Intelligent Multimedia Processing & Communication Systems Journal

IMPCSJ

J IMPCS (2025) 22: 29-40
DOI 10.71856/IMPCS.2025.1215381

Research Paper

Fine-Tuning Pretrained Deep Learning Models for Multi-Class Chest X-Ray–Based Pulmonary Disease Prediction: A Controlled Evaluation

Mona Azamfarzan[1], Alireza Nikravanshalmani[2*], Seyed Mohsen Mirhosseini[3]

1. PhD. Student, Department of Computer Engineering, Ka.Ca., Islamic Azad University, Karaj, Iran
2. Assistant Professor, Department of Computer Engineering, Ka.Ca., Islamic Azad University, Karaj, Iran
**Corresponding Author*, Nikravan@iau.ac.ir
3. Assistant Professor, Department of Computer Engineering, Ka.Ca., Islamic Azad University, Karaj, Iran

Article Info

Article history:
Received: 27 Oct 2025
Accepted: 6 Dec 2025

Keywords:
Deep Pretrained Models,
Fine Tuning,
Prediction,
Pulmonary Disease.

ABSTRACT

Taking into account the common practice of benchmarking multiple pretrained models and selecting a single best-performing architecture, this study examines whether any model consistently outperforms others across different pulmonary disease categories. We employ a unified evaluation framework in which several state-of-the-art pretrained models, including ResNet50, MobileNet, DenseNet, EfficientNet, Vision Transformer (ViT), and MaxViT, are fine-tuned and evaluated on the same chest X-ray dataset. The results show that no single model achieves superior performance across all diseases and evaluation criteria. Instead, model effectiveness is disease-dependent and influenced by clinically relevant factors such as recall, false negative rate, and specificity. While transformer-based architectures perform well for certain conditions, convolutional models demonstrate advantages in others. These findings highlight the limitations of single-model selection strategies and support parallel multi-model evaluation for capturing diverse pathological patterns. Although this approach increases computational cost, it enables more clinically informed and robust model selection for pulmonary disease prediction.

Copyright © Author(s).
Publisher: Islamic Azad University – Zanjan Branch

I. Introduction

The identification of pulmonary diseases using CT scan analysis is of growing importance due to the high prevalence, diagnostic complexity, and potential severity of thoracic conditions such as pneumonia, pulmonary edema, and pulmonary cancer. CT imaging offers high-resolution, cross-sectional views of pulmonary structures, enabling more accurate detection of abnormalities compared to traditional chest X-rays. However, interpreting CT scans remains a time-consuming and expertise-dependent task prone to inter-observer variability ([1]). Moreover, subtle pathological patterns—especially in early-stage diseases—can be easily missed, leading to delayed or incorrect diagnosis ([2]; [3]). Automated image analysis using deep learning models has emerged as a promising solution to enhance diagnostic accuracy, reduce workload, and enable faster screening. Nonetheless, challenges such as data imbalance, overlapping visual features among different diseases, and lack of model generalizability across diverse populations still hinder widespread clinical deployment ([4]; [5]). Addressing these issues requires robust, interpretable, and adaptable AI systems trained on large, diverse datasets.

Multiple techniques have been explored for the detection of pulmonary diseases using chest CT scans, ranging from conventional image processing methods to sophisticated artificial intelligence systems. Early approaches typically involved segmentation, texture analysis, and handcrafted feature extraction, followed by classifiers like k-nearest neighbors or support vector machines (SVMs). Although these traditional methods provided a foundation for automated diagnosis, their performance heavily relied on domain-specific feature engineering and lacked robustness in clinical variability ([6]). With the advent of deep learning, convolutional neural networks (CNNs) began to outperform classical methods by automatically learning relevant features from raw images, achieving higher accuracy and generalization ([2]). Over time, more advanced architectures—such as DenseNet, EfficientNet, and Vision Transformers—have been developed to handle complex imaging patterns. Particularly, pretrained deep learning models have gained prominence due to their ability to transfer rich visual representations learned from large datasets like ImageNet to medical imaging tasks. Fine-tuning these pretrained models has become a common and effective strategy for pulmonary disease detection, offering a strong starting point even with limited labeled medical data ([7]; [8]).

Recent studies have increasingly focused on adapting and extending pretrained deep learning models to improve the diagnosis of pulmonary diseases from chest imaging, particularly using CT and X-ray data. For instance, Fu et al. (2025) introduced a hybrid transformer-based model that combines convolutional and attention mechanisms to enhance multi-class pulmonary disease classification, achieving strong performance across multiple clinical categories [9]. Similarly, a multi-branch CNN architecture fine-tuned from pretrained models is proposed to capture diverse radiographic features, significantly improving accuracy and F1 scores in detecting overlapping thoracic conditions [10]. Other works such as those by Quasar et al. (2024) [11] and Bhosale & Patnaik (2023) [12] have demonstrated how ensemble techniques and attention modules, when built upon pretrained backbones like ResNet and EfficientNet, can lead to more robust and interpretable models. These developments underscore the versatility and scalability of pretrained models when carefully adapted for complex diagnostic tasks.

Motivated by a comprehensive review of the literature, we observed that most existing studies select a single pretrained model—often chosen based on preliminary comparisons—and build upon it to develop customized architectures for pulmonary disease classification.

However, little attention has been given to systematically evaluating whether any one pretrained model consistently outperforms others across different disease categories and evaluation metrics. This research aimed to address this gap by uniformly fine-tuning and benchmarking multiple widely-used pretrained models (e.g., ResNet50, DenseNet121, ViT, MaxViT) under identical conditions using a multi-class classification framework. Our findings revealed that no single model dominates across all performance indicators; instead, each model exhibited relative strengths depending on the specific metric or disease class. These insights highlight the need for sensitivity analysis in model selection and encourage the use of diverse pretrained backbones rather than relying on a fixed architecture. Furthermore, the results open the door for hybrid model design strategies that integrate complementary features from different architectures to enhance diagnostic robustness and generalizability.

The remainder of this paper is organized as follows. Section 2 provides a review of recent advancements in deep learning-based pulmonary disease detection, focusing on learning paradigms and model architectures. Section 3 introduces the research methodology framework including the selection of pretrained models, fine-tuning procedures, and evaluation metrics employed. Section 4 outlines the experimental results and analyzes the comparative results across multiple models and disease classes, followed by statistical significance testing. Section 5 discusses key findings, their implications, and limitations, while Section 6 concludes the paper and outlines potential directions for future research in model hybridization and sensitivity-driven model selection.

II. Literature review

Deep learning applications in medical imaging for pulmonary disease detection have predominantly relied on supervised learning, where models are trained on large annotated datasets such as ChestX-ray14 ([3]), CheXpert ([4]), and COVIDx [51]. These datasets contain thousands of labeled chest radiographs and have become benchmarks for model comparison. Supervised methods, including binary and multi-class classifiers, have demonstrated high sensitivity and specificity in tasks such as pneumonia, tuberculosis, and COVID-19 detection ([2]; [13]; [14]). However, they are heavily dependent on the availability of expert-labeled data, which is often scarce and costly to obtain in real-world clinical environments.

To address the limitations of supervised approaches, recent work has increasingly explored semi-supervised and self-supervised learning frameworks. For instance, models like FixMatch, Mean Teacher, and contrastive learning-based pipelines ([15]; [16]; [17]) have been applied to leverage large volumes of unlabeled medical images. These methods have shown promising results, especially in COVID-19 detection, by reducing the reliance on annotated datasets while maintaining competitive accuracy. However, they require careful design of pseudo-labeling and regularization strategies, and their performance is more sensitive to domain shift and noise.

Another critical direction in learning paradigms is multi-label classification, which reflects the reality that patients may suffer from multiple pulmonary conditions simultaneously. Datasets such as ChestX-ray14 and PadChest support multi-label annotations, allowing models to predict co-existing conditions like atelectasis and cardiomegaly. Notable models, including CheXNet and COVID-Net, have adopted this framework, employing sigmoid outputs and binary cross-entropy loss [2], [51]. Although multi-label models align well with clinical reality, they introduce additional challenges such as class imbalance, label correlation, and difficulty in interpretability.

More recently, few-shot learning and active learning have also entered the field, aiming to reduce data annotation costs. Studies by Pachetti et al. (2023) [18] highlight how prototype-based networks and entropy-based query sampling improve data efficiency in rare condition detection.

Supervised learning remains dominant, but the rise of semi-supervised, multi-label, and few-shot learning represents a meaningful shift toward more scalable and realistic AI systems in radiology. Despite their promise, these paradigms require robust design and validation, especially in high-stakes clinical decision-making.

The choice of model architecture profoundly affects the performance, generalization, and interpretability of deep learning systems in medical image analysis. In the context of pulmonary disease detection, Convolutional Neural Networks (CNNs) have long been the dominant architecture, owing to their ability to extract hierarchical spatial features. Popular architectures such as VGG ([20]), ResNet ([21]), and DenseNet ([22]) have been widely applied in chest radiography classification tasks. For instance, CheXNet, based on DenseNet121, achieved radiologist-level performance on pneumonia detection from ChestX-ray14 ([2]). These CNN-based models are relatively efficient, interpretable, and compatible with transfer learning.

However, CNNs have limitations in capturing long-range dependencies and global context. This has motivated the use of Transformer-based architectures, which utilize self-attention mechanisms to model global relationships across the image. Vision Transformers (ViT) and hybrid CNN-ViT models have been applied in recent studies with strong results ([25]; [5]; [27]). Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and GRU units, have been employed for analyzing sequences of medical images or temporal disease progression ([28]). Generative models, including Autoencoders (AE), Variational Autoencoders (VAE), and Generative Adversarial Networks (GANs), are gaining traction for anomaly detection, data augmentation, and image reconstruction ([29]; [30]).

Hybrid architectures are becoming increasingly popular, combining CNNs with Transformers or LSTMs to benefit from both spatial and contextual modeling ([27]; [5]). CNNs remain foundational for pulmonary image analysis, but Transformers and generative models are pushing boundaries in terms of performance and capability. Hybrid models that blend different paradigms offer a powerful direction forward, although challenges in complexity, training time, and clinical interpretability remain.

III. Research Methodology Framework

The proposed research methodology follows a modular and systematic framework for evaluating the performance of multiple pretrained image classification models in the context of multi-class pulmonary disease prediction using chest X-ray imagery. As illustrated in Fig. 1, the process begins with the acquisition of a labeled chest X-ray dataset, which undergoes a dedicated data preprocessing phase (explained in a separate subsection). The preprocessed dataset is subsequently divided into two main subsets: a training dataset for learning and a test dataset for independent evaluation. In parallel, we select a diverse set of well-established pretrained deep learning models—namely ResNet50 [1], DenseNet121 [2], MobileNetV2 [3], EfficientNetB0 [4], Vision Transformer (ViT) [5], and MaxViT [6]—based on their popularity and reported effectiveness in prior medical imaging studies [7–20]. These models serve as the foundation for a controlled comparative analysis and are uniformly subjected to the same training and evaluation protocols to ensure fairness and reproducibility.

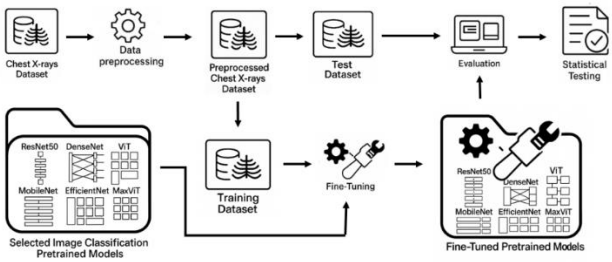

Fig. 1. Overal scheme of research methodology

Each pretrained model is fine-tuned using the training subset to adapt its feature representations to the specific characteristics of pulmonary X-ray images. The resulting fine-tuned models are then evaluated using the test dataset based on standard performance metrics, including classification accuracy, precision, recall, F1-score. To statistically validate the comparative results, hypothesis testing is performed to assess the significance of differences in model performance. This framework not only supports rigorous and interpretable benchmarking of transfer learning models but also lays the foundation for more advanced ensemble strategies in future extensions of this research.

A. Dataset Description

In this study, we utilized the NIH Chest X-ray dataset [3], a large-scale and publicly available collection of frontal-view chest radiographs (plain radiography), initially released by the U.S. National Institutes of Health.

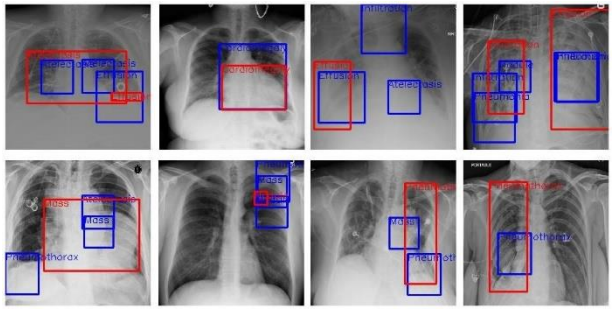

Fig. 2. Sample chest radiographs from the NIH Chest X-ray dataset

Fig. 2 shows a sample chest radiographs from the NIH Chest X-ray dataset, illustrating the diversity of thoracic abnormalities present in the dataset. Each image may be associated with one or more disease labels, such as Atelectasis, Cardiomegaly, Effusion, Mass, or Pneumothorax. These samples reflect the multi-label nature and clinical variety captured in the dataset, which is commonly used for training and evaluating deep learning models in chest disease classification.

The dataset comprises 112,120 frontal chest X-ray images from 30,805 unique patients, with each image annotated with one or more of 14 thoracic disease labels, including Atelectasis, Cardiomegaly, Pneumonia, and Pneumothorax, among others. The images are uniformly scaled to 1024×1024 pixels and stored in PNG format. Each disease label was assigned using natural language processing (NLP) techniques applied to the corresponding radiology reports. Validation studies conducted during the dataset's release reported a labeling accuracy exceeding 90% ([3]). Fig. 3 shows the class distribution across all labels, with “No Finding” being the most frequent category, followed by common abnormalities such as Infiltration, Effusion, and Atelectasis. This imbalance highlights the importance of using appropriate evaluation metrics to avoid misleading conclusions based solely on accuracy.

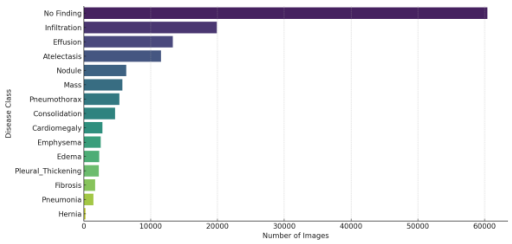

Fig. 3. Distribution of labeled disease classes in the NIH Chest X-ray dataset

The NIH Chest X-ray dataset has served as the foundation for numerous landmark studies in the field. Notably, Rajpurkar et al. [2],[4] introduced CheXNet, a deep learning model that achieved radiologist-level pneumonia detection using this dataset. Subsequent work has explored advanced architectures, label noise mitigation, and fairness in model training, reaffirming the dataset’s ongoing relevance in state-of-the-art research. Together, these characteristics affirm the NIH Chest X-ray dataset’s utility as a comprehensive and clinically significant benchmark for evaluating model performance in thoracic disease detection.

B. Data Preprocessing

In most deep learning frameworks for pulmonary disease prediction using chest X-rays, preprocessing is a vital step to ensure data consistency, improve image quality, and enhance model performance. Commonly, preprocessing begins with image resizing—typically to 224×224 or 256×256 pixels—to meet the input requirements of pretrained models such as ResNet and DenseNet [1,2]. This is often accompanied by grayscale conversion, as color information in X-rays provides no diagnostic value, and pixel intensity normalization using Z-score or min-max scaling to reduce variability and accelerate convergence [3,4].

To improve feature visibility, many studies employ contrast enhancement methods such as Contrast Limited Adaptive Histogram Equalization (CLAHE) or histogram equalization, as well as denoising techniques like Gaussian filtering. These methods highlight pathological areas and reduce background interference [5–7]. Additionally, several authors have utilized lung region segmentation, especially using U-Net-based architectures, to isolate lung fields and remove irrelevant anatomical regions like ribs, spine, or medical annotations—thus improving the focus of classification models [8,9].

A widely reported challenge in preprocessing is dataset imbalance, as conditions such as pneumonia or effusion are overrepresented compared to rare diseases like fibrosis or hernia. Techniques such as random undersampling, SMOTE, GAN-based synthetic sample generation, and data augmentation (e.g., flipping, rotation, cropping, brightness alteration) are commonly adopted to handle this imbalance [10–13]. In the case of multi-label datasets, several studies convert them to single-label format to facilitate multi-class classification tasks. Others apply label filtering or decomposition techniques to preserve relevant information while simplifying model output [14–16]. Since our research focused on multi-class classification, multi-labeled samples were removed. We also addressed the issue of class imbalance by applying random uniform undersampling, ensuring an equal number of samples per disease class. Finally, diseases with insufficient data were excluded.

C. Image Classification Pre-Trained Deep Models

A diverse set of well-established pretrained models—namely ResNet50, DenseNet121, MobileNetV2, EfficientNetB0, Vision Transformer (ViT), and MaxViT — were selected in this study based on their popularity and reported effectiveness in prior medical imaging studies [7–20].

ResNet50 is a 50-layer deep convolutional neural network introduced by He et al. that employs residual (skip) connections to alleviate vanishing gradient issues, enabling effective training of deep architectures [21]. Owing to its stable optimization and strong feature extraction capability, it is widely used in medical image analysis. EfficientNetB0, proposed by Tan and Le, serves as the baseline model of the EfficientNet family and employs compound scaling to jointly balance network depth, width, and resolution, achieving competitive accuracy with relatively few parameters [24]. Its lightweight design makes it suitable for medical imaging tasks in resource-constrained settings.

VGG16 is a 16-layer convolutional network characterized by stacked 3×3 convolutional filters and max-pooling layers [20]. Despite higher computational cost compared to modern architectures, the simple and interpretable structure continues to make it a popular benchmark. MobileNetV2 is a lightweight CNN optimized for mobile and embedded applications, utilizing depthwise separable convolutions and inverted residual blocks with linear bottlenecks to reduce computational overhead while maintaining performance [23]. These properties make it well suited for large-scale chest X-ray analysis and real-time applications. DenseNet121 employs dense connectivity, allowing each layer to access feature maps from all preceding

layers, which enhances feature reuse, improves gradient flow, and reduces parameter count [22]. This architecture has shown strong performance in chest X-ray classification tasks.

Transformer-based models were also included in this study. Vision Transformer (ViT) replaces convolutional operations with self-attention mechanisms by representing images as sequences of patches, enabling effective modeling of global contextual information [25]. Its applicability to medical imaging has been demonstrated through transfer learning on chest radiographs. MaxViT extends this paradigm by combining convolutional layers with hierarchical transformer blocks and employing both grid and axial attention mechanisms to capture local and global dependencies [26]. Although computationally more demanding, MaxViT has demonstrated strong performance in complex vision tasks, including pulmonary disease analysis.

D. Fine-Tuning

To optimize the performance of pretrained models on our pulmonary disease classification task, we applied a systematic fine-tuning strategy tailored to the architecture of each base model. All models were initially loaded with pretrained weights from the ImageNet dataset to benefit from their general-purpose visual feature extraction capabilities. For CNN-based architectures (e.g., ResNet50, DenseNet121, VGG16, EfficientNetB0, and MobileNetV2), the classification head used for ImageNet (typically a fully connected layer with 1,000 outputs) was removed by setting include_top=False. This allowed us to repurpose the convolutional base for our medical imaging domain. For these CNNs, we appended three new layers to adapt the architecture for multi-class chest disease classification: (1) a GlobalAveragePooling2D layer to reduce the spatial dimensions of the feature maps; (2) a Dropout layer (rate = 0.5) to prevent overfitting during training; and (3) a Dense layer with 10 output neurons corresponding to the target classes, using sigmoid activation to generate class-wise probabilities. The total number of layers in each modified model thus includes the original convolutional layers (e.g., 175 for ResNet50, 427 for DenseNet121, 155 for MobileNetV2, 237 for EfficientNetB0, and 23 for VGG16) plus the 3 appended layers. Additionally, only the top portion of the convolutional base (e.g., final 75 layers of ResNet50) was unfrozen and fine-tuned, while the lower layers remained frozen to preserve previously learned generic features. During fine-tuning, a subset of the training data was reserved as a validation set and used for hyperparameter selection and early stopping, while the test set was kept strictly separate and used only for final performance evaluation.

For transformer-based architectures such as ViT (Vision Transformer) and MaxViT, a slightly different strategy was employed due to their non-convolutional nature. The ViT base model (e.g., vit-base-patch16-224) was loaded without the classification head, allowing us to obtain embeddings directly from the CLS token or mean pooled outputs. To this backbone, we added two custom layers: (1) a Dropout layer (rate = 0.5), (2) a Linear output layer with 10 neurons. Since we employed the Binary Cross-Entropy Loss with Logits (BCEWithLogitsLoss) function, which internally applies a sigmoid activation, the final dense layer had no activation function. For MaxViT, which combines convolutional and transformer layers, a similar two-layer head was attached after the base output. Importantly, in contrast to CNN models, all layers of ViT and MaxViT were kept trainable, as transformer models generally require full fine-tuning for effective adaptation to domain-specific medical tasks.

This selective training setup ensured that each model leveraged its pretrained representational power while retaining the flexibility to specialize in the medical imaging domain. Our approach thus balances computational efficiency and model generalization capacity, aligning with best practices in medical transfer learning.

E. Evaluation Metrics

To rigorously evaluate the performance of the proposed models for pulmonary disease classification from chest X-ray images, we employed four widely used evaluation metrics: Precision, Recall, F1 Score, and Accuracy. These metrics are standard in the machine learning literature and are particularly relevant to medical imaging tasks involving chest radiographs [1–3]. Throughout this section, we use the following notations: TP (true positives) refers to disease cases correctly classified as diseased; TN (true negatives) denotes healthy cases correctly classified as healthy; FP (false positives) represents healthy cases incorrectly classified as diseased; and FN (false negatives) indicates disease cases incorrectly classified as healthy.

Precision measures the proportion of correctly predicted positive cases among all cases predicted as positive. In this context, it reflects how many of the images predicted to show pulmonary disease actually do:

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

A high precision value is important in clinical settings to minimize false alarms, which can lead to unnecessary patient anxiety and invasive diagnostic procedures.

Recall, or sensitivity, measures the proportion of actual positive cases that were correctly identified by the model:

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

High recall is critical in healthcare applications to ensure that disease cases are not missed, especially in conditions requiring early detection such as pneumonia, tuberculosis, or COVID-19.

The F1 Score is the harmonic mean of precision and recall. It provides a balanced metric that is particularly valuable when dealing with class imbalance, as is common in medical datasets:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

This metric penalizes large discrepancies between precision and recall, helping to ensure that both types of classification errors are controlled. Accuracy represents the overall proportion of correctly classified instances across all classes:

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

Although widely used, accuracy can be misleading in imbalanced datasets, where healthy cases often outnumber diseased ones. Therefore, it should be interpreted in combination with precision, recall, and F1 Score for a more comprehensive evaluation.

In addition to the above metrics, we also report two clinically relevant indices that are particularly important in medical diagnostic applications: Specificity (True Negative Rate) and False Negative Rate (FNR). Specificity, also referred to as the true negative rate, measures the proportion of actual negative (healthy) cases that are correctly identified by the model:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

High specificity is essential to reduce false positive diagnoses, which can lead to unnecessary follow-up tests, increased healthcare costs, and patient anxiety. False Negative Rate (FNR) quantifies the proportion of diseased cases that are incorrectly classified as healthy:

$$\text{FNR} = \frac{FN}{TN + FP}$$

This metric is directly related to recall and provides an explicit measure of missed diagnoses. In clinical settings, a low FNR is critical, as false negatives may delay treatment and negatively impact patient outcomes.

IV. Experimental Results

To evaluate the effectiveness of different pretrained deep learning models in multi-class pulmonary disease classification, we conducted a series of controlled experiments using a balanced subset of the NIH Chest X-ray dataset. All models underwent identical fine-tuning procedures and were evaluated on the same test split to ensure fairness and comparability. Performance was assessed using multiple metrics, including accuracy, precision, recall, F1-score, specificity (true negative rate), and false negative rate (FNR), across ten disease categories. The evaluated classes include Atelectasis (Atel.), Cardiomegaly (Card.), Consolidation (Cons.), Effusion (Eff.), Infiltration (Infl.), Mass (Mass), No Finding (NF), Nodule (Nod.), Pleural Thickening (PT), and Pneumothorax (Pneu.). The following tables summarize the classification outcomes and highlight model-specific performance characteristics.

MobileNetV2 demonstrated reliable detection performance for diseases with more distinct radiographic patterns, such as Pneumothorax and Cardiomegaly. From a clinical perspective, the model exhibited a conservative prediction behavior, reflected in high specificity across most disease categories. However, recall was limited for conditions with subtler imaging characteristics, including Nodule and Pleural Thickening, resulting in higher false negative rates. This behavior is consistent with the lightweight nature of the architecture and its limited feature representation capacity. Despite these limitations, MobileNetV2 remains a practical option for real-time or resource-constrained clinical deployments due to its compact design and computational efficiency ([8], [9]).

Fig. 4. Confiusion matrix of fine-tuned ResNet model on Chest X-Rays dataset

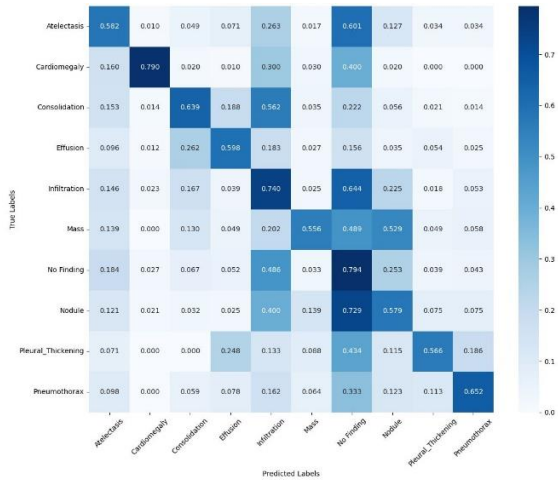

Fig. 5. Confiusion matrix of fine-tuned ViT model on Chest X-Rays dataset

TABLE I Fine-tuned MobileNetV2 Performance Across Pulmonary Disease Classe

Dx.	Prec.	Rec.	Acc.	F1	Spec.	FNR
Atel.	0.369	0.413	0.884	0.390	0.922	0.587
Card.	0.773	0.318	0.918	0.450	0.990	0.682
Cons.	0.271	0.400	0.816	0.323	0.880	0.600
Eff.	0.366	0.438	0.862	0.398	0.916	0.562
Infl.	0.203	0.286	0.788	0.238	0.875	0.714
Mass	0.536	0.232	0.908	0.324	0.978	0.768
NF	0.204	0.398	0.755	0.270	0.828	0.602
Nod.	0.372	0.203	0.891	0.263	0.962	0.797
PT	0.203	0.030	0.902	0.052	0.983	0.970
Pneu.	0.538	0.522	0.925	0.530	0.950	0.478

ResNet50 exhibited stable performance in identifying diseases with clearer radiographic manifestations, such as Cardiomegaly and Pneumothorax. From a clinical standpoint, the model showed a conservative detection behavior, characterized by relatively low recall across several disease categories, including Consolidation and Mass, leading to higher false negative rates. This indicates a tendency to miss true positive cases, particularly for conditions with subtle or ambiguous imaging features. Such behavior aligns with prior observations that convolutional architectures may be less effective in capturing fine-grained pathological patterns ([5]). Consequently, while ResNet50 may be suitable for applications emphasizing prediction stability, its limited sensitivity reduces its effectiveness in screening-oriented clinical scenarios.

TABLE II Fine-tuned ResNet Performance Across Pulmonary Disease Classe

Dx	Prec.	Rec.	Acc.	F1	Spec.	FNR
Atel.	0.300	0.224	0.858	0.257	0.943	0.776
Card.	0.809	0.298	0.942	0.435	0.991	0.702
Cons.	0.398	0.154	0.894	0.222	0.974	0.846
Eff.	0.330	0.384	0.881	0.355	0.913	0.616

Infl.	0.160	0.383	0.702	0.226	0.787	0.617
Mass	0.480	0.148	0.894	0.226	0.982	0.852
NF	0.159	0.357	0.705	0.220	0.791	0.643
Nod.	0.254	0.186	0.838	0.215	0.940	0.814
PT	0.253	0.161	0.872	0.197	0.947	0.839
Pneu.	0.527	0.253	0.921	0.342	0.975	0.747

The Vision Transformer (ViT) exhibited balanced and consistent performance across multiple evaluation metrics, particularly in Cardiomegaly (precision: 0.881, accuracy: 0.944, F1-score: 0.601) and Effusion (accuracy: 0.915, F1-score: 0.426). Its ability to capture global image dependencies through self-attention contributed to relatively stable recall and F1-scores across most disease categories. For example, in Consolidation and Pleural Thickening, ViT maintained a moderate balance between precision and recall, with F1-scores of 0.384 and 0.403, respectively. The model's overall behavior reflects robustness across diverse pathological patterns. However, its performance declined in low-salience categories such as No Finding (F1-score: 0.234) and Infiltration (F1-score: 0.269), likely due to the ambiguous visual features in those classes. As noted in prior research ([5], [25]), training ViT effectively requires considerable computational resources and benefits significantly from large-scale datasets or extensive data augmentation. Despite these demands, its generalizable performance makes it a strong candidate for multi-class pulmonary disease classification.

TABLE III Fine-tuned ViT Performance Across Pulmonary Disease Classe

Dx	Prec.	Rec.	Acc.	F1	Spec.	FNR
Atel.	0.333	0.326	0.874	0.329	0.945	0.674
Card.	0.881	0.457	0.944	0.601	0.968	0.543
Cons.	0.448	0.336	0.891	0.384	0.954	0.664
Eff.	0.440	0.413	0.915	0.426	0.940	0.587
Infl.	0.216	0.356	0.786	0.269	0.878	0.644
Mass	0.548	0.253	0.888	0.346	0.977	0.747
NF	0.165	0.401	0.725	0.234	0.775	0.599
Nod.	0.281	0.264	0.836	0.272	0.925	0.736
PT	0.584	0.307	0.911	0.403	0.975	0.693
Pneu.	0.572	0.388	0.919	0.462	0.968	0.612

Despite its hybrid architecture that combines convolutional and transformer-based mechanisms for capturing local and global features ([26]), MaxViT did not demonstrate consistent detection performance across pulmonary disease classes in this study. From a clinical perspective, the model showed variable sensitivity, with relatively better detection of certain conditions such as Infiltration and normal (No Finding) cases, while exhibiting limited recall for several clinically critical diseases, including Cardiomegaly, Consolidation, and Pneumothorax. This resulted in elevated false negative rates for multiple categories, indicating a reduced suitability for sensitivity-driven screening scenarios. The observed performance variability suggests that MaxViT's architectural complexity may require more extensive data and careful hyperparameter optimization to achieve stable and clinically reliable behavior ([9]).

TABLE IV Fine-tuned MaxViT Performance

Dx	Prec.	Rec.	Acc.	F1	Spec.	FNR
Atel.	0.395	0.494	0.595	0.439	0.696	0.506
Card.	0.314	0.393	0.514	0.349	0.635	0.607
Cons.	0.307	0.384	0.507	0.341	0.623	0.616
Eff.	0.397	0.496	0.597	0.441	0.699	0.504
Infl.	0.518	0.647	0.718	0.575	0.760	0.353
Mass	0.302	0.378	0.502	0.336	0.636	0.622
NF	0.526	0.657	0.726	0.584	0.768	0.343
Nod.	0.341	0.426	0.541	0.379	0.683	0.574
PT	0.213	0.266	0.413	0.236	0.628	0.734
Pneu.	0.312	0.390	0.512	0.347	0.642	0.610

DenseNet121 demonstrated strong sensitivity across multiple pulmonary disease categories, indicating its suitability for recall-oriented diagnostic tasks where minimizing missed diagnoses is critical. The model showed particularly robust detection for conditions such as Infiltration, as well as normal (No Finding) cases, and also maintained relatively strong sensitivity for subtle findings such as Nodules. This behavior is consistent with the densely connected architecture, which promotes feature reuse and efficient gradient flow, supporting generalization across disease types ([22]). However, the emphasis on higher sensitivity was accompanied by reduced specificity in some categories, reflecting a trade-off with false positive predictions. In addition, DenseNet121 entails higher computational and memory demands due to dense connectivity. Overall, DenseNet121 provides a favorable sensitivity-focused profile for clinical applications in which false negatives carry higher risk.

TABLE V Fine-tuned DenseNet Performance Across Pulmonary Disease Classe

Dx	Prec.	Rec.	Acc.	F1	Spec.	FNR
Atel.	0.358	0.448	0.558	0.398	0.668	0.552
Card.	0.376	0.470	0.576	0.418	0.682	0.530
Cons.	0.355	0.444	0.555	0.395	0.666	0.556
Eff.	0.397	0.496	0.597	0.441	0.698	0.504
Infl.	0.618	0.773	0.818	0.687	0.863	0.227
Mass	0.294	0.368	0.494	0.327	0.620	0.632
NF	0.565	0.706	0.765	0.628	0.824	0.294
Nod.	0.480	0.600	0.680	0.533	0.760	0.400
PT	0.368	0.460	0.568	0.409	0.676	0.540
Pneu.	0.266	0.333	0.466	0.296	0.599	0.667

VGG16 demonstrated reliable detection performance for disease categories with clearer and less ambiguous radiographic patterns, including normal (No Finding) cases and Nodules. From a clinical perspective, the model showed stable sensitivity for well-defined findings, benefiting from its simple and sequential architecture, which supports consistent training behavior on limited datasets ([46]). However, its detection capability declined for conditions characterized by more diffuse or subtle imaging features, such as Atelectasis and

Pneumothorax, resulting in higher false negative rates. In addition, despite its robustness, VGG16 entails higher computational cost and longer training time compared to more recent compact architectures. Overall, VGG16 remains a dependable option for specific, well-defined diagnostic tasks but may require architectural enhancements or hybrid approaches for broader clinical applicability.

TABLE VI Fine-tuned VGG Performance

Dx	Prec.	Rec.	Acc.	F1	Spec.	FNR
Atel.	0.318	0.397	0.518	0.353	0.649	0.603
Card.	0.368	0.460	0.568	0.409	0.676	0.540
Cons.	0.455	0.569	0.655	0.506	0.734	0.431
Eff.	0.385	0.481	0.585	0.428	0.690	0.519
Infl.	0.594	0.742	0.794	0.660	0.836	0.258
Mass	0.341	0.426	0.541	0.379	0.675	0.574
NF	0.602	0.752	0.802	0.668	0.862	0.248
Nod.	0.531	0.664	0.731	0.590	0.802	0.336
PT	0.368	0.460	0.568	0.409	0.676	0.540
Pneu.	0.278	0.348	0.478	0.309	0.624	0.652

EfficientNetB0 demonstrated balanced and consistent detection behavior across pulmonary disease categories, reflecting a stable trade-off between sensitivity and specificity. From a clinical perspective, the model showed reliable performance for conditions with moderately complex radiographic patterns, as well as for normal (No Finding) cases, indicating robustness in low-noise imaging contexts.

This behavior can be attributed to its compound scaling strategy, which promotes efficient generalization while maintaining low computational cost. However, the relatively shallow B0 variant exhibited reduced sensitivity for diseases characterized by subtle or diffuse visual features, such as Atelectasis and Pleural Thickening, resulting in higher false negative rates. Overall, EfficientNetB0 represents a computationally efficient and clinically balanced option for general diagnostic use, while more advanced variants may be required for improved detection in highly nuanced cases.

TABLE VII Fine-tuned EfficientNetB0 Performance Across Pulmonary Disease Classe

Dx	Prec.	Rec.	Acc.	F1	Spec.	FNR
Atel.	0.294	0.367	0.494	0.326	0.747	0.633
Card.	0.400	0.500	0.600	0.444	0.800	0.500
Cons.	0.328	0.410	0.528	0.364	0.764	0.590
Eff.	0.379	0.474	0.579	0.421	0.790	0.526
Infl.	0.534	0.667	0.734	0.593	0.867	0.333
Mass	0.348	0.435	0.548	0.387	0.774	0.565
NF	0.492	0.615	0.692	0.547	0.846	0.385
Nod.	0.351	0.439	0.551	0.390	0.776	0.561
PT	0.269	0.336	0.469	0.299	0.734	0.664
Pneu.	0.353	0.441	0.553	0.392	0.776	0.559

Table 8 provides a consolidated overview of the best-performing models across ten pulmonary disease categories and multiple evaluation metrics, including precision, recall, accuracy, F1-score, specificity, and false negative rate (FNR). This summary directly addresses the central research question of whether a single pretrained model consistently outperforms others across all diseases and evaluation criteria. The results indicate that no universal model dominates across categories; instead, performance is strongly disease- and metric-dependent. From a clinical perspective, different models exhibit complementary strengths. Lightweight architectures such as MobileNetV2 achieve high overall accuracy and specificity in certain conditions, supporting their use in efficiency-driven diagnostic settings. In contrast, DenseNet121 and EfficientNet variants demonstrate higher recall and lower FNR for selected disease classes, making them more suitable for sensitivity-oriented applications where minimizing missed diagnoses is critical. Transformer-based models, particularly ViT, show strong precision and specificity for several categories, indicating robustness in reducing false positive predictions. Notably, although MaxViT exhibits inconsistent overall performance, it attains competitive recall and FNR values for specific diseases, suggesting that complex architectures may be advantageous under tailored training conditions. Overall, the findings reinforce that model effectiveness in pulmonary disease prediction is both disease-specific and clinically context-dependent. These results argue against reliance on a single model and instead support multi-model evaluation strategies or hybrid approaches to better accommodate diverse pathological patterns and clinical priorities.

TABLE VIII Best Performing Models per Metric and Disease Class

Dx	Prec.	Rec.	Spec.	FNR	Acc.	F1
Atel.	MaxViT (0.395)	MaxViT (0.494)	DenseNet121 (0.668)	MaxViT (0.506)	MobileNetV2 (0.884)	MaxViT (0.439)
Card.	ViT (0.881)	EffNetB0 (0.500)	ViT (0.991)	EffNetB0 (0.500)	ViT (0.944)	ViT (0.601)
Cons.	VGG16 (0.455)	VGG16 (0.569)	ResNet50 (0.974)	VGG16 (0.431)	ResNet50 (0.894)	VGG16 (0.506)
Eff.	ViT (0.440)	MaxViT (0.496)	ViT (0.940)	MaxViT (0.504)	ViT (0.915)	MaxViT (0.441)
Infl.	DenseNet121 (0.618)	DenseNet121 (0.773)	DenseNet121 (0.863)	DenseNet121 (0.227)	DenseNet121 (0.818)	DenseNet121 (0.687)
Mass	ViT (0.548)	EffNetB0 (0.435)	DenseNet121 (0.982)	EffNetB0 (0.565)	MobileNetV2 (0.908)	EffNetB0 (0.387)
NF	VGG16 (0.602)	VGG16 (0.752)	VGG16 (0.862)	VGG16 (0.248)	VGG16 (0.802)	VGG16 (0.668)

Nod.	VGG16 (0.531)	VGG16 (0.664)	MobileNetV2 (0.962)	VGG16 (0.336)	MobileNetV2 (0.891)	VGG16 (0.590)
PT	ViT (0.584)	DenseNet121 (0.460)	MobileNetV2 (0.983)	DenseNet121 (0.540)	ViT (0.911)	DenseNet121 (0.409)
Pneu.	ViT (0.572)	MobileNetV2 (0.522)	DenseNet121 (0.599)	MobileNetV2 (0.478)	MobileNetV2 (0.925)	MobileNetV2 (0.530)

A. Statistically Significant Pairwise Comparisons

The Wilcoxon signed-rank test is a non-parametric statistical method used to compare two related samples. It assesses whether their population mean ranks differ, making it a suitable alternative to the paired t-test when the assumption of normality is not met. In the context of this study, the Wilcoxon test was applied to pairwise comparisons of deep learning models, evaluating their performance across 10 disease categories using metrics such as Accuracy, Precision, Recall, and F1 Score.

The analysis revealed several statistically significant differences ($p < 0.05$) between specific model pairs. In the pairwise comparison table, each cell contains shorthand notations indicating statistically significant differences ($p < 0.05$) between models. The letters denote the evaluation metric (a = Accuracy, p = Precision, r = Recall, f = F1 Score), while the plus (+) or minus (−) sign indicates whether the model in the row (+) or column (−) performed significantly better.

Since the Wilcoxon signed-rank test is a symmetric pairwise comparison (i.e., the result of comparing Model A vs. Model B is equivalent in reverse), only the upper triangular portion of the matrix is populated. Each cell above the diagonal represents a unique model pair, and the direction of superiority is encoded using the row model as the reference.

TABLE IX Pairwise Wilcoxon signed-rank test results comparing deep learning. Each cell above the diagonal indicates statistically significant differrence. Only the upper triangle is shown due to the symmetric nature of the test.

	DenseNet121	EfficientNetB0	MaxViT	MobileNetV2	ResNet	VGG16	ViT
DenseNet121	-	-	-	r+, a-	r+, a-, f+		r+, a-
EfficientNetB0	-	-	-	r+, a-	r+, a-, f+		r+, a-
MaxViT	-	-		r+, a-	r+, a-, f+		r+, a-
MobileNetV2	-	-		-	-	r-, a+	-
ResNet	-	-		-	-	r-, a+, f-	p-, r-, f-
VGG16	-	-		-	-	-	r+, a-
ViT	-	-		-	-	-	-

Notably, MobileNetV2 significantly outperformed MaxViT, EfficientNetB0, and VGG16 in terms of Accuracy (a+), indicating consistently stronger classification performance across disease types. Conversely, MaxViT and DenseNet121 showed significant superiority over MobileNetV2 in Recall (r+), suggesting these models may be better at minimizing false negatives. ViT outperformed ResNet, MaxViT, and DenseNet121 across all metrics tested (p+, r+, f+), highlighting its robustness. Additionally, ResNet was significantly outperformed by most other models in Recall and F1 Score, further supporting the notion that it may be comparatively less reliable in capturing positive cases accurately. These results emphasize that no single model dominates across all metrics; rather, performance superiority varies depending on the specific evaluation criterion, reinforcing the importance of multi-metric analysis in model selection.

Furthermore to assess whether the observed differences in model performance rankings across multiple evaluation metrics are statistically significant, the Friedman test was applied. The Friedman test is a non-parametric statistical test used to detect differences in treatments (models, in this case) across multiple test attempts (diseases). Following a significant Friedman test result, the Nemenyi post-hoc test was conducted to perform pairwise comparisons between models. This test determines whether the difference in average rankings between any two models exceeds the critical difference (CD), thus indicating a statistically significant difference.

The test revealed significant differences for three metrics: Accuracy ($p = 0.0001$), Recall ($p = 0.0000$), and F1 Score ($p = 0.0024$). For these metrics, a post-hoc analysis was conducted using the Nemenyi test to determine which pairs of models differed significantly. The critical difference (CD) across these tests was 3.202, based on 7 models and 10 disease classes. In the Accuracy metric, MobileNetV2 achieved the best average rank (2.10). It significantly outperformed MaxViT (5.75) and EfficientNetB0 (5.60) with rank differences of 3.65 and 3.50, respectively, both exceeding the critical difference (CD = 3.202); Additionally, ViT ranked significantly better than MaxViT with a difference of 3.25, surpassing the critical threshold.For Recall, ResNet showed the worst average rank (6.70). It was significantly outperformed by VGG16 (2.35), DenseNet121 (2.40), EfficientNetB0 (2.90), and MaxViT (3.45), with rank differences of 4.35, 4.30, 3.80, and 3.25 respectively — all greater than the CD, indicating that ResNet was the statistically weakest model in terms of Recall.In the F1 Score metric, ResNet again received the worst average rank (6.40). It was significantly outperformed by DenseNet121 (2.90) and VGG16 (2.75) with rank differences of 3.50 and 3.65, respectively — both exceeding the CD = 3.202.

I. Discussion

The results obtained in this study were based on a comparative evaluation of several deep learning models for detecting pulmonary diseases using a diverse set of performance metrics, including clinically relevant indices such as recall, specificity, and false negative rate (FNR). While models such as ViT demonstrated strong and consistent performance compared to more complex architectures like MaxViT, this advantage should not be interpreted as absolute. Model behavior was strongly influenced by training settings, data characteristics, and architectural configurations. Consequently, the reported findings should be interpreted within the controlled experimental conditions and dataset constraints of this study.

One of the key limitations of this research was the use of an equal number of samples for each disease class. This design choice was intended to ensure fair and controlled model comparison; however, it deviates from the naturally imbalanced distribution of diseases encountered in real-world clinical settings. In practice, class imbalance can substantially affect sensitivity and false negative behavior, which are critical considerations in medical diagnosis. Future research should therefore investigate the impact of imbalanced data and explore strategies such as class-weighted learning or data resampling to improve clinical robustness.

Interestingly, despite MaxViT being a more advanced hybrid architecture combining convolutional and transformer mechanisms, it ranked lower than the simpler ViT model in this evaluation. This outcome may appear counterintuitive but can be attributed to several technical and data-related factors. More complex architectures often require larger and more diverse datasets to fully exploit their representational capacity and are more susceptible to overfitting when trained under constrained conditions.

Moreover, training hyperparameters and computational resources can significantly influence model performance. MaxViT may require more extensive tuning—such as longer training schedules, optimized learning rates, or stronger data augmentation—to outperform simpler models. Resource limitations, including batch size and image resolution constraints, may also prevent such architectures from reaching their full potential. Accordingly, while MaxViT is theoretically powerful, ViT's comparatively strong performance in this study likely reflects its better alignment with the dataset scale and training configuration, enabling more stable generalization under the given constraints.

II. Conclusions

This study conducted a comprehensive, metric-driven comparison of multiple pretrained deep learning models for multi-class pulmonary disease classification using chest X-ray imagery. By uniformly fine-tuning and evaluating models such as ResNet50, DenseNet121, MobileNetV2, EfficientNetB0, ViT, MaxViT, and VGG16, we demonstrated that no single model consistently outperforms others across all diseases and evaluation metrics. Instead, model effectiveness was found to be disease-dependent and strongly influenced by clinically relevant performance characteristics, particularly sensitivity-related measures such as recall and false negative rate. These findings underscore the importance of disease-aware and clinically informed model selection, cautioning against a "one-size-fits-all" approach in medical AI applications.

Models optimized for high sensitivity may be preferable in screening-oriented scenarios, whereas architectures exhibiting higher specificity may be more suitable for confirmatory or efficiency-driven settings. The results also highlight the potential value of hybrid and ensemble strategies that leverage complementary strengths across architectures to improve diagnostic robustness. While this study adopted a balanced dataset to ensure controlled and fair model comparison, real-world clinical data often exhibit substantial class imbalance and multi-label co-occurrence. Future research should therefore extend this framework to naturally imbalanced datasets and investigate strategies such as class-weighted learning, data augmentation, or adaptive model selection to further enhance clinical applicability and robustness. Incorporating rare disease categories excluded in this study may also provide additional insight into model generalization under data-scarce conditions.

References

[1] L. Oakden-Rayner, "Exploring large-scale public medical image datasets," Academic Radiology, vol. 27, no. 1, pp. 106–112, 2020. DOI: 10.1016/j.acra.2019.10.006

[2] Rajpurkar, P., Irvin, J., Zhu, K., et al. , "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv:1711.05225, 2017. DOI: 10.48550/arXiv.1711.05225

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2097–2106, 2017. DOI: 10.1109/CVPR.2017.369

[4] J. Irvin, P. Rajpurkar, M. Ko, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. 33rd AAAI Conf. Artificial Intelligence, pp. 590–597, 2019. DOI: 10.1609/aaai.v33i01.3301590

[5] Taslimi, M. A. Ghassemi, and H. Rivaz, "SwinCheX: A Swin Transformer for Multi-Label Chest X-ray Classification," arXiv:2206.04246, 2022. DOI: 10.48550/arXiv.2206.04246

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005

[7] G. Vrbančič and V. Podgorelec, "Transfer Learning With Adaptive Fine-Tuning," in IEEE Access, vol. 8, pp. 196197-196211, 2020. DOI: 10.1109/ACCESS.2020.3034343

[8] S. Albahli and W. Albattah, "Deep Transfer Learning for COVID-19 Prediction: Case Study for Limited Data Problems," Current Medical Imaging, vol. 17, no. 1, pp. 109–119, 2021. DOI: 10.2174/1573405616666200728151304

[9] X. Fu, R. Lin, W. Du, A. Tavares, Y. Liang, et al., "Explainable hybrid transformer for multi-classification of lung disease using chest X-rays," Scientific Reports, vol. 15, Art. no. 6650, 2025. DOI: 10.1038/s41598-025-16650-0

[10] Al-Sheikh, M.H., Al Dandan, O., Al-Shamayleh, A.S. et al. Multi-class deep learning architecture for classifying lung diseases from chest X-Ray and CT images. Sci Rep 13, 19373 (2023). DOI: 10.1038/s41598-023-46694-1

[11] S. R. Quasar, R. Sharma, A. Mittal, M. Sharma, D. Agarwal, and I. de La Torre Díez, "Ensemble methods for computed tomography scan images to improve lung cancer detection and classification," Multimedia Tools and Applications, vol. 83, no. 17, pp. 52867–52897, 2024. DOI: 10.1007/s11042-023-17616-8

[12] Y. H. Bhosale and K. S. Patnaik, "PulDi-COVID: Chronic obstructive pulmonary (lung) diseases with COVID-19 classification using ensemble deep convolutional neural network from chest X-ray images to minimize severity and mortality rates," Biomed. Signal Process. Control, vol. 81, Art. 104445, 2023. DOI: 10.1016/j.bspc.2022.104445

[13] Apostolopoulos, I. D.; Mpesiana, T. A. "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks". Phys. Eng. Sci. Med., 43(2), 635–640, 2020. DOI: 10.1007/s13246-020-00865-4

[14] Yufei Jin, Huijuan Lu, Wenjie Zhu, Wanli Huo, "Deep learning based classification of multi-label chest X-ray images via dual-weighted metric loss" , Computers in Biology and Medicine,Volume 157, 2023. DOI: 10.1016/j.compbiomed.2023.106559

[15] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li , et al., "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," Medical Image Analysis, Volume 63, 2020. DOI: 10.1016/j.media.2020.101693

[16] Wang, WC., Ahn, E., Feng, D. et al. A Review of Predictive and Contrastive Self-supervised Learning for Medical Images.

Mach. Intell. Res. 20, 483–513 ,2023. DOI: 10.1007/s11633-022-1406-4
[17] L. Nwosu, X. Li, L.Qian et al., “Semi-supervised Learning for COVID-19 Image Classification via ResNet”, arXiv: 2103.06140., 2021. DOI: 10.48550/arXiv.2103.06140
[18] E. Pachetti, S. Colantonio,"A systematic review of few-shot learning in medical imaging", Artificial Intelligence in Medicine, Volume 156, 2024. DOI:10.1016/j.artmed.2024.102949
[19] Parise, Alec, and Brian Mac Namee. "Exploring Optimal Configurations in Active Learning for Medical Imaging." International Conference on Innovative Techniques and Applications of Artificial Intelligence. Cham: Springer Nature Switzerland, 2023. DOI: 10.1007/978-3-031-47994-6_6
[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv:1409.1556, 2014. DOI: 10.48550/arXiv.1409.1556
[21] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. DOI: 10.1109/CVPR.2016.90
[22] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. DOI: 10.1109/CVPR.2017.243
[23] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. 2018. DOI: 10.1109/CVPR.2018.00474
[24] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In International conference on machine learning, pp. 6105-6114. PMLR, 2019. DOI: 10.1109/CVPR.2019.00336
[25] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020). DOI: 10.48550/arXiv.2010.11929
[26] Tu, Zhengzhong, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. "Maxvit: Multi-axis vision transformer." In European conference on computer vision, pp. 459-479. Cham: Springer Nature Switzerland, 2022. DOI: 10.1007/978-3-031-20053-3_27
[27] Khan, Asifullah, Zunaira Rauf, Abdul Rehman Khan, Saima Rathore, Saddam Hussain Khan, Najmus Saher Shah, Umair Farooq et al. "A recent survey of vision transformers for medical image segmentation." arXiv preprint arXiv:2312.00634 (2023). DOI: 10.48550/arXiv.2312.00634
[28] Shafi, Syed Mohammed, and Sathiya Kumar Chinnappan. "Hybrid transformer-CNN and LSTM model for lung disease segmentation and classification." PeerJ Computer Science 10 (2024): e2444. DOI: 10.7717/peerj-cs.2444
[29] Baur, Christoph, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images." In International MICCAI brainlesion workshop, pp. 161-169. Cham: Springer International Publishing, 2018. DOI: 10.1007/978-3-319-92282-3_18
[30] Frid-Adar, Maayan, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification." Neurocomputing 321, 321-331, 2018. DOI: 10.1016/j.neucom.2018.09.013
[31] Rajeashwari, S., and K. Arunesh. "Enhancing pneumonia diagnosis with ensemble-modified classifier and transfer learning in deep-CNN based classification of chest radiographs." Biomedical Signal Processing and Control 93 (2024): 106130. DOI: 10.1016/j.bspc.2024.106130
[32] Jin, Haoyang, Yufei Tang, Feiyang Liao, Qiang Du, Zhongyi Wu, Ming Li, and Jian Zheng. "Adaptive noise-aware denoising network: Effective denoising for CT images with varying noise intensity." Biomedical Signal Processing and Control 96 (2024): 106548. DOI: 10.1016/j.bspc.2024.106548
[33] Al-Shourbaji, Ibrahim, Pramod H. Kachare, Laith Abualigah, Mohammed E. Abdelhag, Bushra Elnaim, Ahmed M. Anter, and Amir H. Gandomi. "A deep batch normalized convolution approach for improving COVID-19 detection from chest X-ray images." Pathogens 12, no. 1 (2022): 17. DOI: 10.3390/pathogens12010017
[34] Hussein, Fairouz, Ala Mughaid, Shadi AlZu’bi, Subhieh M. El-Salhi, Belal Abuhaija, Laith Abualigah, and Amir H. Gandomi. "Hybrid clahe-cnn deep neural networks for classifying lung diseases from x-ray acquisitions." Electronics 11, no. 19 (2022): 3075. DOI: 10.3390/electronics11193075
[35] Kibria, Hafsa Binte, and Md Ali Hossain. "Lightweight parallel cnn to classify covid-19 associated pneumonia from chest x-ray." In 2023 20th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), pp. 1-6. IEEE, 2023. DOI: 10.1109/CCE56193.2023.10025543
[36] Luo, Yimin, Sophie Majoe, Jiang Kui, Haikun Qi, Kuberan Pushparajah, and Kawal Rhode. "Ultra-dense denoising network: application to cardiac catheter-based X-ray procedures." IEEE Transactions on Biomedical Engineering 68, no. 9 (2020): 2626-2636. DOI: 10.1109/TBME.2020.3041571
[37] Liu, Wufeng, Jiaxin Luo, Yan Yang, Wenlian Wang, Junkui Deng, and Liang Yu. "Automatic lung segmentation in chest X-ray images using improved U-Net." Scientific Reports 12, no. 1 (2022): 8649. DOI: 10.1038/s41598-022-12743-y
[38] Baltruschat, Ivo M., Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. "Comparison of deep learning approaches for multi-label chest X-ray classification." Scientific reports 9, no. 1 (2019): 6381. DOI: 10.1038/s41598-019-42857-3
[39] Jangam, Ebenezer, Chandra Sekhara Rao Annavarapu, and Aaron Antonio Dias Barreto. "A multi-class classification framework for disease screening and disease diagnosis of COVID-19 from chest X-ray images." Multimedia Tools and Applications 82, no. 10 (2023): 14367-14401. DOI: 10.1007/s11042-023-14005-4
[40] Huang, Jie, Zhao-Min Chen, Zhao-Min Chen, Xiaoqin Zhang, Xiaoqin Zhang, Yisu Ge, Yisu Ge et al. "Label decoupling and reconstruction: A two-stage training framework for long-tailed multi-label medical image recognition." In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 2861-2869. 2024. DOI: 10.1145/3611669.3643655
[41] Allaouzi, Imane, and Mohamed Ben Ahmed. "A novel approach for multi-label chest X-ray classification of common thorax diseases." IEEE Access 7 (2019): 64279-64288. DOI: 10.1109/ACCESS.2019.2916848
[42] Powers, David MW. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv preprint arXiv:2010.16061 (2020). DOI: 10.48550/arXiv.2010.16061
[43] Jaeger, Stefan, Sema Candemir, Sameer Antani, Yì-Xiáng J. Wáng, Pu-Xuan Lu, and George Thoma. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases." Quantitative imaging in medicine and surgery 4, no. 6 (2014): 475. DOI: 10.3978/j.issn.2223-4292.2014.11.20

[44] Mujahid, Muhammad, Furqan Rustam, Roberto Álvarez, Juan Luis Vidal Mazón, Isabel de la Torre Díez, and Imran Ashraf. "Pneumonia classification from X-ray images with inception-V3 and convolutional neural network." Diagnostics 12, no. 5 (2022): 1280. DOI: 10.3390/diagnostics12051280

[45] Ali, Mudasir, Mobeen Shahroz, Urooj Akram, Muhammad Faheem Mushtaq, Stefanía Carvajal Altamiranda, Silvia Aparicio Obregon, Isabel De La Torre Díez, and Imran Ashraf. "Pneumonia detection using chest radiographs with novel efficientnetv2l model." IEEE Access 12 (2024): 34691-34707. DOI: 10.1109/ACCESS.2024.3372683

[46] Alvi, Sohaib Bin Khalid, Muhammad Ziad Nayyer, Muhammad Hasan Jamal, Imran Raza, Isabel de la Torre Diez, Carmen Lili Rodriguez Velasco, Jose Manuel Brenosa, and Imran Ashraf. "A lightweight deep learning approach for COVID-19 detection using X-ray images with edge federation." Digital health 9 (2023): 20552076231203604. DOI: 10.1177/20552076231203604

[47] Bressem, Keno K., Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. "Comparing different deep learning architectures for classification of chest radiographs." Scientific reports 10, no. 1 (2020): 13590. DOI: 10.1038/s41598-020-70479-z

[48] Yang, Li, Shasha Liu, Jinyan Liu, Zhixin Zhang, Xiaochun Wan, Bo Huang, Youhai Chen, and Yi Zhang. "COVID-19: immunopathogenesis and Immunotherapeutics." Signal transduction and targeted therapy 5, no. 1 (2020): 128. DOI: 10.1038/s41392-020-00243-2

[49] Bhosale, Rajendra D., and D. M. Yadav. "Customized convolutional neural network for pulmonary multi-disease classification using chest x-ray images." Multimedia Tools and Applications 83, no. 6 (2024): 18537-18571. DOI: 10.1007/s11042-023-17147-0

[50] JZhang, Jianpeng, Yutong Xie, Guansong Pang, Zhibin Liao, Johan Verjans, Wenxing Li, Zongji Sun et al. "Viral pneumonia screening on chest X-rays using confidence-aware anomaly detection." IEEE transactions on medical imaging 40, no. 3 (2020): 879-890. DOI: 10.1109/TMI.2020.3040950

[51] Wang, Linda, Zhong Qiu Lin, and Alexander Wong. "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images." Scientific reports 10, no. 1 (2020): 19549. DOI: 10.1038/s41598-020-76550-z.