; Alvin, A.P.C.; Sree, S.V.; Molinari, F.; Saba, L.; Nicolaides, A.; Suri, J.S. Symptomatic vs. asymptomatic plaque classification in carotid ultrasound. *J. Med. Syst.* **2012**, *36*, 1861–1871. [CrossRef]
64. Acharya, U.R.; Vinitha Sree, S.; Mookiah, M.R.K.; Yantri, R.; Molinari, F.; Zieleźnik, W.; Małyszek-Tumidajewicz, J.; Stępień, B.; Bardales, R.H.; Witkowska, A.; et al. Diagnosis of Hashimoto's thyroiditis in ultrasound using tissue characterization and pixel classification. *Proc. Inst. Mech. Eng. Part H J. Eng. Med.* **2013**, *227*, 788–798. [CrossRef]
65. Acharya, U.R.; Faust, O.; Alvin, A.P.C.; Krishnamurthi, G.; Seabra, J.C.; Sanches, J.; Suri, J.S. Understanding symptomatology of atherosclerotic plaque by image-based tissue characterization. *Comput. Methods Programs Biomed.* **2013**, *110*, 66–75. [CrossRef]
66. Acharya, U.R.; Faust, O.; Sree, S.V.; Alvin, A.P.C.; Krishnamurthi, G.; Sanches, J.; Suri, J.S. Atheromatic™: Symptomatic vs. asymptomatic classification of carotid ultrasound plaque using a combination of HOS, DWT & texture. In Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 3 August–3 September 2011; IEEE: Piscataway, NJ, USA, 2011.
67. Acharya, U.R.; Mookiah, M.R.K.; Sree, S.V.; Afonso, D.; Sanches, J.; Shafique, S.; Nicolaides, A.; Pedro, L.M.; e Fernandes, J.F.; Suri, J.S. Atherosclerotic plaque tissue characterization in 2D ultrasound longitudinal carotid scans for automated classification: A paradigm for stroke risk assessment. *Med. Biol. Eng. Comput.* **2013**, *51*, 513–523. [CrossRef]
68. Molinari, F.; Liboni, W.; Pavanelli, E.; Giustetto, P.; Badalamenti, S.; Suri, J.S. Accurate and automatic carotid plaque characterization in contrast enhanced 2-D ultrasound images. In Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 22–26 August 2007; IEEE: Piscataway, NJ, USA, 2007.
69. Saba, L.; Biswas, M.; Suri, H.S.; Viskovic, K.; Laird, J.R.; Cuadrado-Godia, E.; Nicolaides, A.; Khanna, N.N.; Viswanathan, V.; Suri, J.S. Ultrasound-based carotid stenosis measurement and risk stratification in diabetic cohort: A deep learning paradigm. *Cardiovasc. Diagn. Ther.* **2019**, *9*, 439–461. [CrossRef]
70. Biswas, M.; Kuppili, V.; Saba, L.; Edla, D.R.; Suri, H.S.; Sharma, A.; Cuadrado-Godia, E.; Laird, J.R.; Nicolaides, A.; Suri, J.S. Deep learning fully convolution network for lumen characterization in diabetic patients using carotid ultrasound: A tool for stroke risk. *Med Biol. Eng. Comput.* **2019**, *57*, 543–564. [CrossRef]
71. Chaddad, A.; Hassan, L.; Desrosiers, C. Deep CNN models for predicting COVID-19 in CT and x-ray images. *J. Med. Imaging* **2021**, *8* (Suppl. S1), 014502. [CrossRef] [PubMed]
72. Gunraj, H.; Wang, L.; Wong, A. COVIDNet-CT: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases From Chest CT Images. *Front. Med.* **2020**, *7*, 608525. [CrossRef] [PubMed]
73. Iyer, T.J.; Raj, A.N.J.; Ghildiyal, S.; Nersisson, R. Performance analysis of lightweight CNN models to segment infectious lung tissues of COVID-19 cases from tomographic images. *PeerJ Comput. Sci.* **2021**, *7*, e368. [CrossRef]
74. Ranjbarzadeh, R.; Jafarzadeh Ghoushchi, S.; Bendechache, M.; Amirabadi, A.; Ab Rahman, M.N.; Baseri Saadi, S.; Aghamohammadi, A.; Kooshki Forooshani, M. Lung Infection Segmentation for COVID-19 Pneumonia Based on a Cascade Convolutional Network from CT Images. *BioMed Res. Int.* **2021**, *2021*, 5544742. [CrossRef]
75. Erasmus, J.J.; Gladish, G.W.; Broemeling, L.; Sabloff, B.S.; Truong, M.T.; Herbst, R.S.; Munden, R.F. Interobserver and intraobserver variability in measurement of non–small-cell carcinoma lung lesions: Implications for assessment of tumor response. *J. Clin. Oncol.* **2003**, *21*, 2574–2582. [CrossRef]
76. Joskowicz, L.; Cohen, D.; Caplan, N.; Sosna, J. Inter-observer variability of manual contour delineation of structures in CT. *Eur. Radiol.* **2019**, *29*, 1391–1399. [CrossRef] [PubMed]
77. El-Baz, A.; Suri, J. *Lung Imaging and CADx*; CRC Press: Boca Raton, FL, USA, 2019.
78. El-Baz, A.; Suri, J.S. *Lung Imaging and Computer Aided Diagnosis*; CRC Press: Boca Raton, FL, USA, 2011.
79. Sudeep, P.V.; Palanisamy, P.; Rajan, J.; Baradaran, H.; Saba, L.; Gupta, A.; Suri, J.S. Speckle reduction in medical ultrasound images using an unbiased non-local means method. *Biomed. Signal Process. Control.* **2016**, *28*, 1–8. [CrossRef]
80. Sarker, M.M.K.; Rashwan, H.A.; Akram, F.; Singh, V.K.; Banu, S.F.; Chowdhury, F.U.; Choudhury, K.A.; Chambon, S.; Radeva, P.; Puig, D.; et al. SLSNet: Skin lesion segmentation using a lightweight generative adversarial network. *Expert Syst. Appl.* **2021**, *183*, 115433. [CrossRef]
81. Saba, L.; Agarwal, M.; Sanagala, S.S.; Gupta, S.K.; Sinha, G.R.; Johri, A.M.; Khanna, N.N.; Mavrogeni, S.; Laird, J.R.; Pareek, G.; et al. Brain MRI-based Wilson disease tissue classification: An optimised deep transfer learning approach. *Electron. Lett.* **2020**, *56*, 1395–1398. [CrossRef]
82. El-Baz, A.; Suri, J.S. *Big Data in Multimodal Medical Imaging*; CRC Press: Boca Raton, FL, USA, 2019.

diagnostics

MDPI

Article

Diagnostic Performance of Dual-Energy Subtraction Radiography for the Detection of Pulmonary Emphysema: An Intra-Individual Comparison

Julia A. Mueller [1], Katharina Martini [1,*], Matthias Eberhard [1], Mathias A. Mueller [2], Alessandra A. De Silvestro [1], Philipp Breiding [1] and Thomas Frauenfelder [1]

[1] Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, 8091 Zürich, Switzerland; juliaanna@gmx.ch (J.A.M.); Matthias.eberhard@gmx.de (M.E.); desilvestro.alessandra@usz.ch (A.A.D.S.); philipp.breiding@usz.ch (P.B.); Thomas.frauenfelder@usz.ch (T.F.)

[2] Institute of Radiology, Cantonal Hospital of Frauenfeld, 8501 Frauenfeld, Switzerland; Mahias.mueller@stgag.de

* Correspondence: katharina.martini@usz.ch

Abstract: Purpose/Objectives: To compare the diagnostic performance of dual-energy subtraction (DE) and conventional radiography (CR) for detecting pulmonary emphysema using computed tomography (CT) as a reference standard. Methods and Materials: Sixty-six patients (24 female, median age 73) were retrospectively included after obtaining lateral and posteroanterior chest X-rays with a dual-shot DE technique and chest CT within ±3 months. Two experienced radiologists first evaluated the standard CR images and, second, the bone-/soft tissue weighted DE images for the presence (yes/no), degree (1–4), and quadrant-based distribution of emphysema. CT was used as a reference standard. Inter-reader agreement was calculated. Sensitivity and specificity for the correct detection and localization of emphysema was calculated. Further degree of emphysema on CR and DE was correlated with results from CT. A p-value < 0.05 was considered as statistically significant. Results: The mean interreader agreement was substantial for CR and moderate for DE (k_{CR} = 0.611 vs. k_{DE} = 0.433; respectively). Sensitivity, as well as specificity for the detection of emphysema, was comparable between CR and DE (sensitivity$_{CR}$ 96% and specificity$_{CR}$ 75% vs. sensitivity$_{DE}$ 91% and specificity$_{DE}$ 83%; p = 0.157). Similarly, there was no significant difference in the sensitivity or specificity for emphysema localization between CR and DE (sensitivity$_{CR}$ 50% and specificity$_{CR}$ 100% vs. sensitivity$_{DE}$ 57% and specificity$_{DE}$ 100%; p = 0.157). There was a slightly better correlation with CT of emphysema grading in DE compared to CR (r_{DE} = 0.75 vs. r_{CR} = 0.68; p = 0.108); these differences were not statistically significant, however. Conclusion: Diagnostic accuracy for the detection, quantification, and localization of emphysema between CR and DE is comparable. Interreader agreement, however, is better with CR compared to DE

Keywords: lung; conventional radiography; diagnostic procedure; chronic obstructive pulmonary disease

Citation: Mueller, J.A.; Martini, K.; Eberhard, M.; Mueller, M.A.; De Silvestro, A.A.; Breiding, P.; Frauenfelder, T. Diagnostic Performance of Dual-Energy Subtraction Radiography for the Detection of Pulmonary Emphysema: An Intra-Individual Comparison. *Diagnostics* **2021**, *11*, 1849. https://doi.org/10.3390/diagnostics11101849

Academic Editor: Sameer Antani

Received: 26 August 2021
Accepted: 5 October 2021
Published: 7 October 2021

1. Introduction

Chronic obstructive pulmonary disease (COPD) is defined symptomatically as chronic bronchitis and physiologically as airway obstruction or anatomically as emphysema [1], usually caused by tobacco use [2]. Its course is creeping and progressive with a high impairment in quality of life [3] and COPD is a leading cause of death worldwide [4]. The early detection of emphysematous lung tissue is important to prevent and manage the global disease burden [5,6].

Since the pathogenesis in COPD is not fully understood and pulmonary function tests (PFT) are not sensitive in detecting mild emphysema and fail to register the heterogeneity of the disease, radiological imaging plays a major role in emphysema detection and evaluation [7,8]. Computed tomography (CT) is the most sensitive radiological imaging modality

for the detection, quantification, and phenotyping of emphysema [9,10]. Due to the high sensitivity of HRCT, pulmonary emphysematous changes detected before PFT (forced expiratory volume in 1 s, FEV1) are pathologic [11]. The benefit of earlier therapy of chronic cough with normal FEV1 but conspicuous features in CT is not yet known, but a delay of disease progression is postulated [12]. Conventional radiography (CR) seems not to be as reliable as CT in the detection of emphysema unless the disease is advanced. Indirect signs, such as horizontal standing ribs, extended intercostal space, flattened diaphragms, retrosternal air space, increased radiographic transparency, and rarefication of small blood vessels in the periphery, can give a hint to the underlying disease, however [13].

Nevertheless, CR holds its position as a first diagnostic approach in the daily clinical practice due to its broad availability; fast examination time; low cost and low radiation dose [14,15]; and new developments, such as dual energy subtraction chest X-ray (DE), which might help to increase its diagnostic accuracy.

In DE, besides the standard image, a soft tissue image with bone information removed and a bone image with soft tissue information removed is generated [16,17].

Previous studies have shown that DE images improve the sensitivity for shading lesions, such as the detection of infectious consolidations, tumors, interstitial lung changes, and aortic or tracheal calcification, compared to CR-images [18–21]. We hypothesize that DE might improve the conspicuousness of hyperlucent lung pathologies in a similar way.

Therefore, the aim of this study was to assess the diagnostic performance of DE for detecting pulmonary emphysema compared to CR using CT as a reference standard.

2. Materials and Methods

2.1. Patient Population

The study was approved by the institutional review board and local ethics committee (KEK Zürich: Cantonal ethics committee Zurich Switzerland). Informed consent was waived because of the retrospective setting of this study (blinded for review).

In this observational study, 74 patients (age: 71.6 ± 8.7 years, 26 females) undergoing CR, DE, and chest CT between September 2015 and Mai 2019 were retrospectively included. Inclusion criterion was the presence of a CT ± 3 months within the conventional imaging. Patients were excluded when there was an intervention (interventional or surgical lung resection) between imaging. In- and exclusion criteria are listed in Table 1.

Table 1. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Age > 18 years	Incapability of undergoing upright chest radiography
Existence of chest examinations with DE and CT	Cardiopulmonary decompensation
Description of lung emphysema in radiologic findings or COPD in list of diagnosis	Obscured lung tissue by foreign bodies, e.g., cardiac devices
Short time interval between the two imaging modalities	Consolidation, e.g., empyema, encapsulated pneumonia
	Changes between performed DE and reference CTOccurrence of a pneumothorax
	Lung volume reduction surgery
	Endoscopic lung volume reduction, e.g., valves, coils, sealants

DE = dual energy subtraction radiography; CT = computed tomography.

2.2. Data Acquisition

2.2.1. CR and DE Images

All patients underwent chest radiography in lateral and p.a. projection, whereby the latter was obtained using a dual energy mode (FDR AcSelerate, Fujifilm, Tokyo, Japan) at a tube current of 7 mA and a tube voltage of 120 kV and 60 kV after a delay of 150 ms according to institution's standard protocol. The higher energy exposure was used to produce the CR image. With the use of a post-processing algorithm, the "virtual" soft tissue and bone image were calculated from the two acquisitions (see Figure 1).

Figure 1. Moderate emphysema, right upper lobe and left lower lobe. Left: conventional X-ray postero anterior (p.a.), middle: dual energy X-ray p.a., right: corresponding computed tomography image.

2.2.2. CT Images

Single-energy CT was performed with or without intravenously injected contrast agent at 120 kV/110 mA (Somatom Sensation, Somatom Flash and Somatom Force, Siemens Healthcare, Forchheim, Germany) and reconstructed with a slice thickness of 2.0 mm. A radiation dose was recorded for each scan.

2.3. Image Analysis

2.3.1. CR/DE Image Analysis

All images were anonymized prior to readout. Two experienced readers (5 and 10 years of experience in thoracic imaging, respectively) reviewed the images in two reading-rounds:

Reading-round 1 (CR): Only the conventional p.a. and lateral projections were evaluated.

Reading round 2 (DE): All images (including the p.a. bone and soft tissue images) of the patients were evaluated.

Readers had to evaluate the images for the presence (yes/no) of emphysema. If emphysema was present, readers had to score the degree of emphysema (none, mild, moderate and strong; 1–4) of emphysema. For the quantification of emphysema, readers were trained with four data sets showing the entire range of emphysema manifestations (Figures 2 and 3).

Readers further had to sign the quadrant (upper right, lower right, upper left, lower left) of the most affected area.

If there was a disagreement between the two readers, re-evaluation was performed until consensus was sought. The time frame between the two reading-rounds was of four weeks to avoid a recall bias.

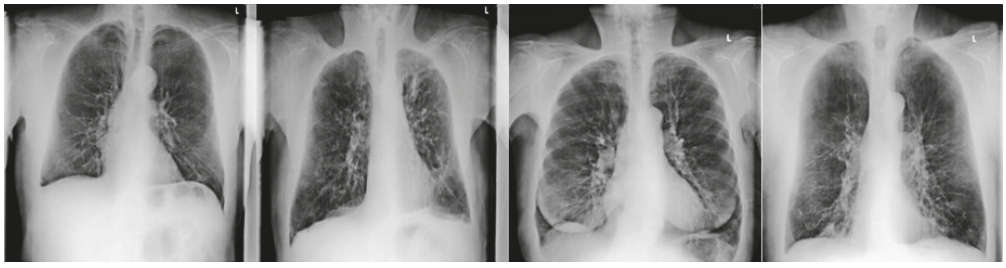


Figure 2. Dual energy soft tissue X-ray image of non-mild-moderate-severe emphysema in the right upper lobe.

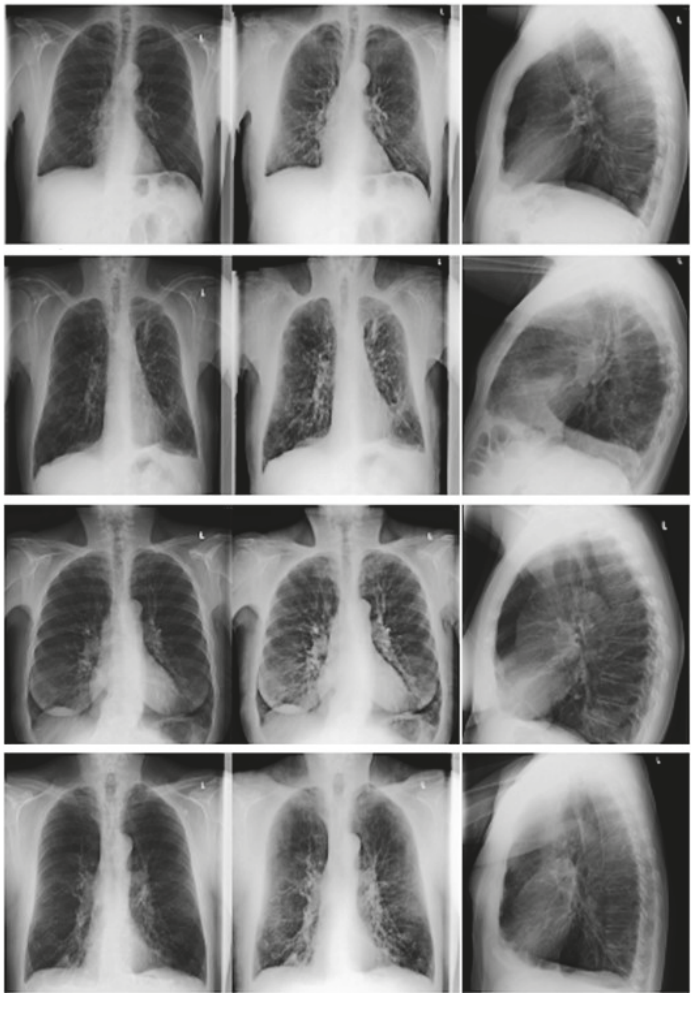


Figure 3. Four patients with varying degrees of emphysema in the right upper lobe.First row: no emphysema; second row: mild emphysema; third row: moderate emphysema; and fourth row: severe emphysema. Left column: conventional postero anterior (p.a.) X-ray; middle column: dual-energy p.a. X-ray; and right column: conventional lateral X-ray.

2.3.2. CT Image Analysis

One expert reader (15 years of experience in thoracic imaging) evaluated the CTs for the presence (yes/no) of emphysema. Further evaluation of CT images was performed with a commercially available software tool (Ziostation, Ziosoft Inc., Tokyo, Japan). The software quantified emphysema using the Goddard score (Figure 4) and assessed the ratio of low attenuation area to lung volume (LAA%) using a threshold of −950 Hounsfield units (HU).

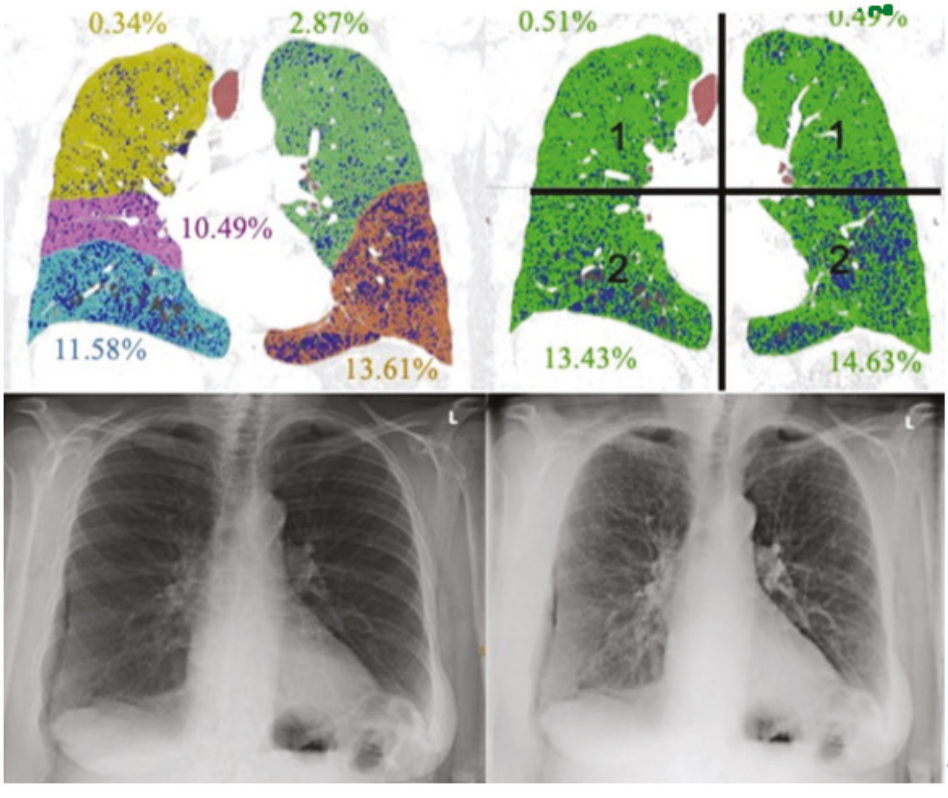


Figure 4. Lobe- and quadrant-based quantification.

The Goddard score is a semi-quantitative assessment score functioning as a surrogate marker for the presence of emphysema based on the evaluation of low attenuation areas in a number of representative lung fields. The total score is defined as the sum of the single scores [22]. The Goddard score was used for overall emphysema grading as well as for defining the most affected lung quadrant.

2.3.3. Statistical Analysis

Statistical analysis was conducted using SPSS (released 2017, version 25.0, Armonk, NY, USA). Interreader agreement for binomial variables was calculated with Cohen's kappa (κ). According to Landis and Koch [23], κ values were defined as follows: slight agreement (κ = 0–0.2), fair agreement (κ = 0.21–0.40), moderate agreement (κ = 0.41–0.60), substantial agreement (κ = 0.61–0.80), and almost perfect agreement (κ = 0.81–1.0). For calculating the interreader agreement in emphysema grading, the interclass correlation coefficient (ICC) was used. An ICC below 0.5 was considered as poor agreement, values between 0.5 and

0.75 were considered as moderate, and values between 0.75 and 0.9 were considered as good agreement [24].

Sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV) for the presence and location of the most affected quadrant were calculated. Further, emphysema grading on CR and DE were correlated with the CT-derived Goddard score using Pearson's correlation (r). A p-value of <0.05 in overall analysis was considered significant.

3. Results

3.1. Patient Population

Indications for conventional radiography and chest CT in the included patents were medical evaluation of potential lung volume reduction surgery (LVRS; n = 23), further assessment of radiological findings (e.g., pulmonary noduli or malignancies) or clarification of clinically persistent symptoms (e.g., chronical cough, recurrent infections or hemoptysis; n = 32), and vascular indications (n = 10).

A total of eight participants were excluded due to the following events within the time interval between CT and DE: lung volume reduction surgery (n = 1), new small pleural effusion (n = 1), new pneumothorax (n = 1), progressive lymphatic spread of leukemia (n = 1), evolution of pericardial effusion (n = 1), paraseptal emphysema (n = 2), and bilateral discrete emphysema (n = 1).

3.2. CT Images: Standard of Reference

The time interval between CT and conventional radiography was of 28 ± 58 days (see Table 2).

Table 2. Patient characteristics.

	Patients	Controls
Number	61	13
Age (years), mean ± SD	71.9 ± 8.2	70.6 ± 10.8
Time between CT and CR/DE (days), mean ± SD	41.8 ± 1.4	20.7 ± 46.8
Male:Female ratio	40:21	8:5

SD = standard deviation; CT = computed tomography; CR = conventional radiography; DE = dual energy subtraction radiography.

3.3. Presence of Emphysema

From the 66 included patients, 81.8% (n = 54) showed emphysema.

3.4. Emphysema Grading

The mean Goddard score based on quantitative CT analysis was 7 (SD ± 4.5; range 0–22). The mean LAA% based on quantitative CT analysis was 16.66% (SD ± 17.1; range 0.01–82.3%). Thirty-two patients had mild emphysema (Goddard-Score 1–7), twenty had moderate emphysema (Goddard-Score 8–15), and two had severe emphysema (Goddard-Score >15).

3.5. CR and DE Image Analysis

3.5.1. Interreader Agreement

Overall, the mean interreader agreement was substantial for CR and moderate for DE (k_{CR} = 0.611 vs. k_{DE} = 0.433; respectively). While the interreader agreement was comparable for emphysema grading (both good), the interreader agreement for the presence of emphysema and for the assignation of the most affected lung quadrant was better in CR (substantial and fair) compared to DE (moderate and slight) (see Table 3).

Table 3. Interreader comparison of assessed features.

Assessment Features	Kappa bzw. ICC CR	Kappa bzw. ICC DE
Presence of emphysema (yes/no)	0.693 (substantial)	0.462 (moderate)
Subjective emphysema score (none = 1, mild = 2, moderate = 3, severe = 4)	0.834 (good)	0.809 (good)
Location of maximal emphysema manifestation	0.306 (fair)	0.027 (slight)

ICC = intra-class correlation; CR = conventional radiography, DE = dual energy subtraction radiography.

3.5.2. Presence of Emphysema and Location of the Most Affected Lung Quadrant

Sensitivity as well as specificity for the detection of emphysema was comparable between CR and DE (sensitivity$_{CR}$ 96% and specificity$_{CR}$ 75% vs. sensitivity$_{DE}$ 91% and specificity$_{DE}$ 83%; $p = 0.157$). Similarly, there was no significant difference in the sensitivity or specificity for emphysema localization between CR and DE (sensitivity$_{CR}$ 50% and specificity$_{CR}$ 100% vs. sensitivity$_{DE}$ 57% and specificity$_{DE}$ 100%; $p = 0.157$) (Table 4).

Table 4. Test characteristics of CR and DE.

	Presence of Emphysema		Location of Maximal Emphysema Manifestation	
Assessment parameter	CR	DE	CR	DE
Sensitivity	96.3%	90.7%	50%	57.4%
Specifity	75%	83.33%	100%	100%
NPV	81.82%	66.67%	30.77%	34.29%
PPV	94.55%	96.08%	100%	100%

CR = conventional radiography, DE = dual energy subtraction radiography, NPV = negative predictive value, PPV = positive predictive value.

3.5.3. Severity of Emphysema between CR/DE and CT

The average subjective emphysema score was rated significantly higher in DE (mean: 2.62 $\pm$ 0.87) versus CR (mean: 2.45 $\pm$ 0.89; $p = 0.003$; controls included). Emphysema grading with DE showed a slightly higher correlation with the Goddard score than with CR; these differences, however, were not statistically significant ($r_{DE} = 0.75$ vs. $r_{CR} = 0.68$; $p = 0.108$). Similarly, emphysema grading with DE showed a slightly higher correlation with LAA% than with CR lacking statistical significance ($r_{DE} = 0.73$ vs. $r_{CR} = 0.71$; $p = 0.586$).

4. Discussion

We compared DE to CR for the evaluation of lung emphysema, and found that diagnostic accuracy for the detection, quantification, and localization of emphysema between CR and DE is comparable. The interreader agreement, however, was better with CR compared to DE.

Clinically, PFT is used to diagnose COPD. PFT, however, is relatively insensitive to the severity and distribution of emphysema. (1) There is no correlation between reduced FEV1 and severity of lung emphysema, leading to a wide range in severity of emphysema despite having clinically the same disease stage [25]. (2) Clinical presentation of emphysema does not definitively relate to the distribution of emphysema on imaging [26–29], and upper lung zones are rather silent regions in PFT, leading to a high percentage of patients with mild to moderate disease being missed by PFT [30,31]. (3) FEV1 depends on the patient's cooperation. These points stress the importance of imaging in early stages of COPD. Further, some patients undergo chest X-ray for other clinical questions (i.e., pre-operative evaluation, evaluation of infective consolidation.) without the suspicion of emphysema or signs of COPD. These patients would otherwise not undergo PFT and could be lost.

Conventional imaging, which is often used as baseline imaging, only yields a moderate sensitivity for detecting emphysema (approximately 40%) [32]. This is due to the slight difference in X-ray absorption of pulmonary parenchyma, resulting in low conspicuity of the disease on conventional imaging [33]. DE is a new imaging modality with the potential to overcome these difficulties. In DE, a post-processing algorithm separates calcium-containing structures from soft-tissue components and overcomes the problem of superimposition of several structures [34].

Further, the less penetrating beam with the lower tube voltage used in DE results in a higher dynamic range of resultant image data, higher intrinsic contrast (i.e., lesion's intensity relative to the surrounding tissue intensity), and hence a better depiction of the lung parenchyma and its pathology [35].

In fact, previous studies could show that DE improves the sensitivity for shading lesions, such as the detection of infectious consolidations, tumors, interstitial lung changes, and aortic or tracheal calcification compared to CR images [18–21]. Other studies have shown that DE can reduce diagnostic errors of chest pathologies and prevent misdiagnosis of consolidations or lung nodules, for example, also by less-experienced radiologists [34,36]. The higher accuracy for detecting focal opacities (i.e., lung nodules or infectious infiltrates) was attributed to the better accentuation of lung abnormalities [37]. Since the better intrinsic contrast should yield also higher diagnostic accuracy for hyperlucencent lung pathologies, also called "minus pathologies", we hypothesized that DE-images emphasize emphysematous lung sections in a similar way and, thus, may aid in earlier detection of pulmonary emphysema.

Our results, however, could not show a higher diagnostic accuracy for the detection and localization of emphysema. On the contrary, interreader agreement seemed to be worse with DE, even though the readers had also the standard CR images side by side when evaluating the DE images. We believe that, quite unusual, soft tissue and bone images confused the readers more in their diagnosis than they helped. Therefore, readers might benefit from training in order to get used to the DE images. Further, differences in the depiction of emphysema might be so subtle that there is no measurable clinical benefit in using DE instead of CR. The results are further hampered by the radiologist inexperience to evaluate DE images, reflected in the worse interreader experience compared to CR.

An interesting observation we made in this study concerns the relatively high sensitivity in the detection of emphysema compared to values reported in the literature for CR [32]. This might be due to the lower kV used for the acquisition of DE images compared to conventional CR images. The higher soft tissue contrast with the lower kV used in DE might yield a better distinction of emphysematous lung changes from normal lung parenchyma. Since CR and the DE images were acquired in our study with the lower kV, both CR and DE benefit from the lower kV and had higher sensitivity for emphysema detection. The acquisition of two consecutive X-rays, first with a conventional CR and then with the DE technique, in order to compare the sensitivities between a conventional CR and DE, would have been unethical. However, previous studies have shown that the use of lower tube voltages resulting in lower beam penetration enhances density differences in the lung [38].

Subjective emphysema grading for both CR and DE correlated well with CT. Even though we could observe a slightly better correlation of DE with CT than with CR, differences were not statistically significant.

The downsides of DE are definitely the higher radiation dose, which is only partially compensated on lateral chest radiography and the risk of motion artifacts which can occur when the patient moves between the two image acquisitions [16].

Even if CR holds its position in initial chest evaluation, it insufficiently quantifies regional lung perfusion and emphysema, evaluates fissural integrity, or stimulates the effect of surgical resection. Therefore, to guide therapeutic options in extended emphysema (e.g., lung volume reduction surgery, endobronchial valves, coils), further imaging examinations are essential [39].

Dual-energy CT imaging methods can not only gain anatomical information but also functional information too. For example, lung iodine perfusion blood volume (iPBV) illustrates regional lung perfusion changes [40], or inhalative xenon tracer gas functions as a surrogate for regional lung ventilation. These modalities correlate with the degree of emphysema and could serve as tools for detecting mild emphysema [41]. Nevertheless, lowering the radiation dose by displaying only a target volume would render overall assessment impossible [42].

Limitations of the study are as follows. First, the quantification of emphysema was based on a subjective scoring system, which may limit interreader comparability. Second, the degree of emphysema in CR and DE may be underestimated due to soft tissue overlay, especially in corpulent patients. Third, we did not distinguish between different types of emphysema (centrilobular, panlobular, paraseptal). Forth, uneven ventilation or hyperinflation might affect the detection of emphysema. While reduced ventilation could lead to underestimation of emphysema due to denser lung parenchyma, overinflation on the other hand could potentially lead to an overestimation of emphysema. These factors similarly affect CT densitometry based on HU values. Fifth, we did not perform a correlation of our findings with PFT. Due to the retrospective nature of our study, PFT was not available in the majority of patients. Sixth, in our study, both the CR and the DE images were acquired with the lower kV, meaning that we could not show a difference in sensitivity for emphysema detection related to different kV in our dataset. The acquisition of two consecutive X-rays, first with a conventional CR and then with the DE technique, would have been unethical. However, previous studies have shown that the use of lower tube voltages resulting in lower beam penetration enhances density differences in the lung [38].

5. Conclusions

In conclusion, diagnostic accuracy for the detection, quantification, and localization of emphysema between CR and DE is comparable. This implies that the presumably higher tissue contrast in DE did not have the expected benefit in the evaluation of emphysema. Besides that, interreader agreement was influenced negatively by the evaluation of DE, which we attribute to the unfamiliarity of the readers with the new technique.

An interesting observation we made in this study was the relatively high sensitivity in the detection of emphysema compared to values reported in the literature for CR [32]. This might be due to the lower kV used for the acquisition of DE images compared to conventional CR images, potentially resulting in a better distinction of emphysematous lung changes from normal lung parenchyma.

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

Funding: This research was funded by a grant from the Lunge Zürich, Switzerland.

Institutional Review Board Statement: The study was approved by the institutional review board and local ethics committee (KEK Zürich: Cantonal ethics committee Zurich Switzerland).

Informed Consent Statement: Informed consent was waived because of the retrospective setting of this study (blinded for review).

Data Availability Statement: The study data is not available in a public database. However, data can be requested at the corresponding author.

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

References

1. Snider, G.L. Nosology for our day: Its application to chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **2003**, *167*, 678–683. [CrossRef]
2. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*; Centers for Disease Control and Prevention (US): Atlanta, GA, USA, 2014.
3. Sullivan, S.D.; Ramsey, S.D.; Lee, T.A. The economic burden of COPD. *Chest* **2000**, *117*, 5S–9S. [CrossRef]
4. Calverley, P.M.; Walker, P. Chronic obstructive pulmonary disease. *Lancet* **2003**, *362* (Suppl. 2), 1053–1061. [CrossRef]
5. Plantier, L.; Boczkowski, J.; Crestani, B. Defect of alveolar regeneration in pulmonary emphysema: Role of lung fibroblasts. *Int. J. Chronic Obstr. Pulm. Dis.* **2007**, *2*, 463–469.
6. Gorbunova, V.; Jacobs, S.S.A.M.; Lo, P.; Dirksen, A.; Nielsen, M.; Bab-Hadiashar, A.; de Bruijne, M. Early Detection of Emphysema Progression. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2010 13th International Conference, Beijing, China, 20–24 September 2010; Volume 13, pp. 193–200.
7. Mohsen, L.A.; Gawad, E.A.A.; Ibrahiem, M.A. CT quantification of emphysema: Is semi-quantitative scoring a reliable enough method? *Egypt. J. Radiol. Nucl. Med.* **2014**, *45*, 673–678. [CrossRef]
8. Irion, K.L.; Marchiori, E.; Hochhegger, B.; da Silva Porto, N.; da Silva Moreira, J.; Anselmi, C.E.; Allen Holemans, J.; Irion, P.O. CT Quantification of Emphysema in Young Subjects with No Recognizable Chest Disease. *Am. J. Roentgenol.* **2008**, *192*, W90–W96. [CrossRef] [PubMed]
9. Madani, A.; Keyzer, C.; Gevenois, P.A. Quantitative computed tomography assessment of lung structure and function in pulmonary emphysema. *Eur. Respir. J.* **2001**, *18*, 720–730. [CrossRef] [PubMed]
10. Thurlbeck, W.M.; Müller, N.L. Emphysema: Definition, imaging, and quantification. *AJR Am. J. Roentgenol.* **1994**, *163*, 1017–1025. [CrossRef] [PubMed]
11. Alkadhi, H.; Marincek, B.; Stolzmann, P. Dual-Energy Bildgebung Nicht nur fürs Handgepäck. *Schweiz. Med. Forum. Schlaglichter* **2008**, *8*, 1019–1020.
12. Welte, T.; Vogelmeier, C.; Papi, A. COPD: Early diagnosis and treatment to slow disease progression. *Int. J. Clin. Pract.* **2015**, *69*, 336–349. [CrossRef]
13. Frauenfelder, T.; Nguyen, T.D.L.; Delaloye, B. CT der Lunge: Von der morphologischen Darstellung zur Quantifizierung. *Swiss Med. Forum* **2013**, 686–688. [CrossRef]
14. Wells, J.M.; Washko, G.R.; Han, M.K.; Abbas, N.; Nath, H.; Mamary, A.J.; Regan, E.; Bailey, W.C.; Martinez, F.J.; Westfall, E.; et al. Pulmonary arterial enlargement and acute exacerbations of COPD. *N. Engl. J. Med.* **2012**, *367*, 913–921. [CrossRef] [PubMed]
15. Santos, S.; Peinado, V.I.; Ramírez, J.; Melgosa, T.; Roca, J.; Rodriguez-Roisin, R.; Barberà, J.A. Characterization of pulmonary vascular remodelling in smokers and patients with mild COPD. *Eur. Respir. J.* **2002**, *19*, 632–638. [CrossRef] [PubMed]
16. Kuhlman, J.E.; Collins, J.; Brooks, G.N.; Yandow, D.R.; Broderick, L.S. Dual-energy subtraction chest radiography: What to look for beyond calcified nodules. *Radiographics* **2006**, *26*, 79–92. [CrossRef] [PubMed]
17. MacMahon, H.; Armato, S.G. Temporal subtraction chest radiography. *Eur. J. Radiol.* **2009**, *72*, 238–243. [CrossRef] [PubMed]
18. Martini, K.; Baessler, M.; Baumueller, S.; Frauenfelder, T. Diagnostic accuracy and added value of dual-energy subtraction radiography compared to standard conventional radiography using computed tomography as standard of reference. *PLoS ONE* **2017**, *12*, e0174285. [CrossRef]
19. Gilkeson, R.C.; Novak, R.D.; Sachs, P. Digital radiography with dual-energy subtraction: Improved evaluation of cardiac calcification. *AJR Am. J. Roentgenol.* **2004**, *183*, 1233–1238. [CrossRef] [PubMed]
20. Fischbach, F.; Freund, T.; Röttgen, R.; Engert, U.; Felix, R.; Ricke, J. Dual-energy chest radiography with a flat-panel digital detector: Revealing calcified chest abnormalities. *AJR Am. J. Roentgenol.* **2003**, *181*, 1519–1524. [CrossRef]
21. Kelcz, F.; Zink, F.E.; Peppler, W.W.; Kruger, D.G.; Ergun, D.L.; Mistretta, C.A. Conventional chest radiography vs dual-energy computed radiography in the detection and characterization of pulmonary nodules. *AJR Am. J. Roentgenol.* **1994**, *162*, 271–278. [CrossRef]
22. Kimura, T.; Kawakami, T.; Kikuchi, A.; Ooev, R.; Akiyama, M.; Horikoshi, H. A Study on Diagnostic Assist Systems of Chronic Obstructive Pulmonary Disease from Medical Images by Deep Learning. *J. Comput. Commun.* **2018**, *6*, 21–31. [CrossRef]
23. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
24. Portney, L.G.; Watkins, M.P. *Foundations of Clinical Research: Applications to Practice*, 3rd ed.; Pearson/Prentice Hall: Washington, DC, USA, 2009.
25. Makita, H.; Nasuhara, Y.; Nagai, K.; Ito, Y.; Hasegawa, M.; Betsuyaku, T.; Onodera, Y.; Hizawa, N.; Nishimura, M.; Group, H.C.C.S. Characterisation of phenotypes based on severity of emphysema in chronic obstructive pulmonary disease. *Thorax* **2007**, *62*, 932–937. [CrossRef]
26. Benoit, T.M.; Straub, G.; von Garnier, C.; Franzen, D. Schweres Lungenemphysem: Noch lange kein Endbahnhof! *Swiss Med. Forum* **2020**, *20*, 7–11. [CrossRef]
27. Celli, B.R.; Cote, C.G.; Marin, J.M.; Casanova, C.; Montes de Oca, M.; Mendez, R.A.; Pinto Plata, V.; Cabral, H.J. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N. Engl. J. Med.* **2004**, *350*, 1005–1012. [CrossRef]

28. Mair, G.; Miller, J.J.; McAllister, D.; Maclay, J.; Connell, M.; Murchison, J.T.; MacNee, W. Computed tomographic emphysema distribution: Relationship to clinical features in a cohort of smokers. *Eur. Respir. J.* **2009**, *33*, 536–542. [CrossRef]
29. de Torres, J.P.; Bastarrika, G.; Zagaceta, J.; Sáiz-Mendiguren, R.; Alcaide, A.B.; Seijo, L.M.; Montes, U.; Campo, A.; Zulueta, J.J. Emphysema presence, severity, and distribution has little impact on the clinical presentation of a cohort of patients with mild to moderate COPD. *Chest* **2011**, *139*, 36–42. [CrossRef]
30. Gurney, J.W.; Jones, K.K.; Robbins, R.A.; Gossman, G.L.; Nelson, K.J.; Daughton, D.; Spurzem, J.R.; Rennard, S.I. Regional distribution of emphysema: Correlation of high-resolution CT with pulmonary function tests in unselected smokers. *Radiology* **1992**, *183*, 457–463. [CrossRef]
31. Šileikienė, V.; Urbonas, M.; Matačiūnas, M.; Norkūnienė, J. Relationships between pulmonary function test parameters and quantitative computed tomography measurements of emphysema in subjects with chronic obstructive pulmonary disease. *Acta Med. Litu.* **2017**, *24*, 209–218. [CrossRef]
32. Thurlbeck, W.M.; Simon, G. Radiographic appearance of the chest in emphysema. *AJR Am. J. Roentgenol.* **1978**, *130*, 429–440. [CrossRef] [PubMed]
33. Linan, D.; Jun, L.; Wushuai, J.; Lu, Z.; Mingschu, W.; Hongli, S.; Shugian, L. Emphysema early diagnosis using X-reay diffraction enhanced imaging at synchrotron light source. *BioMed. Eng. OnLine* **2014**, *13*, 82.
34. Gezer, M.C.; Algin, O.; Durmaz, A.; Arslan, H. Efficiency and reporting confidence analysis of sequential dual-energy subtraction for thoracic X-ray examinations. *Qatar Med. J.* **2019**, *2019*, 9. [CrossRef] [PubMed]
35. Huda, W.; Abrahams, R.B. Radiographic techniques, contrast, and noise in X-ray imaging. *AJR Am. J. Roentgenol.* **2015**, *204*, W126–W131. [CrossRef]
36. Del Ciello, A.; Franchi, P.; Contegiacomo, A.; Cicchetti, G.; Bonomo, L.; Larici, A.R. Missed lung cancer: When, where, and why? *Diagn Interv. Radiol.* **2017**, *23*, 118–126. [CrossRef] [PubMed]
37. MacMahon, H. Improvement in detection of pulmonary nodules: Digital image processing and computer-aided diagnosis. *Radiographics* **2000**, *20*, 1169–1177. [CrossRef] [PubMed]
38. Schaefer-Prokop, C.; Neitzel, U.; Venema, H.W.; Uffmann, M.; Prokop, M. Digital chest radiography: An update on modern technology, dose containment and control of image quality. *Eur. Radiol.* **2008**, *18*, 1818–1830. [CrossRef]
39. McKenna, R.J.; Brenner, M.; Fischel, R.J.; Singh, N.; Yoong, B.; Gelb, A.F.; Osann, K.E. Patient selection criteria for lung volume reduction surgery. *J. Thorac. Cardiovasc. Surg.* **1997**, *114*, 957–964; discussion 964–967. [CrossRef]
40. Lee, C.W.; Seo, J.B.; Lee, Y.; Chae, E.J.; Kim, N.; Lee, H.J.; Hwang, H.J.; Lim, C.H. A pilot trial on pulmonary emphysema quantification and perfusion mapping in a single-step using contrast-enhanced dual-energy computed tomography. *Investig. Radiol.* **2012**, *47*, 92–97. [CrossRef]
41. Koike, H.; Sueyoshi, E.; Sakamoto, I.; Uetani, M. Quantification of Lung Perfusion Blood Volume by Dual-Energy CT in Patients With and Without Chronic Obstructive Pulmonary Disease. *J. Belg. Soc. Radiol.* **2015**, *99*, 62–68. [CrossRef]
42. Lu, G.M.; Zhao, Y.; Zhang, L.J.; Schoepf, U.J. Dual-energy CT of the lung. *AJR Am. J. Roentgenol.* **2012**, *199*, S40–S53. [CrossRef]

diagnostics

MDPI

Systematic Review

The Added Effect of Artificial Intelligence on Physicians' Performance in Detecting Thoracic Pathologies on CT and Chest X-ray: A Systematic Review

Dana Li [1,2,*], Lea Marie Pehrson [1], Carsten Ammitzbøl Lauridsen [1,3], Lea Tøttrup [4], Marco Fraccaro [4], Desmond Elliott [5], Hubert Dariusz Zając [5], Sune Darkner [5], Jonathan Frederik Carlsen [1] and Michael Bachmann Nielsen [1,2]

[1] Department of Diagnostic Radiology, Copenhagen University Hospital, Rigshospitalet, 2100 Copenhagen, Denmark; Lea.marie.pehrson@regionh.dk (L.M.P.); Carsten.ammitzboel.lauridsen.01@regionh.dk (C.A.L.); Jonathan.frederik.carlsen@regionh.dk (J.F.C.); Mbn@dadlnet.dk (M.B.N.)
[2] Department of Clinical Medicine, University of Copenhagen, 2100 Copenhagen, Denmark
[3] Department of Technology, Faculty of Health and Technology, University College Copenhagen, 2200 Copenhagen, Denmark
[4] Unumed Aps, 1055 Copenhagen, Denmark; Leat@unumed.com (L.T.); Mf@unumed.com (M.F.)
[5] Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark; De@di.ku.dk (D.E.); Hdz@di.ku.dk (H.D.Z.); Darkner@di.ku.dk (S.D.)
* Correspondence: Dana.li@regionh.dk

Citation: Li, D.; Pehrson, L.M.; Lauridsen, C.A.; Tøttrup, L.; Fraccaro, M.; Elliott, D.; Zając, H.D.; Darkner, S.; Carlsen, J.F.; Nielsen, M.B. The Added Effect of Artificial Intelligence on Physicians' Performance in Detecting Thoracic Pathologies on CT and Chest X-ray: A Systematic Review. *Diagnostics* **2021**, *11*, 2206. https://doi.org/10.3390/diagnostics11122206

Academic Editors: Sameer Antani and Sivaramakrishnan Rajaraman

Received: 20 October 2021
Accepted: 23 November 2021
Published: 26 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: Our systematic review investigated the additional effect of artificial intelligence-based devices on human observers when diagnosing and/or detecting thoracic pathologies using different diagnostic imaging modalities, such as chest X-ray and CT. Peer-reviewed, original research articles from EMBASE, PubMed, Cochrane library, SCOPUS, and Web of Science were retrieved. Included articles were published within the last 20 years and used a device based on artificial intelligence (AI) technology to detect or diagnose pulmonary findings. The AI-based device had to be used in an observer test where the performance of human observers with and without addition of the device was measured as sensitivity, specificity, accuracy, AUC, or time spent on image reading. A total of 38 studies were included for final assessment. The quality assessment tool for diagnostic accuracy studies (QUADAS-2) was used for bias assessment. The average sensitivity increased from 67.8% to 74.6%; specificity from 82.2% to 85.4%; accuracy from 75.4% to 81.7%; and Area Under the ROC Curve (AUC) from 0.75 to 0.80. Generally, a faster reading time was reported when radiologists were aided by AI-based devices. Our systematic review showed that performance generally improved for the physicians when assisted by AI-based devices compared to unaided interpretation.

Keywords: artificial intelligence; deep learning; computer-based devices; radiology; thoracic diagnostic imaging; chest X-ray; CT; observer tests; performance

1. Introduction

Artificial intelligence (AI)-based devices have made significant progress in diagnostic imaging segmentation, detection, and disease differentiation, as well as prioritization. AI has emerged as the cutting-edge technology to bring diagnostic imaging into the future [1]. AI may be used as a decision support system, where radiologists reject or accept the algorithm's diagnostic suggestions, which was investigated in this review, but there is no AI-based device that fully autonomously diagnose or classify findings in radiology yet. Some products have been developed for the purpose of radiological triage [2]. Triage and notification of a certain finding have been a task that has had some autonomy since there is no clinician assigned to re-prioritize the algorithm's suggestions. Other uses of

AI algorithms could be suggestion of treatment options based on disease specific predictive factors [3] and automatic monitoring and overall survival prognostication to aid the physician in deciding the patient's future treatment plan [4].

The broad application of plain radiography in thoracic imaging and the use of other modalities, such as computed tomography (CT), to delineate abnormalities adds to the number of imaging cases that can provide information to successfully train an AI-algorithm [5]. In addition to providing large quantities of data, chest X-ray is one of the most used imaging modalities. Thoracic imaging has, therefore, not only a potential to provide a large amount of data for developing AI-algorithms successfully, but there is also potential for AI-based devices to be useful in a great number of cases. Because of this, several algorithms in thoracic imaging have been developed—most recently in the diagnosis of COVID-19 [6].

AI has attracted increasing attention in diagnostic imaging research. Most studies demonstrate their AI-algorithm's diagnostic superiority by separately comparing the algorithm's diagnostic accuracy to the accuracy achieved by manual reading [7,8]. Nevertheless, several factors seem to prevent AI-based devices from diagnosing pathologies in radiology without human involvement [9], and only few studies conduct observer tests where the algorithm is being used as a second or concurrent reader to radiologists: a scenario closer to a clinical setting [10,11]. Even though diagnostic accuracy of an AI-based device can be evaluated by testing it independently, this may not reflect the true clinical effect of adding AI-based devices, since such testing eliminates the factor of human-machine interaction and final human decision making.

Our systematic review investigated the additional effect AI-based devices had on physicians' abilities when diagnosing and/or detecting thoracic pathologies using different diagnostic imaging modalities, such as chest X-ray and CT.

2. Materials and Methods

2.1. Literature Search Strategy

The literature search was completed on 24 March 2021, from 5 databases: EMBASE, PubMed, Cochrane library, SCOPUS, and Web of Science. The search was restricted to peer-reviewed publications of original research written in English from 2001–2021, both years included.

The following specific MESH terms were used in PubMed: "thorax", "radiography, thoracic", "lung", "artificial intelligence", "deep Learning", "machine Learning", "neural networks, computer", "physicians", "radiologists", "workflow", "physicians". MESH terms were combined with the following all-fields specific search words and their bended forms: "thorax", "chest", "lung", "AI", "artificial intelligence", "deep learning", "machine learning", "neural networks", "computer", "computer neural networks", "clinician", "physician", "radiologist", "workflow".

To perform the EMBASE search, the following combination of text word search and EMTREE terms were used: ("thorax" (EMTREE term) OR "lung" (EMTREE term) OR "chest" OR "lung" OR "thorax") AND ("artificial intelligence (EMTREE term) OR "machine learning" (EMTREE term) OR "deep learning" (EMTREE term) OR "convolutional neural network" (EMTREE term) OR "artificial neural network" (EMTREE term) OR "ai" OR "artificial intelligence" OR "neural network" OR "deep learning" OR "machine learning") AND ("radiologist (EMTREE term) OR " physician" (EMTREE term) OR "clinician" (EMTREE term) OR "workflow" (EMTREE term) OR "radiologist" OR "clinician" OR "physician" OR "workflow").

We followed the PRISMA guidelines for literature search and study selection. After removal of duplicates, all titles and abstracts retrieved from the search were independently screened by two authors (D.L. and L.M.P.). In case of unresolved disagreements, that could not be determined by consensus vote between D.L. and L.M.P., a third author (J.F.C.) was appointed to assess and resolve the disagreement. Data were extracted by D.L. and L.M.P. using pre-piloted forms. To describe the performance of the radiologists without and with assistance of AI-based devices, we used a combination of narrative synthesis and compared

measures of accuracy, area under the ROC curve (AUC), sensitivity, specificity, and time measurements.

For evaluating the risk of bias and assess quality of research, we used the QUADAS-2 tool [12].

2.2. Study Inclusion Criteria

Peer-reviewed original research articles published in English, between 2001 and 2021, were reviewed for inclusion. Inclusion criteria were set at follows:

1. AI-based devices, either independent or incorporated into a workflow, used for imaging diagnosis and/or detection of findings in lung tissue, regardless of thoracic imaging modality; and
2. an observer test where radiologists or other types of physicians used the AI-algorithm as either a concurrent or a second reader; and
3. within the observer test, the specific observer that diagnosed/detected the findings without AI-assistance must also participate as the observer with AI-assistance; and
4. outcome measurements of observer tests included either sensitivity, specificity, AUC, accuracy, or some form of time measurement recording observers' reading time without and with AI-assistance.

Studies where one set of physicians, with the aid of AI, retrospectively re-evaluate another set of physicians' diagnoses without AI were excluded. AI-based devices that did not detect specific pulmonary tissue findings/pathology, e.g., rib fracture, aneurisms, thyroid enlargements etc. were also excluded.

3. Results

We included a total of 38 studies [13–50] in our systematic review. The QUADAS-2 tool is presented in Figure 1, and a PRISMA flowchart of the literature search is presented in Figure 2.

We divided the studies into two groups: The first group, consisting of 19 studies [13–31], used an AI-based device as a concurrent reader in an observer test, where the observers were tasked with diagnosing images with assistance from an AI-based device, while not being allowed (blinded) to see their initial diagnosis made without assistance from AI (Table 1a). The second group, consisting of 20 studies [19,32–50] used the AI-based device as a second reader in an un-blinded sequential observer test, thus allowing observers to see and change their original un-assisted diagnosis (Table 1b).

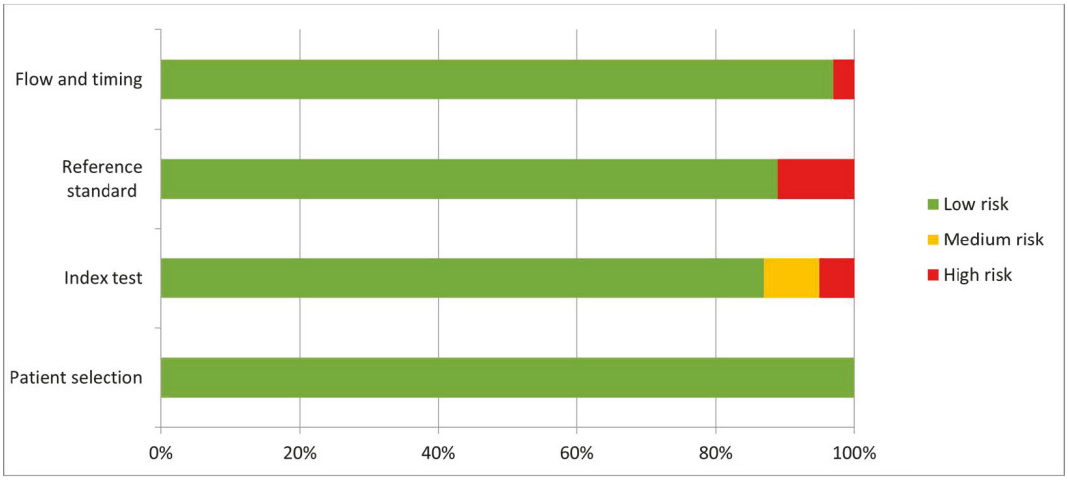


Figure 1. The QUADAS-2 tool for evaluating risk of bias and assess quality of research.

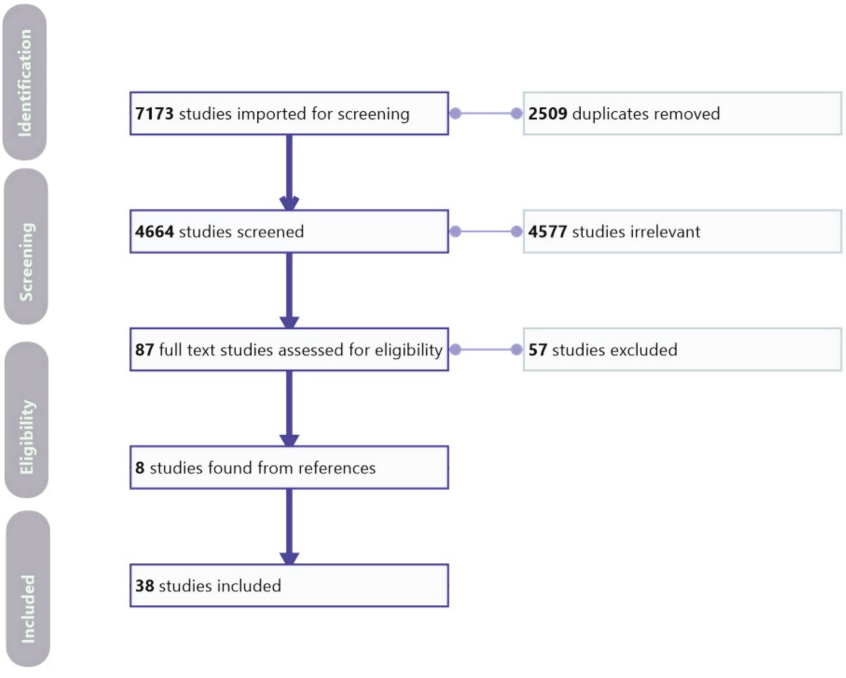


Figure 2. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) flowchart of the literature search and study selection.

Table 1. (**a**) Included studies with artificial intelligence-based devices as concurrent readers in the observer test. (**b**) Included studies with artificial intelligence-based devices in an observer test with a sequential test design.

Author	Year	Standard of Reference	Type of Artificial Intelligence-Based CAD	Pathology	No. of Cases	Test Observers	Image Modality
			a				
Bai et al. [13]	2021	RT-PCR	EfficientNet-B3 Convolutional Neural Network	COVID-19 pneumonia	119	6 radiologists (10–20 years of chest CT experience)	CT
Beyer et al. [19]	2007	Radiologist identified and consensus vote	Commercially available (LungCAD prototype version, Siemens Corporate Research, Malvern, PA, USA)	Pulmonary nodules	50	4 radiologists (2–11 years experience)	CT
de Hoop et al. [20]	2010	Histologically confirmed	Commercially available (OnGuard 5.0; Riverain Medical, Miamisburg, OH, USA)	Pulmonary nodules	111	1 general radiologist, 1 chest radiologist, and 4 residents	Chest X-ray
Dorr et al. [14]	2020	RT-PCR	DenseNet 121 architecture	COVID-19 pneumonia	60	23 radiologists and 31 emergency care physicians	Chest X-ray
Kim et al. [15]	2020	Bacterial culture and RT-PCR for viruses	Commercially available (Lunit INSIGHT for chest radiography, version 4.7.2; Lunit, Seoul, South Korea)	Pneumonia	387	3 emergency department physicians (6–7 years experience)	Chest X-ray
Koo et al. [21]	2020	Pathologically confirmed	Commercially available (Lunit Insight CXR, ver. 1.00; Lunit, Seoul, South Korea)	Pulmonary nodules	434	2 thoracic radiologists and 2 residents	Chest X-ray

Table 1. *Cont.*

Author	Year	Standard of Reference	Type of Artificial Intelligence-Based CAD	Pathology	No. of Cases	Test Observers	Image Modality
			a				
Kozuka et al. [22]	2020	Radiologist identified and majority vote	Faster Region-Convolutional Neural Network	Pulmonary nodules	120	2 radiologists (1–4 years experience)	CT
Lee et al. [23]	2012	Pathologically confirmed	Commercially available (IQQA-Chest, EDDA Technology, Princeton Junction, NJ, USA)	Pulmonary nodules malignant/benign	200	5 chest radiologists and 5 residents	Chest X-ray
Li et al. [24]	2011	CT	Commercially available (SoftView, version 2.0; Riverrain Medical, Miamisburg, OH, USA-Image normalization, feature extraction and regression networks)	Pulmonary nodules	151	3 radiologists (10–25 years experience)	Chest X-ray
Li et al. [25]	2011	Pathologically confirmed and radiology assessed	Commercially available (SoftView, version 2.0; Riverain Medical)	Pulmonary nodules	80	2 chest radiologists, 4 general radiologists, and 4 residents	Chest X-ray
Liu et al. [16]	2020	-	Segmentation model with class attention map including a residual convolutional block	COVID-19 pneumonia	643	-	Chest X-ray
Liu et al. [26]	2019	Radiologist identified and majority vote	DenseNet and Faster Region-Convolutional Neural Network	Pulmonary nodule	271	2 radiologists (10 years experience)	CT
Martini et al. [27]	2021	Radiologist consensus	Commercially available (ClearRead-CT, Riverrain Technologies, Miamisburg, OH, USA)	Pulmonary consolidations/nodules	100	2 senior radiologists, 2 final-year residents, and 2 inexperienced residents	MDCT
Nam et al. [29]	2021	RT-PCR and CT	Deep learning-based algorithm (Deep convolutional neural network)	Pneumonia, pulmonary edema, active tuberculosis, interstitial lung disease, nodule/mass, pleural effusion, acute aortic syndrome, pneumoperitoneum, rib fracture, pneumothorax, mediastinal mass.	202	2 thoracic radiologists, 2 board-certified radiologists, and 2 residents	Chest X-ray
Rajpurkar et al. [31]	2020	Positive culture or Xpert MTB/RIF test	Convolutional Neural Network	Tuberculosis	114	13 physicians (6 months–25 years of experience)	Chest X-ray
Singh et al. [28]	2021	Radiologically reviewed	Commercially available (ClearRead CT Vessel Suppression and Detect, Riverain Technologies TM)	Subsolid nodules (Incl ground-glass and/or part-solid)	123	2 radiologists (5–10 years experience)	CT
Sung et al. [30]	2021	CT and clinical information	Commercially available (Med-Chest X-ray system (version 1.0.0, VUNO, Seoul, South Korea)	Nodules, consolidation, interstitial opacity, pleural effusion, pneumothorax	128	2 thoracic radiologists, 2 board-certified radiologists, 1 radiology resident, and 1 non-radiology resident	Chest X-ray
Yang et al. [17]	2021	RT-PCR	Deep Neural Network	COVID-19 pneumonia	60	3 radiologists (5–20 years experience)	CT
Zhang et al. [18]	2021	RT-PCR	Deep Neural Network using the blur processing method to improve the image enhancement algorithm	COVID-19 pneumonia	15	2 physicians (13–15 years experience)	CT

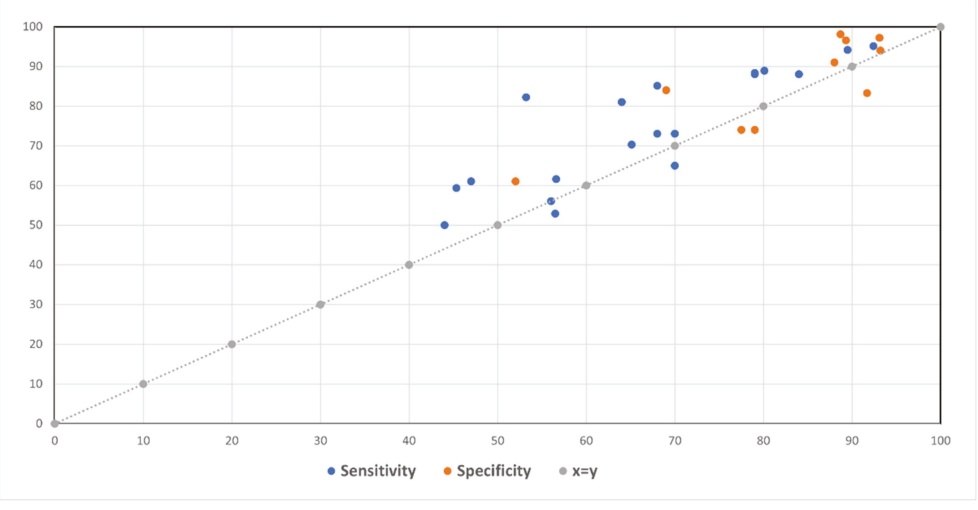
Table 1. *Cont.*

Author	Year	Standard of Reference	Type of Artificial Intelligence-Based CAD	Pathology	No. of Cases	Test Observers	Image Modality
			b				
Abe et al. [47]	2004	Radiological review and clinical correlation	Single three-layer, feed-forward Artificial Neural Network with a back-propagation algorithm	Sarcoidosis, miliary tuberculosis, lymphangitic carcinomatosis, interstitial pulmonary edema, silicosis, scleroderma, P. Carinii pneumonia, Langerhals cell histiocytosis, idiopathic pulmonary fibrosis, viral pneumonia, pulmonary drug toxicity	30	5 radiologists (6–18 years experience)	Chest X-ray
Abe et al. [48]	2003	Radiology consensus	Fourier transformation and Artificial Neural Network	Detection of interstitial lung disease	20	8 chest radiologists, 13 other radiologists, and 7 residents	Chest X-ray
		Clinical correlation and bacteriological	Artificial Neural Network	Differential diagnosis of 11 types of interstitial lung disease	28	16 chest radiologists, 25 other radiologists, and 12 residents	Chest X-ray
		Pathology	Artificial Neural Network	Distinction between malignant and benign pulmonary nodules	40	7 chest radiologists, 14 other radiologists, and 7 residents	Chest X-ray
Awai et al. [33]	2004	Radiological review	Artificial Neural Network	Pulmonary nodules	50	5 board-certified radiologists and 5 residents	CT
Awai et al. [32]	2006	Histology	Neural Network	Pulmonary nodules malignant/benign	33	10 board-certified radiologists and 9 radiology residents	CT
Beyer et al. [19]	2007	Radiologist identified and consensus vote	Commercially available (LungCAD prototype version, Siemens Corporate Research, Malvern, PA, USA)	Pulmonary nodules	50	4 radiologists (2–11 years experience)	CT
Bogoni et al. [34]	2012	Majority of agreement	Commercially available (Lung CAD VC20A, Siemens Healthcare, Malvern, PA, USA)	Pulmonary nodules	43	5 fellowship-trained chest radiologists (1–10 years experience)	CT
Chae et al. [35]	2020	Pathologically confirmed and radiologically reviewed	CT-lungNET (Deep Convolutional Neural Network)	Pulmonary nodules	60	2 medical students, 2 residents, 2 non-radiology physicians, and 2 thoracic radiologists	CT
Chen et al. [36]	2007	Surgery or biopsy	Deep Neural Network	Pulmonary nodules malignant/benign	60	3 junior radiologists, 3 secondary radiologists, and 3 senior radiologists	CT
Fukushima et al. [49]	2004	Pathological, bacteriological and clinical correlation	Single three-layer, feed-forward Artificial Neural Network with a back-propagation algorithm	Sarcoidose, diffuse panbronchioloitis, nonspecific interstitial pneumonia, lymphangitic carcinomatosis, usual interstitial pneumonia, silicosis, BOOP or chronic eopsinophilic pneumonia, pulmonary alveolar proteinosis, miliary tuberculosis, lymphangiomyomatosis, P, carinii pneumonia or cytomegalovirus pneumonia	130	4 chest radiologists and 4 general radiologists	High Resolution CT

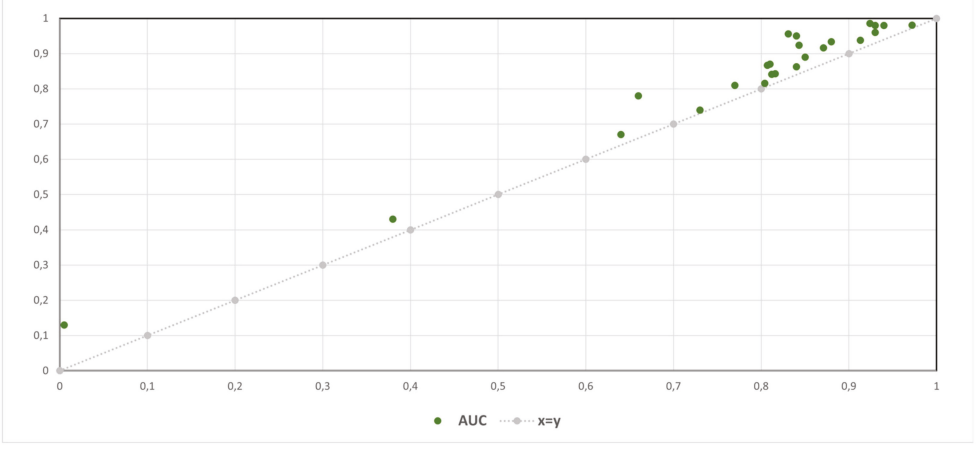
Table 1. *Cont.*

Author	Year	Standard of Reference	Type of Artificial Intelligence-Based CAD	Pathology	No. of Cases	Test Observers	Image Modality
			b				
Hwang et al. [50]	2019	Pathology, clinical or radiological	Deep Convolutional Neural Network with dense blocks	4 different target diseases (pulmonary malignant neoplasms, tuberculosis, pneumonia, pneumothorax) classified in to binary classification of normal/abnormal	200	5 thoracic radiologists, board-certified radiologists, and 5 non-radiology physicians	Chest X-ray
Kakeda et al. [41]	2004	CT	Commercially available (Trueda, Mitsubishi Space Software, Tokyo, Japan)	Pulmonary nodules	90	4 board-certified radiologists and 4 residents	Chest X-ray
Kasai et al. [40]	2008	CT	Three Artificial Neural Networks	Pulmonary nodules	41	6 chest radiologists and 12 general radiologists	Lateral chest X-ray only
Kligerman et al. [42]	2013	Histology and CT	Commercially available (OnGuard 5.1; Riverain Medical, Miamisburg, OH, USA)	Lung cancer	81	11 board-certified general radiologists (1–24 years experience)	Chest X-ray
Liu et al. [37]	2021	Histology, CT, and biopsy/surgical removal	Convolutional Neural Networks	Pulmonary nodules malignant/benign	879	2 senior chest radiologists, 2 secondary chest radiologists, and 2 junior radiologists	CT
Matsuki et al. [38]	2001	Pathology and radiology	Three-layer, feed-forward Artificial Neural Network with a back-propagation algorithm	Pulmonary nodules	50	4 attending radiologists, 4 radiology fellows, 4 residents	High Resolution CT
Nam et al. [43]	2019	Pathologically confirmed and radiologically reviewed	Deep Convolutional Neural Networks with 25 layers and 8 residual connections	Pulmonary nodules malignant/benign	181	4 thoracic radiologists, 5 board-certified radiologists, 6 residents, and 3 non-radiology physicians	Chest X-ray
Oda et al. [44]	2009	Histology, cytology, and CT	Massive training Artificial Neural Network	Pulmonary nodules	60	7 board-certified radiologists and 5 residents	Chest X-ray
Rao et al. [39]	2007	Consensus and majority vote	LungCAD	Pulmonary nodules	196	17 board-certified radiologists	MDCT
Schalekamp et al. [45]	2014	Radiologically reviewed, pathology and clinical correlation	Commercially available (ClearRead +Detect 5.2; Riverain Technologies and ClearRead Bone Suppression 2.4; Riverain Technologies)	Pulmonary nodules	300	5 radiologists and 3 residents	Chest X-ray
Sim et al. [46]	2020	Biopsy, surgery, CT, and pathology	Commercially available (ALND, version 1.00; Samsung Electronics, Suwon, South Korea)	Cancer nodules	200	5 senior chest radiologists, 4 chest radiologists, and 3 residents	Chest X-ray

Visual summaries of the performance change in sensitivity, specificity, and AUC for all studies are shown in Figure 3a,b.



(a)



(b)

Figure 3. Sensitivity and specificity (**a**) and AUC (**b**) without and with the aid of an AI-based device.

3.1. Studies Where Human Observers Used AI-Based Devices as Concurrent Readers

In 19 studies observers were first tasked to diagnose the image without an AI-based device. After a washout period, the same observers were then tasked to diagnose the images again. They were not allowed to see and change their original un-aided radiological diagnosis before making their diagnosis aided by and AI-based device (Table 1a). The results of the observer tests are listed in Table 2a–c for concurrent reader studies.

Table 2. Sensitivity and specificity (**a**); accuracy and AUC (**b**); and time measurement results (**c**) for observer tests without and with AI-based devices as a concurrent reader.

Author	Without AI-Based CAD		With AI-Based CAD		Change	Statistical Significance between Difference
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)		
			a			
Bai et al. [13]	79	88	88	91	↑	$p < 0.001$
Beyer et al. [19]	56.5	-	61.6	-	↑	$p < 0.001$
de Hoop et al. [20]	56 *	-	56 *	-	↑	-
Dorr et al. [14]	47	79	61	75	↑	$p < 0.007$
Kim et al. [15]	73.9	88.7	82.2	98.1	↑	$p < 0.014$
Koo et al. [21]	92.4	93.1	95.1	97.2	↑	-
Kozuka et al. [22]	68	91.7	85.1	83.3	↑	$p < 0.01$ **
Lee et al. [23]	84	-	88	-	↑	-
Rajpurkar et al. [31]	70	52	73	61	↑	-
Singh et al. [28]	68 *	77.5 *	73 *	74 *	↑	-
Sung et al. [30]	80.1	89.3	88.9	96.6	↑	$p < 0.01$
Yang et al. [17]	89.5	-	94.2	-	↑	$p < 0.05$
Author	**Without AI-Based CAD**		**With AI-Based CAD**		**Change**	**Statistical Significance between Difference**
	Accuracy (%)	**AUC**	**Accuracy (%)**	**AUC**		
			b			
Bai et al. [13]	85	-	90	-	↑	$p < 0.001$
Kim et al. [15]	-	0.871	-	0.916	↑	$p = 0.002$
Koo et al. [21]	-	0.93	-	0.96	↑	$p < 0.0001$
Li et al. [24]	-	0.840	-	0.863	↑	$p = 0.01$
Li et al. [25]	-	0.807	-	0.867	↑	$p < 0.001$
Liu et al. [26]	-	0.66 *	-	0.78 *	↑	-
Nam et al. [29]	66.3 *	-	82.4 *	-	↑	$p < 0.05$
Rajpurkar et al. [31]	60	-	65	-	↑	$p = 0.002$
Singh et al. [28]	-	0.73 *	-	0.74 *	↑	Not statistically significant
Sung et al. [30]	-	0.93	-	0.98	↑	$p = 0.003$
Yang et al. [17]	94.1	-	95.1	-	↑	$p = 0.01$
Author	**Without AI-Based CAD**		**With AI-Based CAD**		**Change**	**Statistical Significance between Difference**
	Time		**Time**			
			c			
Beyer et al. [19]	294 s (1)		337 s (1)		↓	$p = 0.04$
Kim et al. [15]	165 min (2)		101 min (2)		↑	-
Kozuka et al. [22]	373 min(2)		331 min (2)		↑	-
Liu et al. [16]	100.5 min (3)		34 min (3)		↑	$p < 0.01$
Liu et al. [26]	15 min (1)		5–10 min (1)		↑	-
Martini et al. [27]	194 s (1)		154 s (1)		↑	$p < 0.001$
Nam et al. [29]	2771.2 s * (1)		1916 s * (1)		↑	$p < 0.002$
Sung et al. [30]	24 s (1)		12 s (1)		↑	$p < 0.001$
Zhang et al. [18]	3.623 min (2)		0.744 min (2)		↑	-

a: * our calculated average; ** for sensitivity only; - not applicable; ↑ positive change. b: * our calculated average; - not applicable; ↑positive change. c: (1) per image/case reading time; (2) total reading time for multiple cases; (3) station survey time; * our calculated average; - not applicable; ↑ positive change; ↓ negative change.

3.1.1. Detection of Pneumonia

Bai et al. [13], Dorr et al. [14], Kim et al. [15] Liu et al. [16], Yang et al. [17], and Zhang et al. [18] had AI-based algorithms to detect pneumonia findings of different kinds, e.g., Covid-19 pneumonia from either non-Covid-19 pneumonia or non-pneumonia. Bai et al. [13], Yang et al. [17], Dorr et al. [14], and Zhang et al. [18] investigated detection of Covid-19 pneumonia. Bai et al. [13], Dorr et al. [14], and Yang et al. [17] all had significant improvement in performance measured in sensitivity after being aided by their AI-based devices (Table 2a), and Zhang et al. [18] reported shorter reading time per image but there was not any mention of statistical significance (Table 2c). Liu et al. [16] incorporated an AI-algorithm into a novel emergency department workflow for Covid-19 evaluations: a clinical quarantine station, where some clinical quarantine stations were equipped with AI-assisted image interpretation, and some did not. They compared the overall median survey time at the clinical quarantine stations in each condition and reported statistically significant shortened time (153 min versus 35 min, $p < 0.001$) when AI-assistance was available. Median survey time specific to the image interpretation part of the clinical quarantine station was also significantly shortened (Table 2c), but they did not report if the shortened reading time were accompanied by the same level of diagnostic accuracy. While the previously mentioned studies specifically investigated Covid-19 pneumonia, Kim et al. [15] used AI-assistance to distinguish pneumonia from non-pneumonia and reported significant improvement in performance measured in sensitivity and specificity after AI-assistance (Table 2a).

Detection of Pulmonary Nodules

Beyer et al. [19], de Hoop et al. [20], Koo et al. [21], Kozuka et al. [22], Lee et al. [23], Li et al. [24], Li et al. [25], Liu et al. [26], Martini et al. [27], and Singh et al. [28] used AI-based devices to assist with detection of pulmonary nodules. Even though de Hoop et al. [20] found a slight increase in sensitivity in residents (49% to 51%) and change in radiologists (63% to 61%) for nodule detection, both changes were not statistically significant (Table 2a). In contrast, Koo et al. [21], Li et al. [24], and Li et al. [25] reported improvement of AUC for every individual participating radiologist when using AI-assistance, regardless of experience level (Table 2b). Lee et al. [23] reported improved sensitivity (84% to 88%) when using AI as assistance (Table 2a) but did not mention if the change in sensitivity was significant. However, their reported increase in mean figure of merit (FOM) was statistically significant. Beyer et al. [19] had performed both blinded and un-blinded observer tests; in the blinded, concurrent reader test, radiologists had significant improved sensitivity (56.6% to 61.6%, $p < 0.001$) (Table 2a) but also significantly increased time for reading when assisted by AI (increase of 43 s per image, $p = 0.04$) (Table 2c). Martini et al. [27] reported improved interrater agreement (17–34%) in addition to improved mean reading time (Table 2c), when assisted by AI. Results for the effects of AI assistance on radiologists by Kozuka et al. [22], Liu et al. [26], and Singh et al. [28] are also shown in Table 2a,b, but only Kozuka et al. [22] reported significant improvement (sensitivity from 68% to 85.1%, $p < 0.01$). In addition to change in accuracy, Liu et al. [26] reported a reduction of reading time per patient from 15 min to 5–10 min without mentioning statistical significance.

Detection of Several Different Findings and Tuberculosis

Nam et al. [29] tested an AI-based device in detecting 10 different abnormalities and measured the accuracy by dividing them into groups of urgent, critical, and normal findings. Radiologists significantly improved their detection of critical (accuracy from 29.2% to 70.8%, $p = 0.006$), urgent (accuracy from 78.2% to 82.7%, $p = 0.04$), and normal findings (accuracy from 91.4% to 93.8%, $p = 0.03$). Reading times per reading session were only significantly improved for critical (from 3371.0 s to 640.5 s, $p < 0.001$) and urgent findings (from 2127.1 to 1840.3, $p < 0.001$) but significantly prolonged for normal findings (from 2815.4 s to 3267.1 s, $p < 0.001$). Even though Sung et al. [30] showed overall improvement

in detection (Table 2a–c), per-lesion sensitivity only improved in residents (79.7% to 86.7%, $p = 0.006$) and board-certified radiologists (83.0% to 91.2%, $p < 0.001$) but not in thoracic radiologists (86.4% to 89.4%, $p = 0.31$). Results from a study by Rajpurkar et al. [31] for the effects of AI-assistance on radiologists detecting tuberculosis show that there were significant improvement in both sensitivity, specificity, and accuracy when aided by AI (Table 2a,b).

3.2. Studies Where Human Observers Used AI-Based Devices as a Second Reader in a Sequential Observer Test Design

In 20 studies, observers were first tasked to diagnose the image without an AI-based device. Immediately afterwards, they were tasked to diagnose the images aided by an AI-based device and were also allowed to see and change their initial diagnosis (Table 1b). The results of the observer tests are listed in Table 3a–c for sequential observer test design studies.

3.2.1. Detection of Pulmonary Nodules Using CT

A total of 16 studies investigated the added value of AI on observers in the detection of pulmonary nodules; nine studies [19,32–39] used CT scans, and seven studies [40–46] used chest X-rays (Table 1b). Although Awai et al. [33], Liu et al. [37], and Matsuki et al. [38] showed statistically significant improvement across all radiologists (Table 3b) when using AI, other studies reported only significant increase in a sub-group of their test observers. Awai et al. [32] and Chen et al. [36] reported only significant improvement in the groups with the more junior radiologists; Awai et al. [32] reported an AUC from 0.768 to 0.901 ($p = 0.009$) in residents but no significant improvement in the board-certified radiologists (AUC 0.768 to 0.901, $p = 0.19$), and Chen et al. [36] reported an AUC from 0.76 to 0.96 ($p = 0.0005$) in the junior radiologists and 0.85 to 0.94 ($p = 0.014$) in the secondary radiologists but no significant improvement in the senior radiologists (AUC 0.91 to 0.96, $p = 0.221$). In concordance, Chae et al. [35] only reported significant improvement in the non-radiologists (AUC from 0.03 to 0.19, $p < 0.05$) but not for the radiologists (AUC from −0.02 to 0.07). While the results from Bogoni et al. [34] confirm the results from Beyer et al.'s [19] concurrent observer test, Beyer et al. [19] showed in the sequential observer test the opposite: decreased sensitivity (56.5 to 52.9, $p < 0.001$) with shortened reading time (294 s to 274 s per image, $p = 0.04$) (Table 3a,c). In addition to overall increase in accuracy (Table 3b), Rao et al. [39] also reported that using AI resulted in greater number of positive actionable management (averaged 24.8 patients), i.e., recommendations for additional images and/or biopsy, that were missed without AI.

3.2.2. Detection of Pulmonary Nodules Using Chest X-ray

As with detection of pulmonary nodules using CT, there were also contrasting results regarding radiologist experience level when using chest X-rays as the test set. Kakeda et al. [41] (AUC 0.924 to 0.986, $p < 0.001$), Kligerman et al. [42] (AUC 0.38 to 0.43, $p = 0.007$), Schalekamp et al. [45] (AUC 0.812 to 0.841, $p = 0.0001$), and Sim et al. [46] (sensitivity 65.1 to 70.3, $p < 0.001$) showed significant improvement across all experience levels when using AI (Table 3a,b). Nam et al. [43] showed significant increase in average among every radiologist experience level (AUC 0.85 to 0.89, $p < 0.001$–0.87), but, individually, there were more observers with significant increase among non-radiologists, residents, and board-certified radiologists than thoracic radiologists. Only one out of four thoracic radiologists had a significant increase. On the other hand, Oda et al. [44] only showed significant improvement for the board-certified radiologists (AUC 0.848 to 0.883, $p = 0.011$) but not for the residents (AUC 0.770 to 0.788, $p = 0.310$). Kasai et al. [40] did not show any statistically significant improvement(Table 3b), but they reported that sensitivity improved when there were only lateral images available (67.9% to 71.6%, $p = 0.01$).

Table 3. Sensitivity and specificity (**a**); accuracy and AUC (**b**); and time measurement results (**c**) for sequential observer tests without and with AI-based devices as a second reader.

Author	Without AI-Based CAD		With AI-Based CAD		Change	Statistical Significance between Difference
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)		
			a			
Abe et al. [48]	64	-	81	-	↑	$p < 0.001$
Beyer et al. [19]	56.5	-	52.9	-	↓	$p < 0.001$
Bogoni et al. [34]	45.34 *	-	59.34 *	-	↑	$p < 0.03$
Chae et al. [35]	70 *	69 *	65 *	84 *	↓	Not statistically significant
Hwang et al. [50]	79 *	93.2 *	88.4 *	94 *	↑	$p = 0.006$–0.99
Kligerman et al. [42]	44	-	50	-	↑	$p < 0.001$
Sim et al. [46]	65.1	-	70.3	-	↑	$p < 0.001$
Author	**Without AI-Based CAD**		**With AI-Based CAD**		**Change**	**Statistical Significance between Difference**
	Accuracy (%)	**AUC**	**Accuracy (%)**	**AUC**		
			b			
Abe et al. [47]	-	0.81	-	0.87	↑	$p = 0.031$
Abe et al. [48]	-	0.94	-	0.98	↑	$p < 0.01$
Abe et al. [48]	-	0.77	-	0.81	↑	$p < 0.001$
Awai et al. [33]	-	0.64	-	0.67	↑	$p < 0.01$
Awai et al. [32]	-	0.843	-	0.924	↑	$p = 0.021$
Chae et al. [35]	69 *	0.005 *	75 *	0.13 *	↑	Not statistically significant
Chen et al. [36]	-	0.84 *	-	0.95 *	↑	$p < 0.221$
Fukushima et al. [49]	-	0.972 *	-	0.982 *	↑	$p < 0.071$
Hwang et al. [50]	-	0.880 *	-	0.934 *	↑	$p < 0.002$
Kakeda et al. [41]	-	0.924	-	0.986	↑	$p < 0.001$
Kasai et al. [40]	-	0.804	-	0.816	↑	Not statistically significant
Kligerman et al. [42]	-	0.38	-	0.43	↑	$p = 0.007$
Liu et al. [37]	-	0.913	-	0.938	↑	$p = 0.0266$
Matsuki et al. [38]	-	0.831	-	0.956	↑	$p < 0.001$
Nam et al. [43]	-	0.85 *	-	0.89 *	↑	$p < 0.001$-0.87
Oda et al. [44]	-	0.816	-	0.843	↑	$p = 0.011$–0.310
Rao et al. [39]	78	-	82.8	-	↑	$p < 0.001$
Schalekamp et al. [45]	-	0.812	-	0.841	↑	$p = 0.0001$
Author	**Without AI-Based CAD**		**With AI-Based CAD**		**Change**	**Statistical Significance between Difference**
	Time		**Time**			
			c			
Beyer et al. [19]	294 s (1)		274 s (1)		↑	$p = 0.04$
Bogoni et al. [34]	143 s (1)		225 s (1)		↓	-

a:* our calculated average; - not applicable; ↑ positive change; ↓ negative change. b: * our calculated average; - not applicable; ↑ positive change. c: (1) per image/case reading time; - not applicable;↑ positive change; ↓ negative change.

3.2.3. Detection of Several Different Findings

Abe et al. [47], Abe et al. [48], Fukushima et al. [49], and Hwang et al. [50] explored the diagnostic accuracy in detection of several different findings besides pulmonary nodules with their AI-algorithm (Table 1b). While Abe et al. [47] found significant improvement in all radiologists (Table 3b), Fukushima et al. [49] only found significant improvement in the group of radiologists that had more radiological task experience (AUC 0.958 to 0.971,

$p < 0.001$). In contrast, Abe et al. [48] found no significant improvement in the more senior radiologists for detection of interstitial disease ($p > 0.089$), and Hwang et al. [50] found no significant improvement in specificity for the detection of different major thoracic diseases in the more senior radiologists ($p > 0.62$). However, there were significant improvements in average among all observers for both studies (Table 3a,b).

4. Discussion

The main finding of our systematic review is that human observers assisted by AI-based devices had generally better detection or diagnostic performance using CT and chest X-ray, measured as sensitivity, specificity, accuracy, AUC, or time spent on image reading compared to human observers without AI-assistance.

Some studies suggest that physicians with less radiological task experience benefit more from AI-assistance [30,32,35,36,48,50], while others showed that physicians with greater radiological task experience benefitted the most from AI-assistance [44,49]. Gaube et al. [51] suggested that physicians with less experience were more likely to accept and deploy the suggested advice given to them by AI. They also reported that observers were generally not averse to following advice from AI compared to advice from humans. This suggests that the lack of improvement in the radiologists' performance with AI-assistance, was not caused by lack of trust in the AI-algorithm but more by the presence of confidence in own abilities. Oda et al. [44] did not find that the group of physicians with less task experience improved from assistance by AI-based device and had two possible explanations. Firstly, the less experienced radiologists had a larger interrater variation of diagnostic performance, leading to insufficient statistical power to show statistical significance. This was also an argument used by Fukushima et al. [49]. Secondly, they argued that the use of AI-assistance lowers false-negative more than false-positive findings, and radiologists with less task experienced generally had more false-positive findings. However, Nam et al. [43] found that physicians with less task experience were more inclined to change their false-negative diagnosis' and not their false-positive findings; therefore, they benefitted more from AI-assistance. Nam et al. [43], confirmed Oda et al.'s [44] finding in that there was a higher acceptance rate for false-negative findings. Brice [52] also confirmed this and suggested that correcting false-negative findings could have the most impact on reducing errors in radiological diagnosis. Although Oda et al. [44], Nam et al. [43], and Gaube et al. [51] had different reports on which level of physicians could improve their performance the most from the assistance of AI-based devices, they all confirm that AI-assistance lowers false-negative findings, which warrants advancing development and implementation of AI-based devices in to the clinics.

A limitation of our review is the heterogeneity of our included studies, e.g., the different methods for observer testing; some of our studies used a blinded observer test where AI-based devices was used as a concurrent reader (Table 1a), some studies used an un-blinded, sequential observer test (Table 1b), and some used both [19]. To the best of our knowledge, Kobayashi et al. [53] was one of the first to use and discuss both test types. Even though they concluded that there was no statistical significance in the difference of the results obtained from the two methods, they argue that an un-blinded, sequential test type would be less time consuming and practically easier to perform. Since then, others have adopted this method of testing [54] not only in thoracic diagnostic imaging and accepted it as a method for comparing effect of diagnostic tests [55]. Beyer et al. [19] also performed both methods of testing, but they did not come to the same conclusions about the results as Kobayashi et al. [53]. Their results of the two test methods were not the same; In the blinded concurrent reader test, they used more reading time per image (294 s to 337 s, $p = 0.04$) but achieved higher sensitivity (56.5 to 61.6, $p < 0.001$), and, in the un-blinded sequential reader test, they were quicker to interpret each image (294 s to 274 s, $p = 0.04$) but had worse sensitivity (56.5 to 52.9, $p < 0.001$) when assisted by AI. The test observers in the study by Kobayashi et al. [53] did not experience prolonged reading time, even though Bogoni et al. [34] confirmed the results by Beyer et al. [19] and also argued

that correcting false-positives would prolong the time spent on an image. Roos et al. [56] also reported prolonged time spent on rejecting false positive cases when testing their computer-aided device and explained that false-positive cases may be harder to distinguish from true-positive cases. This suggests that the sequential observer test design could result in prolonged time spent on reading an image when assisted by a device since they are forced to decide on previous findings. Future observer test studies must, therefore, be aware of this bias, and more studies are needed to investigate this aspect of observer tests.

A pre-requisite for AI-based devices to have a warranted place in diagnostic imaging is that it has higher accuracy than the intended user, since human observers with less experience may have a higher risk of also being influenced by inaccurate advice due to availability bias [57] and premature closure [58]. To be able to include a larger number of studies, we allowed the possibility of some inter-study variability in the performance of the AI-based devices because of different AI-algorithms being used. We recognize this as a limitation adding to the heterogeneity of our systematic review. In addition, we did not review the diagnostic performance of the AI-algorithm by itself, and we did not review the training or test dataset that was used to construct the AI-algorithm. Because of the different AI-algorithms, the included studies may also have been subjected to publication bias since there may be a tendency to only publish well-performing AI-algorithms.

Improved performance in users is a must before implementation can be successful. Our systematic review focused on observer tests performed in highly controlled environments where they were able to adjust their study settings to eliminate biases and variables. However, few prospective clinical trials have been published where AI-based devices have been used, in a more dynamic and clinically realistic environment [59,60]. No clinical trials have been published using AI-based devices on thoracic CT or chest X-rays, whether it be as a stand-alone diagnostic tool or as an additional reader to humans [61]. Our systematic review has, therefore, been a step towards the integration of AI in the clinics by showing that it generally has a positive influence on physicians when used as an additional reader. Further studies are warranted not only on how AI-based devices influence human decision making but also on their performance and integration into a more dynamic, realistic clinical setting.

5. Conclusions

Our systematic review showed that sensitivity, specificity, accuracy, AUC, and/or time spent on reading diagnostic images generally improved when using AI-based devices compared to not using them. Disagreements still exist, and more studies are needed to uncover factors that may inhibit an added value by AI-based devices on human decision-making.

Author Contributions: Conceptualization, D.L., J.F.C., S.D., H.D.Z., L.T., D.E., M.F. and M.B.N.; methodology, D.L., L.M.P., C.A.L. and J.F.C.; formal analysis, D.L., L.M.P. and J.F.C.; investigation, D.L., L.M.P., J.F.C. and M.B.N.; writing—original draft preparation, D.L.; writing—review and editing, D.L., L.M.P., C.A.L., H.D.Z., D.E., L.T., M.F., S.D., J.F.C. and M.B.N.; supervision, J.F.C., S.D. and M.B.N.; project administration, D.L.; funding acquisition, S.D. and M.B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovation Fund Denmark (IFD) with grant no. 0176-00013B for the AI4Xray project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Sharma, P.; Suehling, M.; Flohr, T.; Comaniciu, D. Artificial Intelligence in Diagnostic Imaging: Status Quo, Challenges, and Future Opportunities. *J. Thorac. Imaging* **2020**, *35*, S11–S16. [CrossRef]
2. Aidoc. Available online: https://www.aidoc.com/ (accessed on 11 November 2021).
3. Mu, W.; Jiang, L.; Zhang, J.; Shi, Y.; Gray, J.E.; Tunali, I.; Gao, C.; Sun, Y.; Tian, J.; Zhao, X.; et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat. Commun.* **2020**, *11*, 5228. [CrossRef]
4. Trebeschi, S.; Bodalal, Z.; Boellaard, T.N.; Bucho, T.M.T.; Drago, S.G.; Kurilova, I.; Calin-Vainak, A.M.; Pizzi, A.D.; Muller, M.; Hummelink, K.; et al. Prognostic Value of Deep Learning-Mediated Treatment Monitoring in Lung Cancer Patients Receiving Immunotherapy. *Front. Oncol.* **2021**, *11*. [CrossRef]
5. Willemink, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing Medical Imaging Data for Machine Learning. *Radiology* **2020**, *295*, 4–15. [CrossRef]
6. Laino, M.E.; Ammirabile, A.; Posa, A.; Cancian, P.; Shalaby, S.; Savevski, V.; Neri, E. The Applications of Artificial Intelligence in Chest Imaging of COVID-19 Patients: A Literature Review. *Diagnostics* **2021**, *11*, 1317. [CrossRef]
7. Pehrson, L.M.; Nielsen, M.B.; Lauridsen, C. Automatic Pulmonary Nodule Detection Applying Deep Learning or Machine Learning Algorithms to the LIDC-IDRI Database: A Systematic Review. *Diagnostics* **2019**, *9*, 29. [CrossRef]
8. Li, D.; Vilmun, B.M.; Carlsen, J.F.; Albrecht-Beste, E.; Lauridsen, C.; Nielsen, M.B.; Hansen, K.L. The Performance of Deep Learning Algorithms on Automatic Pulmonary Nodule Detection and Classification Tested on Different Datasets That Are Not Derived from LIDC-IDRI: A Systematic Review. *Diagnostics* **2019**, *9*, 207. [CrossRef] [PubMed]
9. Strohm, L.; Hehakaya, C.; Ranschaert, E.R.; Boon, W.P.C.; Moors, E.H.M. Implementation of artificial intelligence (AI) applications in radiology: Hindering and facilitating factors. *Eur. Radiol.* **2020**, *30*, 5525–5532. [CrossRef] [PubMed]
10. Wagner, R.F.; Metz, C.E.; Campbell, G. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. *Acad. Radiol.* **2007**, *14*, 723–748. [CrossRef] [PubMed]
11. Gur, D. Objectively Measuring and Comparing Performance Levels of Diagnostic Imaging Systems and Practices. *Acad. Radiol.* **2007**, *14*, 641–642. [CrossRef]
12. Whiting, P.F.; Rutjes, A.W.S.; Westwood, M.E.; Mallett, S.; Deeks, J.J.; Reitsma, J.B.; Leeflang, M.M.; Sterne, J.A.; Bossuyt, P.M.; QUADAS-2 Group. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann. Intern. Med.* **2011**, *155*, 529–536. [CrossRef] [PubMed]
13. Bai, H.X.; Wang, R.; Xiong, Z.; Hsieh, B.; Chang, K.; Halsey, K.; Tran, T.M.L.; Choi, J.W.; Wang, D.-C.; Shi, L.-B.; et al. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. *Radiology* **2021**, *299*, E225. [CrossRef] [PubMed]
14. Dorr, F.; Chaves, H.; Serra, M.M.; Ramirez, A.; Costa, M.E.; Seia, J.; Cejas, C.; Castro, M.; Eyheremendy, E.; Slezak, D.F.; et al. COVID-19 pneumonia accurately detected on chest radiographs with artificial intelligence. *Intell. Med.* **2020**, *3-4*, 100014. [CrossRef] [PubMed]
15. Kim, J.H.; Kim, J.Y.; Kim, G.H.; Kang, D.; Kim, I.J.; Seo, J.; Andrews, J.R.; Park, C.M. Clinical Validation of a Deep Learning Algorithm for Detection of Pneumonia on Chest Radiographs in Emergency Department Patients with Acute Febrile Respiratory Illness. *J. Clin. Med.* **2020**, *9*, 1981. [CrossRef] [PubMed]
16. Liu, P.-Y.; Tsai, Y.-S.; Chen, P.-L.; Tsai, H.-P.; Hsu, L.-W.; Wang, C.-S.; Lee, N.-Y.; Huang, M.-S.; Wu, Y.-C.; Ko, W.-C.; et al. Application of an Artificial Intelligence Trilogy to Accelerate Processing of Suspected Patients With SARS-CoV-2 at a Smart Quarantine Station: Observational Study. *J. Med. Internet Res.* **2020**, *22*, e19878. [CrossRef]
17. Yang, Y.; Lure, F.Y.; Miao, H.; Zhang, Z.; Jaeger, S.; Liu, J.; Guo, L. Using artificial intelligence to assist radiologists in distinguishing COVID-19 from other pulmonary infections. *J. X-ray Sci. Technol.* **2021**, *29*, 1–17. [CrossRef]
18. Zhang, D.; Liu, X.; Shao, M.; Sun, Y.; Lian, Q.; Zhang, H. The value of artificial intelligence and imaging diagnosis in the fight against COVID-19. *Pers. Ubiquitous Comput.* **2021**, 1–10. [CrossRef]
19. Beyer, F.; Zierott, L.; Fallenberg, E.M.; Juergens, K.U.; Stoeckel, J.; Heindel, W.; Wormanns, D. Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader. *Eur. Radiol.* **2007**, *17*, 2941–2947. [CrossRef]
20. De Hoop, B.; de Boo, D.W.; Gietema, H.A.; van Hoorn, F.; Mearadji, B.; Schijf, L.; van Ginneken, B.; Prokop, M.; Schaefer-Prokop, C. Computer-aided Detection of Lung Cancer on Chest Radiographs: Effect on Observer Performance. *Radiology* **2010**, *257*, 532–540. [CrossRef] [PubMed]
21. Koo, Y.H.; Shin, K.E.; Park, J.S.; Lee, J.W.; Byun, S.; Lee, H. Extravalidation and reproducibility results of a commercial deep learning-based automatic detection algorithm for pulmonary nodules on chest radiographs at tertiary hospital. *J. Med. Imaging Radiat. Oncol.* **2020**, *65*, 15–22. [CrossRef]
22. Kozuka, T.; Matsukubo, Y.; Kadoba, T.; Oda, T.; Suzuki, A.; Hyodo, T.; Im, S.; Kaida, H.; Yagyu, Y.; Tsurusaki, M.; et al. Efficiency of a computer-aided diagnosis (CAD) system with deep learning in detection of pulmonary nodules on 1-mm-thick images of computed tomography. *Jpn. J. Radiol.* **2020**, *38*, 1052–1061. [CrossRef] [PubMed]
23. Lee, K.H.; Goo, J.M.; Park, C.M.; Lee, H.J.; Jin, K.N. Computer-Aided Detection of Malignant Lung Nodules on Chest Radiographs: Effect on Observers' Performance. *Korean J. Radiol.* **2012**, *13*, 564–571. [CrossRef] [PubMed]

24. Li, F.; Hara, T.; Shiraishi, J.; Engelmann, R.; MacMahon, H.; Doi, K. Improved Detection of Subtle Lung Nodules by Use of Chest Radiographs with Bone Suppression Imaging: Receiver Operating Characteristic Analysis With and Without Localization. *Am. J. Roentgenol.* **2011**, *196*, W535–W541. [CrossRef] [PubMed]
25. Li, F.; Engelmann, R.; Pesce, L.L.; Doi, K.; Metz, C.E.; MacMahon, H. Small lung cancers: Improved detection by use of bone suppression imaging-comparison with dual-energy subtraction chest radiography. *Radiology* **2011**, *261*, 937–949. [CrossRef]
26. Liu, K.; Li, Q.; Ma, J.; Zhou, Z.; Sun, M.; Deng, Y.; Tu, W.; Wang, Y.; Fan, L.; Xia, C.; et al. Evaluating a Fully Automated Pulmonary Nodule Detection Approach and Its Impact on Radiologist Performance. *Radiol. Artif. Intell.* **2019**, *1*, e180084. [CrossRef]
27. Martini, K.; Blüthgen, C.; Eberhard, M.; Schönenberger, A.L.N.; De Martini, I.; Huber, F.A.; Barth, B.K.; Euler, A.; Frauenfelder, T. Impact of Vessel Suppressed-CT on Diagnostic Accuracy in Detection of Pulmonary Metastasis and Reading Time. *Acad. Radiol.* **2020**, *28*, 988–994. [CrossRef]
28. Singh, R.; Kalra, M.K.; Homayounieh, F.; Nitiwarangkul, C.; McDermott, S.; Little, B.P.; Lennes, I.T.; Shepard, J.-A.O.; Digumarthy, S.R. Artificial intelligence-based vessel suppression for detection of sub-solid nodules in lung cancer screening computed tomography. *Quant. Imaging Med. Surg.* **2021**, *11*, 1134–1143. [CrossRef]
29. Nam, J.G.; Kim, M.; Park, J.; Hwang, E.J.; Lee, J.H.; Hong, J.H.; Goo, J.M.; Park, C.M. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *Eur. Respir. J.* **2020**, *57*, 2003061. [CrossRef]
30. Sung, J.; Park, S.; Lee, S.M.; Bae, W.; Park, B.; Jung, E.; Seo, J.B.; Jung, K.-H. Added Value of Deep Learning-based Detection System for Multiple Major Findings on Chest Radiographs: A Randomized Crossover Study. *Radiology* **2021**, *299*, 450–459. [CrossRef]
31. Rajpurkar, P.; O'Connell, C.; Schechter, A.; Asnani, N.; Li, J.; Kiani, A.; Ball, R.L.; Mendelson, M.; Maartens, G.; Van Hoving, D.J.; et al. CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digit. Med.* **2020**, *3*, 1–8. [CrossRef]
32. Awai, K.; Murao, K.; Ozawa, A.; Nakayama, Y.; Nakaura, T.; Liu, D.; Kawanaka, K.; Funama, Y.; Morishita, S.; Yamashita, Y. Pulmonary Nodules: Estimation of Malignancy at Thin-Section Helical CT—Effect of Computer-aided Diagnosis on Performance of Radiologists. *Radiology* **2006**, *239*, 276–284. [CrossRef]
33. Awai, K.; Murao, K.; Ozawa, A.; Komi, M.; Hayakawa, H.; Hori, S.; Nishimura, Y. Pulmonary Nodules at Chest CT: Effect of Computer-aided Diagnosis on Radiologists' Detection Performance. *Radiology* **2004**, *230*, 347–352. [CrossRef] [PubMed]
34. Bogoni, L.; Ko, J.P.; Alpert, J.; Anand, V.; Fantauzzi, J.; Florin, C.H.; Koo, C.W.; Mason, D.; Rom, W.; Shiau, M.; et al. Impact of a Computer-Aided Detection (CAD) System Integrated into a Picture Archiving and Communication System (PACS) on Reader Sensitivity and Efficiency for the Detection of Lung Nodules in Thoracic CT Exams. *J. Digit. Imaging* **2012**, *25*, 771–781. [CrossRef] [PubMed]
35. Chae, K.J.; Jin, G.Y.; Ko, S.B.; Wang, Y.; Zhang, H.; Choi, E.J.; Choi, H. Deep Learning for the Classification of Small ($\leq$2 cm) Pulmonary Nodules on CT Imaging: A Preliminary Study. *Acad. Radiol.* **2020**, *27*, e55–e63. [CrossRef] [PubMed]
36. Chen, H.; Wang, X.-H.; Ma, D.-Q.; Ma, B.-R. Neural network-based computer-aided diagnosis in distinguishing malignant from benign solitary pulmonary nodules by computed tomography. *Chin. Med. J.* **2007**, *120*, 1211–1215. [CrossRef]
37. Liu, J.; Zhao, L.; Han, X.; Ji, H.; Liu, L.; He, W. Estimation of malignancy of pulmonary nodules at CT scans: Effect of computer-aided diagnosis on diagnostic performance of radiologists. *Asia-Pacific J. Clin. Oncol.* **2020**, *17*, 216–221. [CrossRef]
38. Matsuki, Y.; Nakamura, K.; Watanabe, H.; Aoki, T.; Nakata, H.; Katsuragawa, S.; Doi, K. Usefulness of an Artificial Neural Network for Differentiating Benign from Malignant Pulmonary Nodules on High-Resolution CT. *Am. J. Roentgenol.* **2002**, *178*, 657–663. [CrossRef]
39. Rao, R.B.; Bi, J.; Fung, G.; Salganicoff, M.; Obuchowski, N.; Naidich, D. LungCAD: A clinically approved, machine learning system for lung cancer detection. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, USA, 12–15 August 2007; pp. 1033–1037.
40. Kasai, S.; Li, F.; Shiraishi, J.; Doi, K. Usefulness of Computer-Aided Diagnosis Schemes for Vertebral Fractures and Lung Nodules on Chest Radiographs. *Am. J. Roentgenol.* **2008**, *191*, 260–265. [CrossRef]
41. Kakeda, S.; Moriya, J.; Sato, H.; Aoki, T.; Watanabe, H.; Nakata, H.; Oda, N.; Katsuragawa, S.; Yamamoto, K.; Doi, K. Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer-Aided Diagnosis System. *Am. J. Roentgenol.* **2004**, *182*, 505–510. [CrossRef]
42. Kligerman, S.; Cai, L.; White, C.S. The Effect of Computer-aided Detection on Radiologist Performance in the Detection of Lung Cancers Previously Missed on a Chest Radiograph. *J. Thorac. Imaging* **2013**, *28*, 244–252. [CrossRef]
43. Nam, J.G.; Park, S.; Hwang, E.J.; Lee, J.H.; Jin, K.-N.; Lim, K.Y.; Vu, T.H.; Sohn, J.H.; Hwang, S.; Goo, J.M.; et al. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology* **2019**, *290*, 218–228. [CrossRef] [PubMed]
44. Oda, S.; Awai, K.; Suzuki, K.; Yanaga, Y.; Funama, Y.; MacMahon, H.; Yamashita, Y. Performance of Radiologists in Detection of Small Pulmonary Nodules on Chest Radiographs: Effect of Rib Suppression With a Massive-Training Artificial Neural Network. *Am. J. Roentgenol.* **2009**, *193*. [CrossRef] [PubMed]
45. Schalekamp, S.; van Ginneken, B.; Koedam, E.; Snoeren, M.M.; Tiehuis, A.M.; Wittenberg, R.; Karssemeijer, N.; Schaefer-Prokop, C.M. Computer-aided Detection Improves Detection of Pulmonary Nodules in Chest Radiographs beyond the Support by Bone-suppressed Images. *Radiology* **2014**, *272*, 252–261. [CrossRef]

46. Sim, Y.; Chung, M.J.; Kotter, E.; Yune, S.; Kim, M.; Do, S.; Han, K.; Kim, H.; Yang, S.; Lee, D.-J.; et al. Deep Convolutional Neural Network-based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. *Radiology* **2020**, *294*, 199–209. [CrossRef] [PubMed]
47. Abe, H.; Ashizawa, K.; Li, F.; Matsuyama, N.; Fukushima, A.; Shiraishi, J.; MacMahon, H.; Doi, K. Artificial neural networks (ANNs) for differential diagnosis of interstitial lung disease: Results of a simulation test with actual clinical cases. *Acad. Radiol.* **2004**, *11*, 29–37. [CrossRef]
48. Abe, H.; MacMahon, H.; Engelmann, R.; Li, Q.; Shiraishi, J.; Katsuragawa, S.; Aoyama, M.; Ishida, T.; Ashizawa, K.; Metz, C.E.; et al. Computer-aided Diagnosis in Chest Radiography: Results of Large-Scale Observer Tests at the 1996–2001 RSNA Scientific Assemblies. *RadioGraphics* **2003**, *23*, 255–265. [CrossRef]
49. Fukushima, A.; Ashizawa, K.; Yamaguchi, T.; Matsuyama, N.; Hayashi, H.; Kida, I.; Imafuku, Y.; Egawa, A.; Kimura, S.; Nagaoki, K.; et al. Application of an Artificial Neural Network to High-Resolution CT: Usefulness in Differential Diagnosis of Diffuse Lung Disease. *Am. J. Roentgenol.* **2004**, *183*, 297–305. [CrossRef] [PubMed]
50. Hwang, E.J.; Park, S.; Jin, K.-N.; Kim, J.I.; Choi, S.Y.; Lee, J.H.; Goo, J.M.; Aum, J.; Yim, J.-J.; Cohen, J.G.; et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw. Open* **2019**, *2*, e191095. [CrossRef]
51. Gaube, S.; Suresh, H.; Raue, M.; Merritt, A.; Berkowitz, S.J.; Lermer, E.; Coughlin, J.F.; Guttag, J.V.; Colak, E.; Ghassemi, M. Do as AI say: Susceptibility in deployment of clinical decision-aids. *Npj Digit. Med.* **2021**, *4*, 1–8. [CrossRef]
52. Brice, J. To Err is Human; Analysis Finds Radiologists Very Human. Available online: https://www.diagnosticimaging.com/view/err-human-analysis-finds-radiologists-very-human (accessed on 1 October 2021).
53. Kobayashi, T.; Xu, X.W.; MacMahon, H.; E Metz, C.; Doi, K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology* **1996**, *199*, 843–848. [CrossRef]
54. Petrick, N.; Haider, M.; Summers, R.M.; Yeshwant, S.C.; Brown, L.; Iuliano, E.M.; Louie, A.; Choi, J.R.; Pickhardt, P.J. CT Colonography with Computer-aided Detection as a Second Reader: Observer Performance Study. *Radiology* **2008**, *246*, 148–156. [CrossRef]
55. Mazumdar, M.; Liu, A. Group sequential design for comparative diagnostic accuracy studies. *Stat. Med.* **2003**, *22*, 727–739. [CrossRef] [PubMed]
56. Roos, J.E.; Paik, D.; Olsen, D.; Liu, E.G.; Chow, L.C.; Leung, A.N.; Mindelzun, R.; Choudhury, K.R.; Naidich, D.; Napel, S.; et al. Computer-aided detection (CAD) of lung nodules in CT scans: Radiologist performance and reading time with incremental CAD assistance. *Eur. Radiol.* **2009**, *20*, 549–557. [CrossRef] [PubMed]
57. Gunderman, R.B. Biases in Radiologic Reasoning. *Am. J. Roentgenol.* **2009**, *192*, 561–564. [CrossRef]
58. Busby, L.P.; Courtier, J.L.; Glastonbury, C.M. Bias in Radiology: The How and Why of Misses and Misinterpretations. *RadioGraphics* **2018**, *38*, 236–247. [CrossRef] [PubMed]
59. Wang, P.; Liu, X.; Berzin, T.M.; Brown, J.R.G.; Liu, P.; Zhou, C.; Lei, L.; Li, L.; Guo, Z.; Lei, S.; et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): A double-blind randomised study. *Lancet Gastroenterol. Hepatol.* **2020**, *5*, 343–351. [CrossRef]
60. Lin, H.; Li, R.; Liu, Z.; Chen, J.; Yang, Y.; Chen, H.; Lin, Z.; Lai, W.; Long, E.; Wu, X.; et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. *EClinicalMedicine* **2019**, *9*, 52–59. [CrossRef]
61. Nagendran, M.; Chen, Y.; Lovejoy, C.A.; Gordon, A.; Komorowski, M.; Harvey, H.; Topol, E.J.; A Ioannidis, J.P.; Collins, G.; Maruthappu, M. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **2020**, *368*, m689. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Diagnostics Editorial Office
E-mail: diagnostics@mdpi.com
www.mdpi.com/journal/diagnostics

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34

www.mdpi.com

ISBN 978-3-0365-6435-7