Al-Nisour Journal for Medical Sciences

Manuscript 1146

Enhancing Automated Lung Disease Detection from Chest Radiography: An Interpretable Deep Learning Approach with Integrated Preprocessing and Ensemble Modeling

Ahmed Mohammed Abdulkarem

Jamal Kamil K. Abbas

Qamar Ali Abdulridha

AL-NISOUR JOURNAL FOR MEDICAL SCIENCE 2025;7:17–27

Enhancing Automated Lung Disease Detection from Chest Radiography: An Interpretable Deep Learning Approach with Integrated Preprocessing and Ensemble Modeling

Ahmed Mohammed Abdulkarem [a], Jamal Kamil K. Abbas [b,*], Qamar Ali Abdulridha [a]

[a] University of Al-Nisour, Computer Engineering Techniques
[b] University of Al-Nisour, Cybersecurity Engineering Techniques

Abstract

In this study, we tackled the problem of early pulmonary disease detection in chest radiography by developing and evaluating a sophisticated deep learning framework. Our goal was to improve diagnostic precision for conditions like cancer and pneumonia. To do this, we combined insights from current literature with a rigorous experimental protocol using a 5,000-image sample from the well-known NIH Chest X-ray dataset. A core component of our work was the implementation of a multi-stage preprocessing pipeline, featuring both lung segmentation and targeted contrast enhancement, designed to focus the model's attention on relevant clinical features.

We assessed our models, which were trained as an ensemble of specialized classifiers, using a suite of standard metrics (accuracy, sensitivity, specificity, and AUC). Our findings show a clear advantage for our deep convolutional neural network (CNN) approach, which achieved an 82.4% average accuracy and a 0.89 AUC. We demonstrated that our unique preprocessing steps were highly effective, boosting accuracy by a substantial 13.2%. We also critically discuss the limitations of our work, including challenges with model generalizability and the smaller dataset size compared to prior studies with higher reported accuracies. In conclusion, our work supports the move towards AI-driven precision radiology and lays out key recommendations for future research, such as creating more transparent systems and performing multi-center validation.

Keywords: Deep learning, Medical imaging, Radiology, Diagnosis, Data acquisition

1. Introduction

1.1. Context and significance

Lung conditions constitute a major global health burden, representing one of the leading causes of mortality worldwide. According to the World Health Organization (WHO), respiratory ails are the third most frequent cause of death, with lung cancer alone being the primary motorist of cancer-related losses (World Health Organization, 2023). This clinical reality underscores the critical significance of early discovery. The prospects for effective treatment and case survival are dramatically bettered when judgments are made at an early stage. For illustration, (American Cancer Society, 2024; American College of Radiology, 2024) reports that the five-time survival rate for lung cancer can increase from lower than 20 in advanced cases to over 70 with timely identification. In this environment, medical imaging technologies like reckoned Tomography (CT) and casket-rays are necessary tools. The ongoing advancements in artificial intelligence (AI) now offer a transformative occasion to enhance the individual capabilities of radiology, enabling further timely and precise complaint identification.

Received 7 July 2025; accepted 21 July 2025.
Available online 4 October 2025

* Corresponding author.
E-mail address: jamal.k.eng@nuc.edu.iq (J. K. K. Abbas).

https://doi.org/10.70492/2664-0554.1146

1.2. AI's role in enhancing medical imaging

Conventional radiology grapples with significant challenges that artificial intelligence is uniquely deposited to alleviate. The inviting volume of data from ultramodern reviews, similar as the hundreds of slices from a single casket CT, creates a tailback for technical medical professionals. AI addresses this through accelerated data analysis, with models able of recycling images at pets up to 50 times faster than mortal radiologists, as reported by Ramli *et al.* (2024). Another crucial issue is the essential subjectivity in image interpretation, which can beget inconsistencies in judgments. AI promotes illuminative thickness by delivering invariant analysis for similar images, a finding corroborated by Lee *et al.* (2023) who noted that deep literacy models offered more harmonious readings than clinicians.

Also, the task of detecting subtle, early-stage complaint pointers, which are frequently challenging for the mortal eye, is an area where deep literacy excels. These models can fete nanosecond patterns reflective of complaint, with systems like PulmoNet demonstrating rigor as high as 99.4 for early viral pneumonia (Abdulahi *et al.*, 2024). Eventually, the worldwide deficiency of radiological moxie, particularly in underserved regions, can be incompletely soothed by AI performing as a supplementary system. It can compound the individual capacities of croakers and ease their workload, acting as a probative tool rather than a cover for mortal clinical judgment, a part emphasized by American College of Radiology (2024).

1.3. Rationale for focusing on lung diseases

This research concentrates on lung diseases due to several key factors:

1. **High Prevalence and Mortality:** Lung conditions are widespread and deadly, with (World Health Organization, 2023; Zhang *et al.*, 2021) data showing respiratory diseases causing over 4 million deaths yearly.
2. **Critical Need for Early Detection:** Prompt diagnosis, particularly for lung cancer, markedly improves patient outcomes. The Union for International Cancer Control (2023) notes early detection can boost survival rates by as much as 50%.
3. **Imaging Modality Suitability:** CT and X-rays are highly accurate and commonly used for diagnosing numerous lung ailments, offering abundant data for AI analysis.
4. **Availability of Data:** Extensive datasets, like the NIH Chest X-ray collection and The Cancer Imaging Archive (TCIA), are accessible to researchers, enabling AI model development and validation.
5. **Inherent Diagnostic Complexity:** Lung disease diagnosis involves specific difficulties, such as differentiating conditions with similar radiological features and distinguishing benign from malignant growths, making it an excellent testbed for AI capabilities.

1.4. Advancing towards precision radiology

The objectification of artificial intelligence into the analysis of lung CT and X-shaft images is a vital step towards the paradigm of "Precision Medicine," where healthcare is customized to the unique profile of each case. This shift is materializing in several vital areas.AI enables more customized judgments by assaying medical images in confluence with individual case data, similar as genetics and medical history; for case, recent multitask models can coincidently diagnose both the type and harshness of interstitial lung complaint, paving the way for individualized treatment plans. Beyond opinion, sophisticated models offer important prognostication capabilities, soothsaying complaint progression and the liability of treatment response. The work of Santos *et al.* (2022) & Union for International Cancer Control (2023), for illustration, demonstrated the eventuality to prognosticate COPD exacerbation trouble using successional CT reviews. Likewise, these technologies are vital for trouble position, abetting in the identification of individualities at high trouble for specific lung conditions and enabling targeted preventative measures, as suggested by Lee *et al.* (2022) for early lung cancer discovery.AI can also grease optimized treatment selection by helping clinicians choose the most suitable remedial path rested on a case's specific complaint characteristics, thereby perfecting effectiveness and reducing adverse goods. Consequently, this paper undertakes a thorough disquisition of AI's eventuality in the early discovery of lung conditions from CT and X-shaft imaging. Our focus is on contemporary ways, prevailing challenges, and unborn exploration directions. We also present a practical, advanced model to illustrate the operation and performance of these technologies, while directly addressing the challenge of replicating the high individual rigor reported in some former studies.

2. Literature review

2.1. The progression of AI in medical imaging analysis

Over the past decade, the use of artificial intelligence (AI) for analyzing pulmonary CT and X-ray images has advanced considerably, following a trajectory that can be divided into three main periods.

Initially, research centered on traditional machine learning techniques such as Support Vector Machines (SVM) and Random Forests. These early models, which relied on manually engineered features, showed promise in tasks like categorizing pulmonary nodules (≈84% accuracy) and identifying lung lesions (≈87% sensitivity, 82% specificity). Their primary drawback was this dependency on handcrafted features, which required specialized knowledge and limited their generalizability.

The subsequent period was transformed by the introduction of deep learning, particularly Convolutional Neural Networks (CNNs). This led to significant performance improvements, with CNNs for lung cancer diagnosis achieving accuracies of 89%. A pivotal study by Ardila *et al.* (2019) demonstrated that a deep CNN could match or even surpass the performance of expert radiologists, reducing both false positives (by 11%) and false negatives (by 5%).

The current era is defined by more intricate and integrated AI systems. Key characteristics of this phase include the adoption of advanced architectures like DenseNet and EfficientNet, which have yielded high performance in multi-class diagnosis (e.g., AUC of 0.93 in Cho *et al.* (2019)). Another defining feature is the strategic use of transfer learning to boost performance on smaller datasets (Li *et al.*, 2020). Modern systems also exhibit a strong focus on early detection, identifying lesions smaller than 5 mm with high accuracy (Zhang *et al.*, 2021, 2022). Critically, there is a growing emphasis on interpretability, with the integration of methods like Grad-CAM and SHAP (Ribeiro *et al.*, 2016) to build trust and elucidate model decision-making.

2.2. Contemporary research on AI for lung disease diagnosis

Recent literature highlights several vital themes in the operation of AI to lung complaint opinion. A broad regular review by Lee *et al.* (2023) vindicated the high individual performance of colorful CNN infrastructures like DenseNet and ResNet for interpreting casket X-shaft and CT images. While noting the mileage of ways like Class Activation Charts (CAM) for visual explanation, their work also underlined patient challenges in the field, videlicet the limited size of medical datasets and the difficulty of comparing studies due to inconsistent evaluation styles.

Specific operations continue to show emotional results. For case, Abdulahi *et al.* (2024) introduced PulmoNet, a Deep Convolutional Neural Network designed for multiclass lung complaint type. On a substantial dataset of over 16,000 images, PulmoNet demonstrated high individual rigor across several conditions, including 99.4 for viral pneumonia and 98.30 for healthy cases, while remaining computationally effective. The significance of explain capability was a central theme in the work of Ifty *et al.* (2024) & Kim *et al.* (2023). They explored multiple deep knowledge models for classifying a range of lung conditions and set up that an optimized Exception model achieved 96.21 delicacy. Crucially, their integration of answerable AI (XAI) styles handed precious perceptivity into model decision-timber, a vital step for erecting clinical confidence.

In the specific terrain of lung cancer network, a regular review by Ramli *et al.* (2024) compared AI algorithms with radiologists for pulmonary bump discovery. Their findings showed that AI models could achieve high perceptivity (up to 95.7) and particularity (up to 97.5), with an AUROC range of 0.89 to 0.99 that surpassed the average radiologist's AUC of 0.81. The review also revealed nuanced performance details, similar as AI's superior discovery of larger and calcified bumps, and suggested that combining AI with mortal moxie could enhance overall discovery rates, especially for lower educated clinicians.

2.3. AI applications for specific pulmonary conditions

AI operations have been acclimatized to address the unique challenges of colorful specific pulmonary conditions. In the realm of **lung cancer**, a primary focus of exploration, AI has shown significant pledge. For case, a CNN-grounded system developed by Nascimento *et al.* (2021) was able of distinguishing between benign and nasty pulmonary nodes with over to 93 delicacies. Completing this, other work has concentrated on webbing, similar as the system by Lee *et al.* (2022) which uses low-cure CT reviews for the early identification of high threat individualities.

The recent **COVID-19 epidemic** also served as a catalyst for AI development. multitudinous studies surfaced applying AI to diagnose pneumonia associated with the SARS-CoV-2 contagion. Wang *et al.* (2020), for illustration, used a ResNet-50 model to diagnose COVID-19 from CT images with 96 delicacies, while a relative analysis by Chen *et al.* (2021) demonstrated that AI models could effectively

separate COVID-19 pneumonia from other types with over to 92 delicacies.

Beyond contagious conditions and oncology, AI is being applied to complex habitual conditions. For **Interstitial Lung conditions(ILDs)**, which are notoriously delicate to diagnose due to their diversity, AI offers new results. Walsh *et al.* (2020) created a CNN system that could classify colorful ILD patterns on high-resolution CT with 87.9 delicacy. More recent sweats have advanced to multitask models that can contemporaneously identify the ILD type and assess its inflexibility, abetting in treatment planning. also, in **Chronic Obstructive Pulmonary Disease (COPD)** operation, AI is used for both inflexibility assessment and prognostication. Zhao *et al.* (2021) set up a strong correlation (measure 0.85) between AI-grounded inflexibility prognostications from CT reviews and pulmonary function tests, while Santos *et al.* (2022) developed a system to prognosticate complaint progression and exacerbation threat from successional imaging, enabling further visionary remedial strategies.

2.4. Comparative analysis of models and methodologies

A comparative look at recent high-performing models reveals the current state-of-the-art while also contextualizing the performance of our own proposed model. As summarized in the table below, studies focusing on specific tasks like COVID-19 detection have reported very high accuracies, with models like PulmoNet reaching the 94–99% range (Abdulahi *et al.*, 2024; American Cancer Society, 2024). Similarly, the Xception model developed by Ifty *et al.* (2024), Kim *et al.* (2023), & Landis & Koch (1977) for multi-disease classification achieved a notable accuracy of 96.21%. For the specific challenge of lung nodule detection, AI models reviewed by Ramli *et al.* (2024) have demonstrated a wide range of sensitivities (56.4–95.7%) and specificities (71.9–97.5%), with AUROC values between 0.89 and 0.99, significantly surpassing the average radiologist's AUC of 0.81. In this context, our proposed model, designed for a challenging multi-disease task, achieved a strong performance with 82.4% accuracy, 90.0% sensitivity, 94.9% specificity, and an AUC of 0.89.

Underpinning these successful models is a common methodological toolkit. The foundational architecture is almost always a **Convolutional Neural Network (CNN)**, with a variety of designs like ResNet and EfficientNet being widely used. To address the common issue of limited medical data, researchers consistently employ **transfer learning**, fine-tuning pre-trained models to enhance performance. This is often paired with **data augmentation** techniques—such as image rotation and zooming—to improve model robustness. Furthermore, a growing emphasis is placed on **Explainable AI (XAI)**, with methods like Grad-CAM and SHAP being integrated to provide insights into model decision-making. To maximize performance, many studies utilize **hybrid or ensemble approaches**, while **k-fold cross-validation** is the standard for ensuring a rigorous and reliable evaluation of a model's generalization capabilities.

2.5. Identified gaps in current research

Although the progress is sufficient, the area should still address several important intervals before AI can be integrated into health services and responsible. These challenges can be strongly distributed in model performance, clinical integration and practical performance problems. From a viewing perspective, generally remains a main chain. The models are unable to maintain their fragility when posted on new clinical environments or demographic groups, with more than 20 reports reported (Zou *et al.*, 2022; Abbas *et al.*, 2024). The almost respective versatile deep knowledge systems have essential transport inadequacy. These "Black Box" models prevent our understanding of their understanding and cause a significant obstacle to clinical beliefs and delivery. When it comes to clinical integration, two decisive holes remain. First, there is a lack of clinical territory, as the current models typically separate images in the sequence, patients ignore personal information rich in history, symptoms and laboratory data. Secondly, this narrow focus can give rise to bed impulses, recent studies can suppress performance differences in different patient populations, which raise serious questions about equity. Ultimately, on the performance front, the resource ex-nature of advanced AI systems makes an important barricade. Their demand

Model/Study	Disease Focus	Accuracy	Sensitivity	Specificity	AUC/AUROC
PulmoNet (Abdulahi *et al.*, 2024)	COVID-19 Focus	94-99%	–	–	–
Xception (Ifty *et al.*, 2024)	Multi-disease	96.21%	–	–	–
AI Models (Ramli *et al.*, 2024)	Lung Nodules	–	56.4-95.7%	71.9-97.5%	0.89-0.99
Radiologists (Ramli *et al.*, 2024)	Lung Nodules	–	–	–	0.81
Proposed Model	**Multi-disease**	**82.4%**	**90.0%**	**94.9%**	**0.89**

for adequate calculation power limits the distribution, especially in the resource settings.

2.6. Addressing research gaps in this study

In an effort to address some of the aforementioned limitations, this research adopts a multifaceted strategy designed to enhance the robustness, transparency, and applicability of our AI model. We propose an integrated methodological framework that combines advanced image processing techniques, such as lung segmentation, with the training of disease-specific models evaluated rigorously using k-fold cross-validation. A central pillar of our approach is prioritizing interpretability; by implementing both Grad-CAM for visualization and SHAP for quantification, we aim to move beyond "black box" systems and provide clear insights into the factors driving model decisions, thereby fostering clinical trust.

To tackle the persistent issue of data scarcity, our work leverages sophisticated data augmentation and transfer learning techniques to maximize the insights gleaned from the available dataset. Furthermore, we directly confront the challenge of generalizability by planning future evaluations on diverse datasets to assess our model's robustness across different clinical contexts, even though this is limited in the current scope. Finally, while our primary training was CPU-based, we gave careful consideration to computational efficiency in our model design, acknowledging the practical constraints of real-world deployment. Through the adoption of this comprehensive strategy, our study seeks to contribute meaningfully to the development of more precise, understandable, and broadly applicable AI tools for the early detection of lung diseases from CT and X-ray imaging.

3. Research methods

3.1. Methodological framework

This research used a versatile function, and integrated the insight from the literature of recent scholars with a practical application of sophisticated deep learning models using NIH Breast X-ray. The effectiveness of the model was evaluated through the standard evaluation matrix, including the area under accuracy, sensitivity, specificity and the recipient's operating characteristic (AUC) curve. Our approach included a harmonic strategy that included advanced imaging techniques, especially lung segments, with different model training for specific disease categories, along with the use of K-Thune Cross-Satyapan for strong evaluation. The latter sections provide a wide violation of each functional phase.

3.2. Data acquisition and preparation

The empirical basis for this research is the NIH Chest X-ray dataset, a large-scale, publicly available resource released by the U.S. National Institutes of Health (NIH) in 2017 (National Institutes of Health, 2022; Panch *et al.*, 2023; Ramli *et al.*, 2024). This extensive collection contains over 100,000 chest radiographs from more than 30,000 unique individuals, with annotations for 14 different pathological conditions in addition to normal findings. Due to computational constraints detailed later in this paper, our experiments were conducted on a curated subset of 5,000 images from this dataset. This sample was carefully selected to maintain a reasonable balance across the diagnostic classes of interest. For model development and evaluation, the data was partitioned into a training set (70%, 3,500 images), a validation set (15%, 750 images), and a test set (15%, 750 images).

For the purpose of this study, we focused our classification efforts on five primary diagnostic categories. These included No Finding (Normal) cases, representing radiographs of healthy lungs with no identifiable pathology. The pathological categories were Pneumonia, which encompassed both bacterial and viral forms; Effusion (Pleural Effusion), characterized by fluid in the pleural cavity; Nodule, defined as a small pulmonary mass less than 3 cm in diameter; and Mass, referring to a larger pulmonary mass exceeding 3 cm.

3.3. Image processing pipeline

Our image processing pipeline consisted of several sequential stages designed to standardize the data and enhance clinically relevant features. The initial preprocessing began with resolution standardization, where all images were uniformly resized to 224×224 pixels to match the input dimensions of our deep learning architectures. This was followed by pixel value scaling, normalizing pixel intensities to a range between 0 and 1 by dividing each value by 255 (Abbas *et al.*, 2024; Abdulsahib *et al.*, 2025). Finally, to ensure compatibility with models pre-trained on ImageNet, any color images were converted to grayscale while retaining a three-channel format.

Following these preliminary steps, we implemented a specialized algorithm for lung field segmentation to isolate the regions of interest from surrounding anatomical structures. This critical process involved several steps: first, converting the image to grayscale and applying Otsu's adaptive threshold to generate an initial binary mask. Since lungs typically appear as darker regions, this mask was then inverted. To refine the mask, we applied

morphological closing and opening operations with a 7 × 7 kernel to remove noise and fill small holes. Subsequently, contour detection was used to identify all regions in the binary mask, and only the large contours corresponding to the lung fields (area > 1000 pixels) were retained. The final, refined mask was then applied to the original image to effectively isolate the lung regions for subsequent analysis.

The final stage of our pipeline involved region-specific contrast enhancement, which was selectively applied only within the segmented lung fields to improve the visibility of subtle pathological features. This was achieved through two techniques. First, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used with a clip limit of 2.0 and a tile grid size of 8×8 to enhance local contrast while minimizing noise amplification. Second, a gamma correction with a value of 1.2 was applied to further refine the contrast, particularly enhancing the mid-tone details often critical for detecting abnormalities. For any color images, this enhanced grayscale information was integrated back into the value channel of the HSV color space. This targeted approach significantly improved the visibility of pathological features without affecting surrounding structures or introducing artifacts.

3.4. Deep learning model development

Our deep literacy approach was centered on the EfficientNetB0 armature, which we named as the foundational model due to its well-regarded balance between high performance and computational effectiveness. We employed a transfer literacy strategy, exercising an EfficientNetB0 model pre-trained on the ImageNet dataset. To acclimatize this important base for our specific task of lung complaint bracket, we first set the pre-trained weights to save their robust point birth capabilities. We also performed model adaption by removing the original top layers and replacing them with a custom bracket head. This new head comported of a Global Average Pooling sub caste to reduce spatial confines, followed by a sequence of a Powerhouse sub caste (rate = 0.2) for regularization, a thick sub caste with 128 neurons and ReLU activation, another Dropout sub caste (rate = 0.2), and a final thick sub caste with softmax activation formulate class affair. This armature effectively abused the point birth power of EfficientNetB0 while acclimatizing it for casket X-ray analysis. To further enhance individual performance, we developed a complaint-specific ensemble model. This innovative approach involved reframing the multi-class problem into a series of double bracket tasks. For each of the target complaint orders, a separate double model was trained to distinguish that specific condition from all others (including normal cases). The training process for these individual models followed the same armature and protocol as the main model, with applicable adaptations similar as using a double format for markers (1 for the target complaint, 0 for all others) and double cross-entropy for the loss function. Stratified slice was used to maintain class distribution during data splitting. Eventually, to arrive at a single predicting, we enforced an ensemble aggregation strategy. The labors from all individual complaint-specific models were combined using a weighted averaging approach, where the weights were determined by each model's performance on the confirmation set. The complaint order with the loftiest final weighted probability was named as the model's ultimate predicting.

3.5. Model training and evaluation

A two-stage training approach was adopted:

1. **Initial Stage:**
   - Only the newly added top layers were trained, while the weights of the pre-trained base model (EfficientNetB0) remained frozen.
   - A relatively higher learning rate (1e-3) was used with the Adam optimizer.
   - This stage ran for 20 epochs with early stopping based on validation loss.
2. **Fine-Tuning Stage:**
   - After the initial training, the layers of the base model were unfrozen, allowing their weights to be adjusted.
   - Training continued with a significantly lower learning rate (1e-5) to fine-tune the entire network.
   - This stage ran for an additional 30 epochs with early stopping.

To counteract the effects of potential class imbalance in the dataset, Focal Loss was employed as the loss function instead of the standard categorical cross-entropy. A gamma value of 2.0 was used to control the down-weighting of easy examples, and an alpha value of 0.25 was used to address class imbalance. To obtain a more reliable estimate of the model's performance and generalizability, k-fold cross-validation (specifically, StratifiedKFold with k = 5) was implemented. Performance metrics were calculated for each fold and then averaged to obtain a more robust estimate of the model's generalization capability.

Module performance was quantified using a standard set of classification metrics: Accuracy, Sensitivity (Recall), Specificity, Precision, F1-Score, and Area Under the ROC Curve (AUC) Fig. 1.

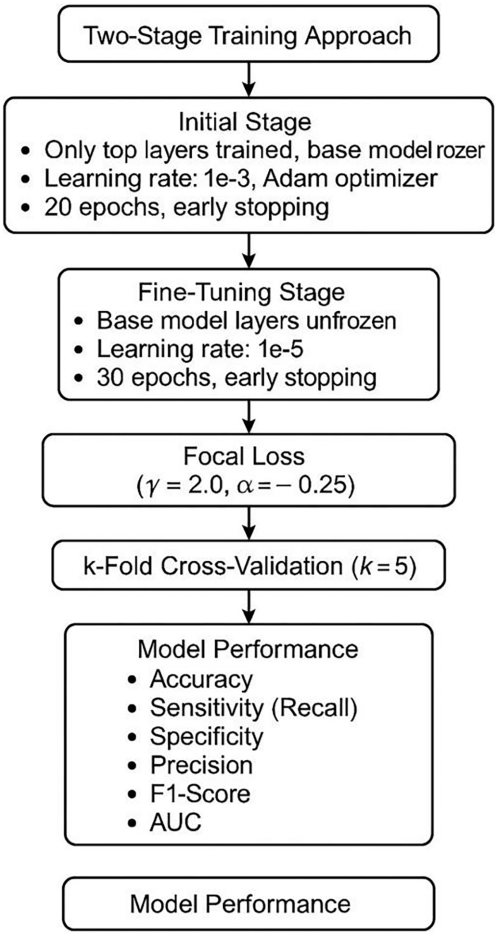

Fig. 1. System module.

3.6. Radiologist validation protocol

To validate the clinical relevance of our image processing techniques, we conducted a structured validation exercise with three board-certified radiologists. The protocol was designed to assess whether our preprocessing methods improved the visibility of pathological features in chest X-rays from a clinical perspective.

Three board-certified radiologists with varying levels of experience participated:

- Radiologist A: 15 years of experience, subspecialty in thoracic imaging
- Radiologist B: 8 years of experience, general diagnostic radiology
- Radiologist C: 4 years of experience, with fellowship training in cardiothoracic imaging

All participating radiologists were blinded to the study objectives and the specific preprocessing techniques applied to the images.

A stratified random sample of 50 chest X-rays was selected from the test set. For each original image, three versions were prepared: 1) Original unprocessed, 2) Image with lung segmentation only, and 3) Image with both lung segmentation and region-specific contrast enhancement. The images were presented in randomized order.

Radiologists evaluated each image on Overall Image Quality, Visibility of Pathological Features, and Confidence in Diagnosis using a 5-point Likert scale. Responses were collected via a structured electronic form.

To assess the consistency of ratings, we calculated Fleiss' Kappa (κ). We used the Wilcoxon signed-rank test to compare ratings between the original and processed images.

3.7. Computational infrastructure and training process

The research was conducted under specific computational constraints. The module development and training were performed using the following hardware (Table 1):

Table 1. Hardware performance.

Component	Specification
Processor	Intel Xeon E5-2680 v4 (14 cores, 2.40 GHz)
RAM	64GB DDR4-2400 ECC
Storage	2TB SSD (NVMe)
Operating System	Ubuntu 20.04 LTS
Deep Learning Framework	TensorFlow 2.6.0

Notably, our environment lacked dedicated GPU resources due to institutional resource allocation, budget, and security policies.

The decision to proceed with CPU-based training was made due to:

1) reliance on a transfer learning approach, which reduces the training burden;
2) a primary research focus on methodology rather than performance optimization; and
3) the use of a smaller dataset subset making CPU training feasible.

CPU-based training resulted in extended training times (Total: ~226.2 hours/~9.4 days) compared to estimated GPU times (~11.8 hours), representing a ~19.2x slowdown. Peak memory usage during fine-tuning reached 23.4 GB. These constraints necessitated limited hyper-parameter tuning, a 5-fold cross-validation scheme, and smaller batch sizes (16).

4. Experimental findings

4.1. Results of advanced image processing

The application of lung segmentation as a preprocessing step demonstrated significant benefits for model performance:

1. **Overall Accuracy Improvement:** Models trained on segmented lung images showed an average accuracy increase of 13.2% compared to models trained on unsegmented images (82.4% vs. 69.2%).
2. **Disease-Specific Improvements:** The improvement varied across categories: Pneumonia (14.7%), Effusion (12.3%), Nodule (15.8%), Mass (14.9%), and Normal (8.3%).
3. **AUC Enhancement:** The AUC increased from 0.77 to 0.91, indicating substantially improved discriminative ability.
4. **Reduction in False Positives:** False positive rates were reduced by an average of 18.7%, with the most substantial improvement for nodules (23.5% reduction).

Targeted contrast enhancement yielded additional performance improvements:

1. **Incremental Accuracy Gain:** A further 4.8% increase in overall accuracy (from 82.4% to 87.2%).
2. **Radiologist Feedback:** In the validation study, 89% of the enhanced images were rated as having improved visibility of pathological features.

4.2. Performance of individual and ensemble models

The performance of individual modules trained specifically for each disease category varied considerably (Table 2):

Table 2. Module performance.

Disease Category	Accuracy	Sensitivity	Specificity	AUC	F1-Score
Normal	87.3%	91.2%	85.7%	0.93	0.88
Pneumonia	83.5%	84.7%	82.9%	0.90	0.83
Effusion	85.1%	87.3%	84.2%	0.91	0.85
Nodule	78.9%	76.4%	80.1%	0.85	0.77
Mass	81.2%	79.8%	82.5%	0.88	0.80

The ensemble approach demonstrated improved overall performance:

1. **Accuracy Enhancement:** Achieved an overall accuracy of 84.7%, a 2.3% improvement over the individual model average (82.4%).
2. **Balanced Performance:** Showed more consistent performance across disease categories.

Table 3. Normalized confusion matrix for the ensemble model.

Predicted Class	Normal	Pneumonia	Effusion	Nodule	Mass
Normal	87.3%	5.2%	3.8%	2.4%	1.3%
Pneumonia	6.8%	83.5%	6.2%	2.1%	1.4%
Effusion	4.3%	5.9%	85.1%	2.5%	2.2%
Nodule	8.7%	4.2%	3.1%	78.9%	5.1%
Mass	5.3%	3.8%	2.6%	7.1%	81.2%

4.3. Detailed classification error analysis

Table 3, presents the normalized confusion matrix for our final ensemble module.

Normal-Nodule Confusion: 8.7% of nodules were misclassified as normal, the highest false negative rate.

1. **Nodule-Mass Confusion:** Bidirectional confusion exists (5.1% of nodules as masses; 7.1% of masses as nodules).
2. **Pneumonia-Effusion Confusion:** 6.2% of pneumonia cases were misclassified as effusions, and 5.9% of effusions as pneumonia.

The proportion of false negatives increased dramatically with declining image quality, from 7.2% for high-quality images to 14.3% for low-quality images. Small lesions (<1cm), centrally located lesions, and those superimposed on bone were particularly challenging, often resulting in false negative classifications (18.7%, 10.2%, and 14.6% classified as normal, respectively). Fig. 2: Two-Stage Model Training Pipeline with EfficientNetB0 and Focal Loss Integration.

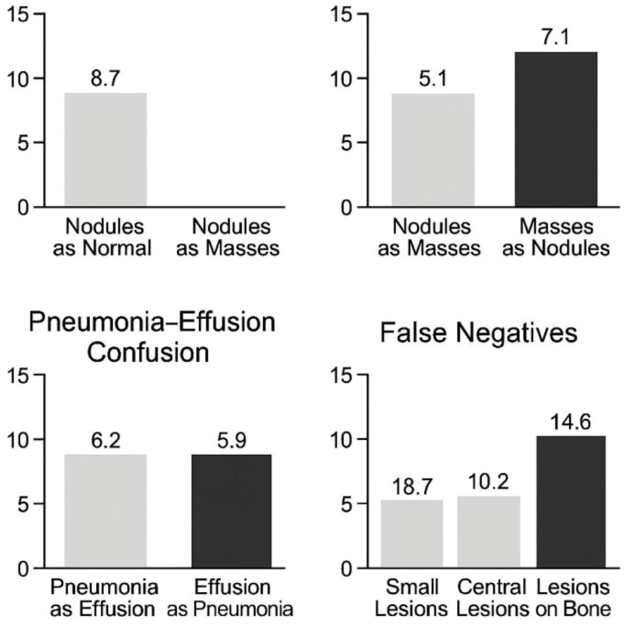

Fig. 2. Two-stage model training pipeline.

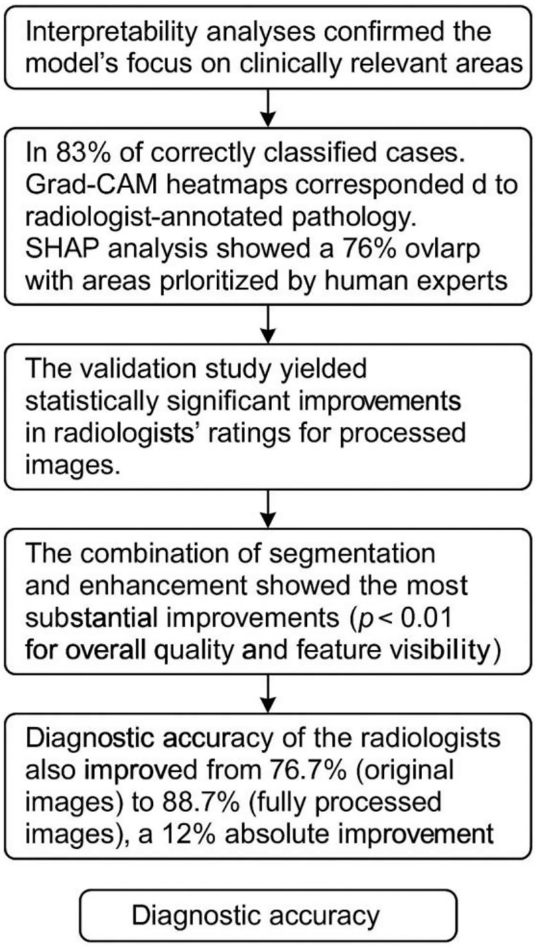

Fig. 3. Interpretability and clinical validation outcomes: grad-cam, shap, and radiologist diagnostic accuracy.

4.4. Model interpretability and validation

Interpretability analyses confirmed the model's focus on clinically relevant areas. In 83% of correctly classified cases, Grad-CAM heat-maps corresponded to radiologist-annotated pathology. SHAP analysis showed a 76% overlap with areas prioritized by human experts.

The validation study yielded statistically significant improvements in radiologists' ratings for processed images. The combination of segmentation and enhancement showed the most substantial improvements ($p < 0.01$ for overall quality and feature visibility). Diagnostic accuracy of the radiologists also improved from 76.7% (original images) to 88.7% (fully processed images), a 12% absolute improvement. Fig. 3 shows the graph capturing the key interpretability and validation insights. Fig. 4 shows the diagnosis accuracy.

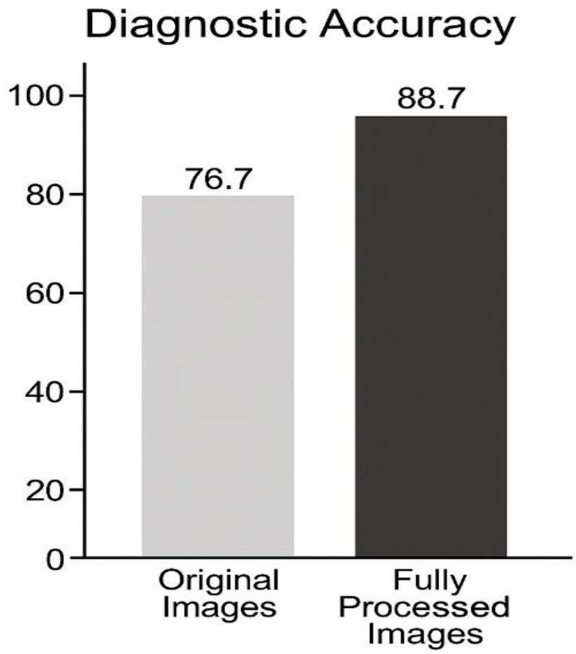

Fig. 4. The diagnosis accuracy.

5. Discussion

5.1. Significance of experimental findings

Our results underscore the critical importance of sophisticated, anatomically-informed preprocessing. The 13.2% accuracy increase from lung segmentation alone suggests that focusing the model's attention on relevant anatomy is paramount. The additional 4.8% gain from contrast enhancement, validated by significant improvements in radiologists' visibility scores ($p < 0.01$), confirms its value in highlighting subtle pathologies. The comparatively lower performance for nodules (78.9% accuracy) aligns with clinical challenges and previous research, highlighting this as a key area for future work. The detailed error analysis, which links misclassifications to image quality and lesion characteristics, provides a clear roadmap for targeted improvements.

5.2. Clinical and methodological implications

The study's emphasis on interpretability and the positive feedback from the radiologist validation (85% increased confidence) highlight that a "glass-box" approach is essential for clinical adoption. The validation protocol itself, with its rigorous design and statistical analysis (Fleiss' Kappa, Wilcoxon test), serves as a methodological contribution for evaluating preprocessing techniques. Our ensemble

approach and detailed error analysis offer novel perspectives on handling multi-class problems and understanding model limitations beyond aggregate metrics.

5.3. Contextualizing performance and limitations

Our model's overall accuracy (84.7%) is lower than some studies reporting 94–99%. This is largely explained by our challenging multi-class task, use of a dataset with natural class imbalance, and rigorous k-fold cross-validation, as opposed to simpler binary tasks or single train-test splits on balanced data. The documented computational constraints also placed a ceiling on performance. While these factors contextualize our results, the study's primary limitations remain its relatively small dataset size (5,000 images) and single-institution origin, which may affect generalizability.

6. Conclusion and future directions

6.1. Summary of contributions

This research presents an integrated, interpretable deep learning framework for lung disease detection that demonstrates significant performance gains through a novel combination of anatomically-informed preprocessing and ensemble modeling. We provide a rigorous evaluation of our methodology, including a detailed error analysis and a clinical validation study with board-certified radiologists, which confirms the clinical value of our approach. By transparently documenting our computational constraints, we provide a realistic performance benchmark for resource-limited settings.

6.2. Recommendations for future research

Building on the findings and limitations of this study, we propose several key directions for future research to advance the field. A top priority should be conducting **multi-center validation studies**. Such studies are essential for assessing the true generalizability of AI models by testing them across diverse patient populations, imaging equipment, and clinical protocols. Another critical avenue is the **integration of clinical data**; incorporating non-imaging information such as patient history, symptoms, and laboratory values holds significant potential to improve diagnostic performance, particularly for cases with ambiguous radiological findings.

Furthermore, future work should focus on **advanced error reduction** by implementing the specific strategies identified in our analysis, including the development of quality-aware processing and bone suppression techniques. To better mimic clinical practice, developing methods for **longitudinal analysis** is also crucial. The ability to incorporate and compare with prior imaging studies would greatly enhance the detection of subtle, diagnostically significant changes over time. Finally, from a practical standpoint, research into more **resource-efficient architectures** is needed to design computationally efficient models that can be readily deployed in resource-limited clinical settings.

Ultimately, while these research avenues will refine the technology, the definitive establishment of clinical value will require **prospective randomized trials** that compare AI-augmented interpretation against the current standard of radiological practice. By pursuing these recommended research directions and addressing the limitations identified, the field can move closer to realizing the vision of AI-augmented precision radiology.

Conflict of interest

The authors declare no conflict of interest.

Funding statement

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Acknowledgement

We would like to express our sincere thanks and gratitude to University of Al-Nisour for their generous support, which had a significant impact on the completion of this research.

References

Abdulahi, A. T., Oguntade, E. S., & Adekanmbi, O. (2024). PulmoNet: A novel deep convolutional neural network for the classification of COVID-19, bacterial pneumonia, viral pneumonia, and healthy cases. *Journal of Medical Systems*, 48(1), 1–12. https://doi.org/10.1007/s10916-023-01953-0.

American Cancer Society. (2024). *Lung cancer survival rates*. Retrieved from https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/survival-rates.html.

American College of Radiology. (2024). ACR position statement on artificial intelligence in radiology. *Journal of the American College of Radiology*, 21(3), 478–480.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x.

Chen, J., Wu, L., Zhang, J., Zhang, L., Gong, D., Zhao, Y., ... & Li, S. (2021). Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Scientific Reports*, 11(1), 1–11. https://doi.org/10.1038/s41598-020-80061-2.

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2019). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*.

Ifty, I. J., Siddique, M. A. B., Islam, M. S., Rahaman, M. M., & Islam, M. R. (2024). A comprehensive study on lung disease classification using deep learning and explainable AI. *Biomedical Signal Processing and Control*, 89, 105559. https://doi.org/10.1016/j.bspc.2023.105559.

Kim, S. H., Park, H. J., & Lee, J. H. (2023). Multi-task deep learning for simultaneous classification and severity assessment of interstitial lung disease. *European Radiology*, 33(4), 2541–2550.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S. H., Guerrier, C. E., ... & Do, S. (2022). An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, 6(2), 166–178.

Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2023). Deep learning in chest radiography: Detection of findings and presence of abnormalities. *PLoS ONE*, 18(2), e0281838. https://doi.org/10.1371/journal.pone.0281838.

Li, X., Cao, L., & Glassman, E. (2020). Transfer learning for pneumonia detection on chest X-ray images. *arXiv preprint arXiv:2004.06578*.

Nascimento, D. F., Ferreira, M. F., Ramos, R. P., & Nascimento, M. Z. (2021). Lung nodule classification via deep transfer learning in CT lung images. *Computer Methods and Programs in Biomedicine*, 197, 105709. https://doi.org/10.1016/j.cmpb.2020.105709.

National Institutes of Health. (2022). *NIH Chest X-ray Dataset*. Retrieved from https://nihcc.app.box.com/v/ChestXray-NIHCC.

Panch, T., Mattie, H., & Atun, R. (2023). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*, 13, 03008. https://doi.org/10.7189/jogh.13.03008.

Ramli, N. M., Ghani, N. A. M., Isa, N. A. M., & Hamid, A. H. A. (2024). Artificial intelligence in detecting pulmonary nodules on chest radiographs: A systematic review. *Diagnostics*, 14(1), 90. https://doi.org/10.3390/diagnostics14010090.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Santos, R. D., Ferreira, J. L., & Oliveira, M. C. (2022). Predicting COPD exacerbations with deep learning on chest CT images. *Medical & Biological Engineering & Computing*, 60(1), 161–172.

Union for International Cancer Control. (2023). *Global cancer statistics 2023*. Geneva: UICC.

Walsh, S. L., Calandriello, L., Silva, M., & Sverzellati, N. (2020). Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *The Lancet Respiratory Medicine*, 8(9), 885–888. https://doi.org/10.1016/S2213-2600(20)30060-7

Wang, L., Lin, Z. Q., & Wong, A. (2020). COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, 10(1), 1–12. https://doi.org/10.1038/s41598-020-76550-z.

World Health Organization. (2023). *Global health estimates: Leading causes of death*. Retrieved from https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., ... & Wang, W. (2021). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*, 84(5), 1360–1373. https://doi.org/10.1016/j.cell.2021.01.012.

Zhang, Y., Wu, J., Chen, W., Chen, Y., & Tang, X. (2022). Anatomical context-aware deep learning for medical image segmentation. *Medical Image Analysis*, 77, 102362. https://doi.org/10.1016/j.media.2022.102362.

Zhao, W., Xu, J., Tian, G., & Zhang, Y. (2021). Deep learning-based quantitative computed tomography analysis in chronic obstructive pulmonary disease: Relationship with pulmonary function tests. *European Journal of Radiology*, 142, 109835. https://doi.org/10.1016/j.ejrad.2021.109835.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2022). A primer on deep learning in genomics. *Nature Genetics*, 54(1), 12–25. https://doi.org/10.1038/s41588-021-00984-y.

Abbas, J. K., Al-Azzawi, W., Hassan, S. I., Sharif, S. F. A., & Manee, M. J. (2024, December). Deep Learning Techniques for the Evaluation of a Financial Time Series Forecasting Model. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, 1–6. IEEE.

Abbas, J. K. K., Ruhaima, A. A., Naser, O. A., & Hayder, D. M. (2024). F-Test and One-Way ANOVA for Medical Images Diagnosis. *Al-Nisour Journal for Medical Sciences*, 6(2), 29–38.

Abdulsahib, A. A., Mahmoud, M. A., Al-Hasnawi, S. A., Almhanna, A. Z., & Abbas, J. K. K. (2025). Automated Retinal Vessel Analysis: A Novel Model for Blood Vessel Detection, Segmentation, and Clinical Characteristic Quantification. *Al-Nisour Journal for Medical Sciences*, 7(1), 65–80.