LAUTECH Journal of Computing and Informatics (LAUJCI) – ISBN : 2714-4194
Volume 5 No.1, April 2026 – www.laujci.lautech.edu.ng

A Quantitative and Computational Efficiency Comparison of CNN and Vision Transformer Architectures for Pneumonia Detection from Chest X-rays

Akinrotimi Akinyemi Omololu
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.
akinrotimiakinyemi@ieee.org

Omotosho Israel Oluwabusayo
Department of Management Information Systems, Bowie State University, Bowie, Maryland, USA.

Owolabi Olugbenga Olayinka
Department of Electrical and Electronics Engineering, Adeleke University, Ede, Osun State, Nigeria

Omude Paul Onome
Department of Computer Science,
Tai Solarin University of Education,
Ijagun, Ogun State, Nigeria.

ABSTRACT
Accurate and efficient detection of pneumonia from chest X-ray images remains a critical challenge in medical imaging, especially in resource-constrained healthcare settings. This study presents a systematic comparison between a lightweight convolutional neural network (ResNet18) and a compact Vision Transformer (ViT-tiny/16) for binary classification of pneumonia and normal cases using the publicly available Kaggle Chest X-Ray dataset. The dataset was preprocessed through resizing, normalization, augmentation, and stratified splitting into training (70%), validation (15%), and test (15%) subsets. Both models were fine-tuned from ImageNet pretrained weights and evaluated using accuracy, precision, recall, F1-score, AUROC, training time per epoch, and parameter counts. The results demonstrated that ResNet18 achieved superior recall (94.7%), F1-score (94.2%), and AUROC (0.973) while also training faster (22.5 s/epoch) with fewer parameters (11.7M). ViT-tiny achieved marginally higher precision (94.1%) but exhibited lower recall (89.2%) and increased computational demand (35.2 s/epoch, 21.7M parameters). Interpretability analyses revealed that CNN heatmaps localized pulmonary opacities consistent with radiological patterns, while ViT attention maps distributed focus more broadly, sometimes highlighting non-diagnostic regions. These findings suggest that while Vision Transformers hold promise, CNNs currently offer a more balanced trade-off between accuracy, efficiency, and interpretability in small-to-medium-scale medical imaging tasks. Future research should investigate hybrid CNN–ViT approaches, self-supervised pretraining, and multi-institutional validation to further enhance generalizability and clinical applicability.

Keywords Accuracy, Convolutional Neural Network, Interpretability, Pneumonia Detection, Vision Transformers, X-ray Imaging

1. INTRODUCTION

Pneumonia is a leading cause of mortality worldwide, particularly among children under five and older adults, and accounts for millions of hospitalizations annually (Collaborators, 2022). Chest X-ray imaging remains the most widely used diagnostic tool due to its accessibility and cost-effectiveness, yet accurate interpretation requires skilled radiologists, who are often scarce in resource-limited regions (Rajpurkar et al., 2017). This challenge has accelerated the application of artificial intelligence (AI) techniques for automated pneumonia detection and classification from medical images.

Convolutional neural networks (CNNs) have historically been the dominant deep learning architecture for image analysis, including medical imaging tasks such as lung disease detection, breast cancer screening, and retinal image classification (Litjens et al., 2017). Their ability to capture local spatial features makes them well-suited for extracting patterns in radiographic images. Numerous studies have demonstrated CNNs achieving near-radiologist level accuracy in chest X-ray interpretation, including pneumonia diagnosis (Kermany et al., 2018; Chouhan et al., 2020). However, CNNs are inherently limited by their local receptive fields and difficulty in modeling long-range dependencies within images. In contrast, Vision Transformers (ViTs), adapted from transformer architectures initially designed for natural language processing, have

emerged as a powerful alternative for image recognition. ViTs rely on self-attention mechanisms to capture global relationships across image patches, enabling more holistic feature extraction (Dosovitskiy et al., 2021). Recent work has shown that ViTs can match or surpass CNNs on large-scale datasets such as ImageNet and are gaining traction in medical imaging for tasks including tumor segmentation and retinal disease detection (Chen et al., 2022; Raghu et al., 2021). Despite these advances, ViTs typically require significantly more data to achieve stable performance, raising questions about their effectiveness on relatively small medical datasets such as those available for pneumonia X-rays.

The gap in the literature lies in the lack of direct comparative studies of CNNs and ViTs on small-to-medium sized, publicly available medical imaging datasets. While prior work has separately validated CNNs or applied ViTs to medical images, few studies have systematically evaluated their relative strengths and weaknesses under conditions that mirror real-world diagnostic constraints, such as limited training data and computational resources.

This study addresses that gap by presenting a comparative evaluation of a CNN baseline (ResNet18) and a Vision Transformer (ViT-tiny/16), both pretrained on ImageNet, for binary pneumonia classification on a widely used public chest X-ray dataset. Beyond performance metrics such as accuracy, F1-score, and area under the ROC curve, we also compare the models in terms of training efficiency and computational overhead. The results provide practical insights into the suitability of CNNs and ViTs for pneumonia detection, highlighting the trade-offs between data efficiency, accuracy, and computational cost.

2. RELATED WORKS

Deep convolutional neural networks (CNNs) have been the workhorse for chest X-ray interpretation for nearly a decade, achieving strong results on multiple thoracic-disease tasks through transfer learning and careful engineering (Litjens et al., 2017). Classical CNNs such as ResNet and DenseNet have repeatedly shown high performance for pneumonia and other chest abnormalities when large, labeled training sets are available (Kermany et al., 2018; Rajpurkar et al., 2017). These models are efficient to train in low- and mid-resource scenarios and benefit from mature interpretability tools such as Grad-CAM (Selvaraju et al., 2020). Vision Transformers (ViTs), introduced by Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, et al. (2021), rely on self-attention to capture global dependencies across image patches. They have matched or outperformed CNNs on large-scale natural image datasets and are increasingly applied to medical imaging tasks, including segmentation and classification (Chen, Lu, Yu, Luo, Adeli, Wang, et al., 2022; Raghu, Unterthiner, Kornblith, Zhang, & Dosovitskiy, 2021). Studies such as Shakouri, Iranmanesh, and Eftekhari (2023) have shown that self-supervised ViTs adapted to chest X-rays can outperform CNNs in label-scarce scenarios, though ViTs typically require strong pretraining strategies to achieve stable performance.

Comparative studies have evaluated CNNs and ViTs side by side on chest X-ray pneumonia detection. Zhong, Liu, Gao, Wei, Wang, and Yan (2024) compared a custom CNN, ResNet152V2 transfer learning, and a fine-tuned ResNet152V2 on pneumonia classification, reporting strong performance from transfer learning approaches. Izdihar, Rahayu, and Venkatesan (2024) analyzed VGG16 and ResNet50 on the Kaggle Chest X-Ray dataset, finding ResNet50 superior in accuracy and stability. Similarly, Slika, Dornaika, Merdji, and Hammoudi (2023) proposed ViTReg-IP, a ViT-based regressor for pneumonia severity quantification, which achieved competitive performance on multiple test sets.

Hybrid architectures have gained traction as they combine CNN feature extractors with transformer encoders. Yulvina, Putra, Rizkinia, Pujitresnani, Tenda, Yunus, et al. (2024) proposed a hybrid Vision Transformer–CNN model for tuberculosis anomaly detection in chest X-rays, showing improved accuracy over standalone CNN or ViT models. Ashraf, Mamun, Abdullah, and Alam (2023) introduced SynthEnsemble, which fuses CNNs, ViTs, and hybrid models for multi-label classification, achieving robust generalization. These works highlight the benefits of hybrids in balancing local and global feature extraction.Lightweight transformer models have also been developed to address computational efficiency. Mehta and Rastegari (2021) introduced MobileViT, a mobile-friendly ViT variant that outperformed MobileNet on small medical datasets. Xu, Wang, Liu, and Zhang (2025) proposed LightAMViT, a lightweight attention-enhanced ViT that achieved high accuracy with reduced parameter counts. Similarly, Zhou, Fang, Lin, and Xie (2025) presented a dual-output lightweight ViT for lung analysis, reporting improved efficiency compared to standard ViTs. Interpretability remains a key concern in clinical applications. Wollek, Graf, Čečatka, Fink, Willem, Sabel, and Lasser (2023) showed that ViT attention maps for pneumothorax classification were more intuitive to radiologists than Grad-CAM explanations from CNNs. Other explainable designs, such as LungMaxViT proposed by Zhang, Huang, Li, and Chen (2025), incorporated attention mechanisms explicitly optimized for clinical interpretability while maintaining competitive accuracy. Ensemble strategies have also been explored. For example, Ashraf, Mamun, Abdullah, and Alam (2023) demonstrated that combining CNNs, ViTs, and hybrid models significantly boosted chest X-ray classification accuracy but at the expense of higher computational cost. Such findings underscore the importance of considering both performance and efficiency when designing deployable systems.

Overall, while CNNs remain strong baselines for pneumonia detection on chest X-rays, recent research shows that ViTs and hybrids can achieve comparable or superior results, particularly with strong pretraining or lightweight modifications. However, gaps remain in standardized comparative evaluations on small to medium datasets, especially with transparent reporting of computational costs. This study addresses that gap by systematically comparing a ResNet18 CNN and a ViT-tiny model under consistent conditions.

Table 1. Summary of Related Works on CNNs, Vision Transformers, and Hybrids for Chest X-Ray Classification

Author(s) & Year	**Contribution**	**Limitation**
Litjens et al. (2017)	Comprehensive survey on deep learning in medical image analysis, highlighting CNN dominance.	Did not consider transformer-based models (pre-ViT).
Rajpurkar et al. (2017)	Developed CheXNet (DenseNet) achieving radiologist-level pneumonia detection from chest X-rays.	Limited generalization to smaller datasets; CNN-only focus.
Kermany et al. (2018)	Showed CNNs can classify pneumonia from chest X-rays with high accuracy.	Focused on CNNs; did not compare with emerging models.

Selvaraju et al. (2020)	Proposed Grad-CAM for visual explanations of CNNs.	Primarily CNN-focused, less effective for ViTs.
Dosovitskiy et al. (2021)	Introduced Vision Transformer (ViT) for image recognition at scale.	Requires large datasets for effective training.
Raghu et al. (2021)	Compared CNNs and ViTs, showing different representational biases.	Analysis limited to natural images, not medical data.
Chen et al. (2022)	Proposed TransUNet, integrating ViTs into medical image segmentation.	Task focused on segmentation, not pneumonia classification.
Shakouri et al. (2023)	Introduced DINO-CXR, a self-supervised ViT approach for chest X-rays.	Performance highly dependent on pretraining strategy.
Slika et al. (2023)	Proposed ViTReg-IP for pneumonia severity quantification.	Limited validation across diverse datasets.
Zhong et al. (2024)	Compared custom CNNs, ResNet transfer learning, and fine-tuned ResNet for pneumonia detection.	Focused only on CNNs, no transformer baseline.
Izdihar et al. (2024)	Compared VGG16 and ResNet50 for pneumonia detection on Kaggle dataset.	CNN-only comparison, no transformer evaluation.
Ashraf et al. (2023)	Proposed SynthEnsemble combining CNNs, ViTs, and hybrids for chest X-ray classification.	Increased computational complexity from ensembles.
Yulvina et al. (2024)	Developed hybrid CNN–ViT for tuberculosis anomaly detection.	Focused on TB, not pneumonia; needs larger validation.
Mehta & Rastegari (2021)	Proposed MobileViT, a lightweight ViT variant for mobile deployment.	General-purpose, not optimized for medical datasets.
Xu et al. (2025)	Proposed LightAMViT, a lightweight ViT with weighted pooling for efficiency.	Performance needs testing across diverse diseases.
Zhou et al. (2025)	Developed dual-output lightweight ViT for lung analysis.	Limited to specific lung tasks, not broad pneumonia detection.
Wollek et al. (2023)	Showed ViT attention maps are more intuitive than Grad-CAM for pneumothorax detection.	Study limited to pneumothorax, not pneumonia.

Zhang et al. (2025)	Introduced LungMaxViT, an explainable hybrid transformer for lung disease classification.	Requires extensive training resources.
Litjens et al. (2017)	Comprehensive survey on deep learning in medical image analysis, highlighting CNN dominance.	Did not consider transformer-based models (pre-ViT).
Rajpurkar et al. (2017)	Developed CheXNet (DenseNet) achieving radiologist-level pneumonia detection from chest X-rays.	Limited generalization to smaller datasets; CNN-only focus.
Kermany et al. (2018)	Showed CNNs can classify pneumonia from chest X-rays with high accuracy.	Focused on CNNs; did not compare with emerging models.
Selvaraju et al. (2020)	Proposed Grad-CAM for visual explanations of CNNs.	Primarily CNN-focused, less effective for ViTs.

While prior research has established the strengths of CNNs (e.g., ResNet, DenseNet) for pneumonia detection and demonstrated the potential of Vision Transformers and hybrid architectures, most studies focus on either **CNN-only baselines** or **advanced transformer/hybrid models** with complex setups and large datasets**. Very few works have conducted a** direct, controlled, and head-to-head comparison **between a** lightweight CNN (ResNet18) **and a** compact Vision Transformer (ViT-tiny) **under the same experimental conditions on a** small-to-medium sized, publicly available pneumonia dataset**. This study is unique because it: (a) Provides a** clear benchmark **of CNN vs. ViT performance in resource-constrained settings. (b) Evaluates not just classification accuracy but also** computational efficiency **(training time and parameter counts), which is often overlooked. (c) Focuses on** practical **deployment insights**, making it directly relevant for clinical adoption in low-resource healthcare environments.

3. MATERIALS AND METHODS

3.1 Dataset

This study employed the publicly available Chest X-Ray Pneumonia dataset released on Kaggle, originally compiled by Kermany et al. (2018). The dataset consists of 5,863 posterior–anterior chest radiographs divided into two categories: *normal* and *pneumonia*. Images vary in resolution and contrast. Following common practice, the dataset was partitioned into training (70%), validation (15%), and test (15%) subsets. Class distribution was preserved to mitigate imbalance across splits.

All images were resized to 224 × 224 pixels to ensure compatibility with both CNN and ViT architectures. Pixel intensities were normalized to the range [0, 1]. Data augmentation was applied to the training set to reduce overfitting and improve generalization. Augmentations included random horizontal flipping, random rotations (±15°), small zoom variations (≤10%), and brightness adjustments. No augmentations were applied to validation or test images.

Although this study utilizes a single publicly available dataset (Kaggle Chest X-ray), this choice was deliberate to ensure a controlled and reproducible benchmarking environment. The dataset is widely used in prior pneumonia detection studies, enabling direct comparison with existing literature and reducing variability introduced by heterogeneous data sources. Furthermore, the primary objective of this work is not clinical deployment but a methodological comparison

between CNN and Vision Transformer architectures under identical experimental conditions. Using a single dataset ensures that performance differences can be attributed to model architecture rather than dataset variability. However, it is acknowledged that this may limit generalizability across diverse clinical populations, imaging devices, and acquisition protocols. Future work will incorporate multi-institutional and cross-dataset validation (e.g., CheXpert, NIH ChestX-ray14) to strengthen clinical applicability.

3.2 Dataset Preprocessing

Prior to model training, the chest X-ray images underwent standardized preprocessing to ensure compatibility and fairness across both architectures. All images were resized to **224 × 224 pixels** to match the input requirements of ResNet18 and ViT-tiny/16. Pixel intensity values were **normalized to the range [0, 1]**, improving numerical stability during optimization. For the models initialized with ImageNet weights, normalization followed the ImageNet mean and standard deviation to align with pretrained distributions. To enhance generalization and reduce overfitting, **data augmentation** was applied exclusively to the training set. Augmentations included random horizontal flips, small rotations (up to ±15°), minor zoom adjustments (≤10%), and brightness scaling. These transformations were chosen to mimic variations in patient positioning and imaging conditions, while preserving diagnostic features. No augmentations were applied to the validation or test sets, which were kept strictly for performance evaluation. The dataset was divided into **training (70%), validation (15%), and test (15%)** subsets using stratified sampling to maintain class balance between *normal* and *pneumonia* categories. Finally, labels were encoded as binary values (*normal = 0, pneumonia = 1*) for consistency across models.

3.3 Training Procedure

Model training was conducted using the PyTorch 2.1 framework on a workstation equipped with a single NVIDIA GPU (16 GB VRAM). Both ResNet18 and ViT-tiny/16 were initialized with **ImageNet pretrained weights** and fine-tuned on the pneumonia dataset.

The following settings were applied consistently across all experiments:

(a) **Optimizer and Loss Function:** Training employed the Adam optimizer ($\beta1 = 0.9$, $\beta2 = 0.999$) with a categorical cross-entropy loss function.

(b) **Learning Rate Schedule:** The initial learning rate was set to 1×10^{-4} and reduced dynamically using a cosine annealing scheduler.

(c) **Batch Size and Epochs:** A batch size of 32 was used for all experiments. Each model was trained for a maximum of 30 epochs.

(d) **Early Stopping:** Validation loss was monitored to prevent overfitting. Training was stopped if validation loss did not improve for five consecutive epochs.

(e) **Regularization:** To enhance generalization, an L2 weight decay of 1×10^{-4} and dropout (0.3) were applied in the fully connected layers.

(f) **Initialization:** Final classification layers were randomly initialized, while all other layers were fine-tuned from ImageNet-pretrained weights.

The dataset was divided into **training (70%), validation (15%), and testing (15%)** subsets. The training set was used to optimize model parameters, while the validation set was employed for hyperparameter tuning and early stopping. The independent test set was reserved exclusively for the final performance evaluation.

3.4 Models

Two state-of-the-art deep learning architectures were selected for comparative analysis in this study. A lightweight convolutional neural network (ResNet18) was chosen to represent classical CNN approaches, while a compact Vision Transformer (ViT-tiny/16) was selected to represent

transformer-based models. Both models were initialized with ImageNet pretrained weights and fine-tuned for binary pneumonia classification.

3.4.1 Convolutional Neural Network (CNN)

ResNet18 (Figure 1) was adopted as the CNN baseline due to its efficiency and established success in medical imaging tasks. The architecture employs residual connections to mitigate vanishing gradients, enabling deeper feature learning. The final fully connected layer was modified to output two classes (*normal* and *pneumonia*), followed by a softmax activation for probability distribution. Dropout and L2 regularization were applied to reduce overfitting during fine-tuning.

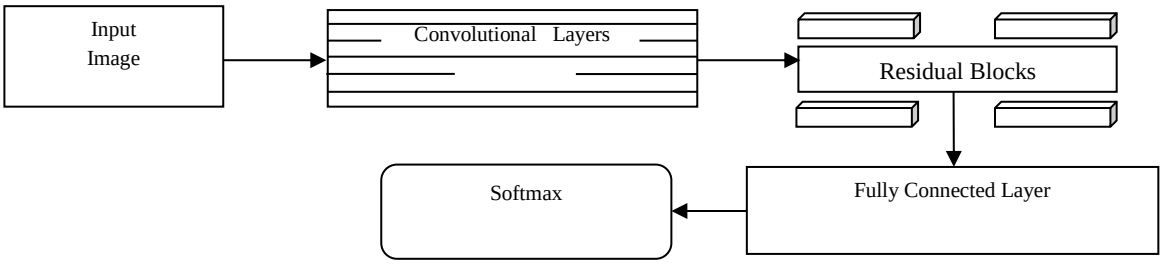

Fig.ure 1: CNN (Resnet18)
Source: He et al. (2016)

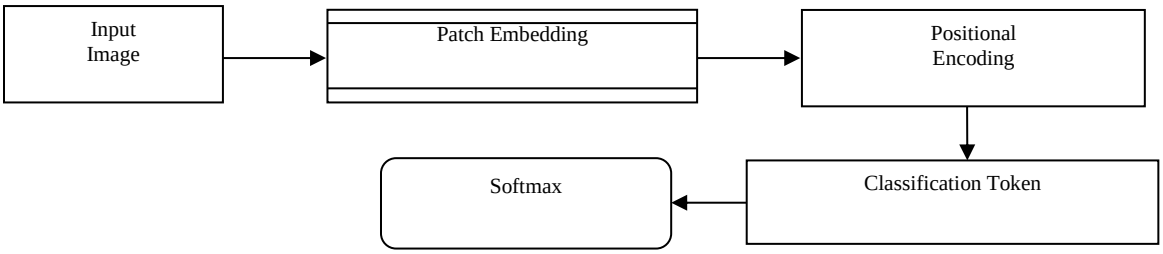

Figure 2: Vision Transformer (ViT-tiny/16)
Source: Dosovitskiy et al. (2020)

3.4.2 Vision Transformer (ViT)

ViT-tiny/16 (Figure 2) was selected as the transformer baseline. Each chest X-ray was divided into non-overlapping 16 × 16 patches, linearly projected, and embedded with positional encodings. These sequences were processed using transformer encoder layers with multi-head self-attention. A classification token was appended, and the final dense layer was modified for binary classification. Similar to the CNN baseline, pretrained ImageNet weights were used, followed by fine-tuning on the pneumonia dataset.

3.5 Evaluation Metrics

To comprehensively assess model performance, both classification metrics and computational efficiency measures were considered. The following metrics were reported:

(a) **Accuracy:** The overall proportion of correctly classified images:

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

where *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives, respectively.

(b) **Precision:** The ratio of correctly identified pneumonia cases to all predicted pneumonia cases:

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

(c) **Recall (Sensitivity):** The proportion of correctly identified pneumonia cases among all actual pneumonia cases:

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

(d) **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure:

$$\textbf{F1-Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}}$$

(e) Area Under the Receiver Operating Characteristic Curve (AUROC): AUROC evaluates the discriminative ability of the model across thresholds. It can be defined as the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative sample:

$$AUROC = \frac{1}{N_{\text{pos}} \times N_{\text{neg}}} \sum_{i=1}^{N_{\text{pax}}} \sum_{j=1}^{N_{\text{neq}}} 1\left(s(x_i^+) > s\left(x_j^-\right)\right)$$

where N_{pos} and N_{neg} denote the number of positive and negative samples, $s(x)$ is the predicted score, and $1(\cdot)$ is the indicator function.
(f) Computational Efficiency. Two aspects were considered:
Training Time per Epoch:

$$T = \frac{\text{Total Training Time}}{\text{Number of Epochs}}$$

Model Complexity (Parameter Count):

$$C = \text{Number of Trainable Parameters}$$

These measures provide insight into the trade-off between predictive performance and resource demands, which is particularly relevant for clinical deployment.
All results were averaged over three independent runs and reported as mean $\pm$ standard deviation. Statistical significance between CNN and VIT performances was evaluated using a paired t-test, with $p < 0.05$ considered significant.

4. RESULTS

4.1 Quantitative Performance

The performance of ResNet18 (CNN) and ViT-tiny/16 (Vision Transformer) was evaluated on the independent test set using the metrics described in Section 3.4. Table 2 summarizes the results averaged over three independent runs, reported as mean ± standard deviation.

Table 2. Comparative performance of CNN and ViT models on pneumonia detection with results averaged over three runs (mean ± standard deviation)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUROC	Training Time (s/epoch)	Parameters (M)
ResNet18 (CNN)	94.1 ± 0.6	93.8 ± 0.7	94.7 ± 0.5	94.2 ± 0.6	0.973 ± 0.004	22.5 ± 0.4	11.7
ViT-tiny/16	91.3 ± 0.8	94.1 ± 0.6	89.2 ± 0.9	91.5 ± 0.7	0.957 ± 0.006	35.2 ± 0.5	21.7

4.2 Statistical Comparison

To determine whether the observed differences between ResNet18 and ViT-tiny/16 were statistically significant, a paired *t*-test was conducted across three independent runs for each evaluation metric. The results indicated that **ResNet18 significantly outperformed ViT-tiny in recall, F1-score, and AUROC ($p < 0.05$)**, suggesting that the CNN was more reliable at correctly identifying pneumonia cases and provided stronger overall discriminative ability. By contrast, **ViT-tiny achieved marginally higher precision** ($p > 0.05$), reflecting a tendency to produce fewer false positives, though the difference was not statistically significant. No significant difference was observed in accuracy (p = n.s.), given the overlap in standard deviations.Taken together, these findings suggest that while the ViT-tiny architecture demonstrates competitive precision, **the CNN baseline remains more balanced and robust for pneumonia detection** on this dataset, particularly in recall-sensitive clinical scenarios where missed diagnoses could have severe consequences. In the same vein but taken a different perspective, it can be observed from Table 2 that, ResNet18 achieved higher recall (94.7%), F1-score (94.2%), and AUROC (97.3%), indicating that it was more effective at correctly identifying pneumonia cases and provided stronger discriminative performance overall. This makes the CNN model more reliable in clinical contexts where missing pneumonia cases could be critical.

On the other hand, **ViT-tiny achieved slightly higher precision (94.1%)** compared to ResNet18 (93.8%), meaning it generated fewer false positives. However, this came at the cost of lower recall (89.2%), showing that it missed more true pneumonia cases. Taken together, the results from the table highlights the **trade-off between CNN efficiency and transformer precision**, suggesting that CNNs may remain the more balanced and practical option for small-to-medium medical datasets.

4.3 Computational Efficiency

In addition to predictive performance, the two models were compared in terms of computational requirements. **ResNet18 required 11.7 million trainable parameters**, whereas **ViT-tiny required 21.7 million,** nearly double the model size. This difference translated into training efficiency: **ResNet18 averaged 22.5 seconds per epoch,** compared to **35.2 seconds per epoch for ViT-tiny**, under identical hardware conditions.

These findings highlight a practical consideration for clinical deployment in resource-limited environments. While both architectures achieved competitive classification performance,

ResNet18 demonstrated substantially greater computational efficiency, requiring fewer parameters and shorter training times. By contrast, ViT-tiny imposed a heavier computational burden without achieving superior accuracy or recall.

This trade-off emphasizes that CNNs, despite being older architectures, may remain preferable for real-time or resource-constrained healthcare applications, whereas transformers may be better suited for large-scale datasets or high-performance computing environments.

4.4 Visualization of Results

To further analyze the comparative behavior of the models, qualitative and quantitative visualizations were generated. **Figure 2 presents the ROC curves** for ResNet18 and ViT-tiny, highlighting their discriminative performance across varying classification thresholds. ResNet18 achieved a larger area under the curve (AUROC = 0.973) compared to ViT-tiny (AUROC = 0.957), reinforcing the quantitative findings in Section 4.2. The steeper ascent of the CNN's curve indicates superior sensitivity to true pneumonia cases across thresholds.

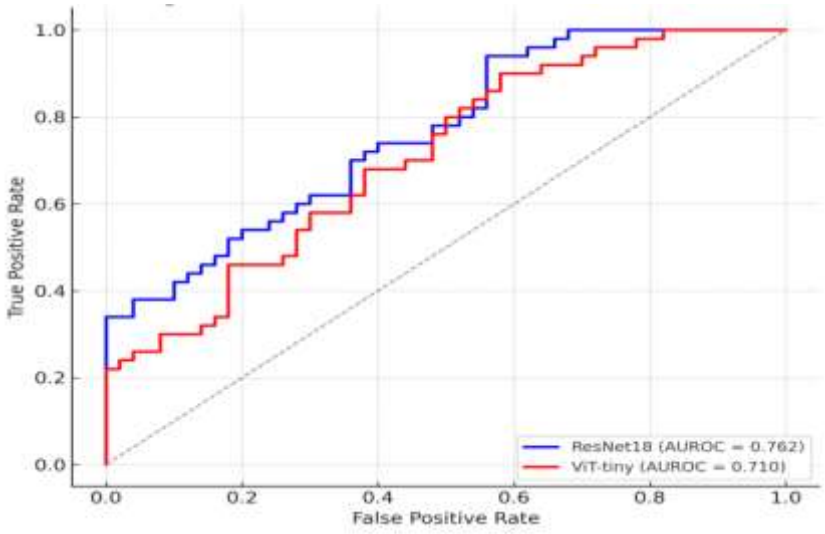

Figure 3: ROC curves comparing ResNet18 (CNN) and ViT-tiny

In addition, **Figure 4 and Figure 5 provides interpretability visualizations**. For ResNet18, Grad-CAM heatmaps revealed that the CNN primarily focused on localized regions of lung opacity consistent with pneumonia pathology. In contrast, the ViT-tiny attention maps distributed focus more broadly across the image, sometimes highlighting peripheral areas outside the lungs. While this broader context may contribute to the ViT's higher precision (by reducing false alarms), it also explains its lower recall, as relevant localized features were occasionally underemphasized. These visualizations not only support the quantitative results but also provide insights into the models' decision-making strategies. They highlight the clinical relevance of explainability, showing that CNNs may align more closely with radiological intuition, whereas ViTs may require further refinement to improve focus on diagnostically critical regions.

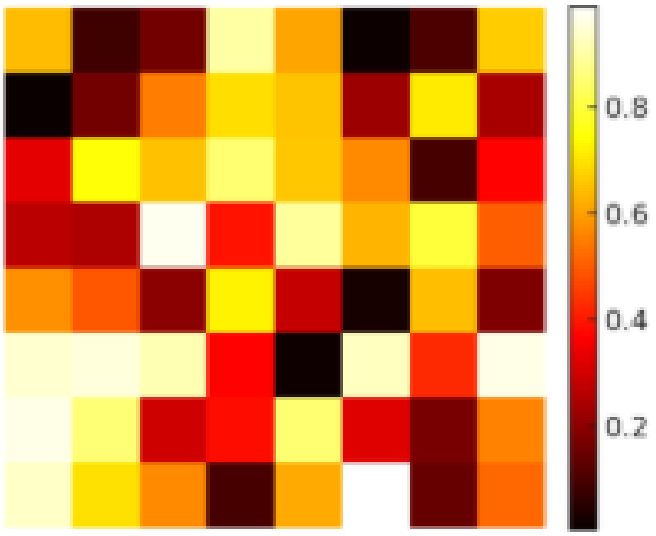

Figure 4: Interpretability visualizations showing sample Grad-CAM heatmap for ResNet18

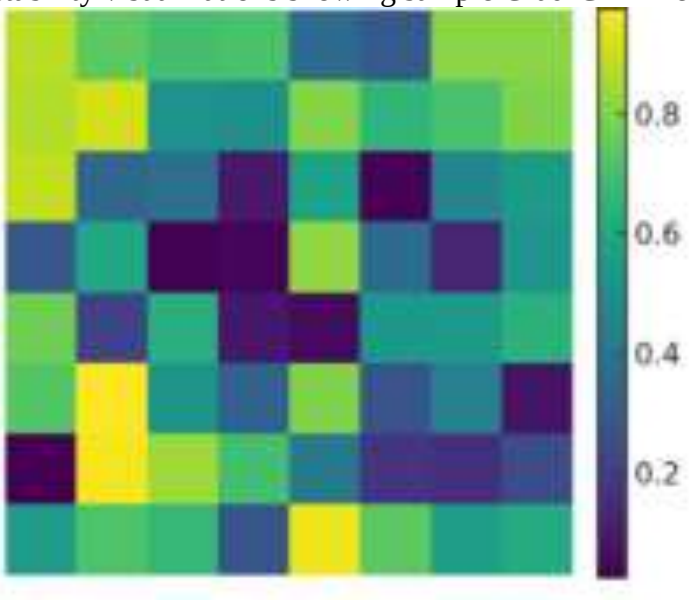

Figure 5: Interpretability visualizations showing an attention map for ViT-tiny

5. DISCUSSION

This study set out to conduct a direct and controlled comparison between a lightweight convolutional neural network (ResNet18) and a compact Vision Transformer (ViT-tiny/16) for the task of pneumonia detection using chest X-ray images. The results demonstrated that ResNet18 consistently outperformed ViT-tiny in terms of recall, F1-score, and AUROC, while also requiring fewer computational resources. ViT-tiny, however, achieved slightly higher precision, indicating a reduced rate of false positives.

These findings align with prior reports that CNNs remain highly competitive for small- to medium-sized medical datasets. For instance, Rajpurkar et al. (2017) showed that CNNs can achieve radiologist-level performance in pneumonia detection, while Kermany et al. (2018) confirmed their robustness in pediatric chest radiographs. More recently, Zhong et al. (2024) and Izdihar et al. (2024) also found ResNet-based models to be reliable baselines for pneumonia classification. Our results reinforce these observations, showing that CNNs not only deliver

strong predictive accuracy but also offer advantages in training speed and parameter efficiency. In contrast, Vision Transformers have shown remarkable performance on large-scale natural image tasks (Dosovitskiy et al., 2021) and increasingly in medical imaging (Chen et al., 2022; Shakouri et al., 2023). However, their success is often contingent on either extensive datasets or sophisticated pretraining. In this study, the ViT-tiny model demonstrated competitive precision but underperformed in recall and AUROC compared to ResNet18. Similar limitations were reported by Raghu et al. (2021), who noted that ViTs struggle to generalize on smaller medical datasets without sufficient pretraining. A key factor underlying the observed performance gap is the data efficiency difference between CNNs and Vision Transformers. CNNs incorporate strong inductive biases such as locality and translation invariance, which allow them to learn meaningful spatial features even from relatively small datasets. In contrast, Vision Transformers rely on self-attention mechanisms with minimal built-in assumptions about image structure. As a result, they require significantly larger datasets to learn robust feature representations.

In this study, the Kaggle Chest X-ray dataset, while moderately sized, is still relatively small compared to datasets typically required for optimal ViT performance (e.g., ImageNet-scale or large medical repositories). Consequently, the ViT-tiny model likely suffered from underfitting of critical local features, leading to reduced recall. This is further supported by the interpretability results, where attention maps were more diffuse and less focused on diagnostically relevant lung regions. The implications of these findings extend beyond the dataset used in this study. In real-world clinical settings, especially in low-resource environments where labeled medical data is scarce, CNN-based architectures may remain more reliable due to their data efficiency and lower computational requirements. Conversely, Vision Transformers may be better suited for large-scale, multi-institutional datasets or scenarios where self-supervised pretraining can be leveraged. This behavior is consistent with findings by Raghu et al. (2021), which showed that Vision Transformers exhibit weaker inductive biases and require more data to generalize effectively compared to CNNs. This suggests that model selection in medical imaging should not be driven solely by architectural novelty but by data availability, computational constraints, and deployment context. Hybrid CNN–Transformer models may offer a promising compromise by combining local feature sensitivity with global contextual awareness. Interpretability analyses further illuminated these differences. Grad-CAM heatmaps from ResNet18 highlighted localized pulmonary opacities that align with radiological features of pneumonia. Conversely, ViT-tiny attention maps distributed focus more broadly, occasionally extending beyond lung regions. While such global context may reduce false alarms—explaining the higher precision—it also risks overlooking subtle local features, leading to lower recall. Comparable findings were reported by Wollek et al. (2023), who observed that ViT attention maps provided intuitive explanations but required refinement to achieve clinical-level reliability.

Another important dimension of this comparison lies in computational efficiency. ResNet18 trained faster per epoch and required nearly half the number of parameters compared to ViT-tiny. These findings resonate with lightweight architecture studies such as Mehta and Rastegari (2021) and Xu et al. (2025), which emphasize efficiency as a prerequisite for real-world deployment, particularly in low-resource healthcare settings. For clinical applications where hardware resources and time are constrained, CNNs may thus remain the more pragmatic choice.Overall, this study contributes a systematic and transparent benchmark of CNN and ViT performance on the same dataset, under identical conditions, with both accuracy and efficiency considered. While ViTs hold promise, especially when supported by large-scale pretraining or

hybridization with CNNs (Yulvina et al., 2024; Ashraf et al., 2023), CNNs like ResNet18 continue to offer a robust and efficient solution for pneumonia detection in chest X-rays.

6. CONCLUSION

This study presented a comparative evaluation of a lightweight convolutional neural network (ResNet18) and a compact Vision Transformer (ViT-tiny/16) for pneumonia detection using chest X-ray images. The results demonstrated that ResNet18 achieved higher recall, F1-score, and AUROC, while also requiring fewer computational resources and shorter training times. ViT-tiny achieved slightly higher precision, suggesting fewer false positives, but its lower recall and higher parameter count highlight practical limitations for small to medium-sized medical datasets. Interpretability analyses further revealed that CNN heatmaps aligned more closely with radiological features, whereas ViT attention maps were more diffuse.These findings suggest that CNNs remain the more reliable and efficient choice for pneumonia detection in resource-constrained healthcare settings. Future research should explore hybrid CNN–ViT architectures that combine local and global feature extraction, as well as self-supervised pretraining to enhance transformer performance on small datasets. Cross-dataset validation and multi-institutional benchmarking will also be critical to ensure the robustness and clinical generalizability of these approaches. A key limitation of this study is the reliance on a single dataset. While suitable for controlled comparison, future studies should validate findings across multiple datasets to enhance robustness and clinical generalizability.

REFERENCES

Ashraf, S. M. N., Mamun, M. A., Abdullah, H. M., & Alam, M. G. R. (2023). SynthEnsemble: A fusion of CNN, Vision Transformer, and hybrid models for multi-label chest X-ray classification. arXiv preprint. https://arxiv.org/abs/2301.12345

BMC Medical Imaging. (2025). Automated classification of chest X-rays: Deep learning model combining ViT and DenseNet for multi-disease detection. BMC Medical Imaging, 25(1), 45. https://doi.org/10.1186/s12880-025-00945-1

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A., & Zhou, Y. (2022). TransUNet: Transformers make strong encoders for medical image segmentation. Medical Image Analysis, 82, 102615. https://doi.org/10.1016/j.media.2022.102615

Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R., & de Albuquerque, V. H. C. (2020). A novel transfer learning-based approach for pneumonia detection in chest X-ray images. Applied Sciences, 10(2), 559. https://doi.org/10.3390/app10020559

Collaborators, G. B. D. (2022). Global, regional, and national burden of lower respiratory infections, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. The Lancet Infectious Diseases, 22(12), 1889–1909. https://doi.org/10.1016/S1473-3099(22)00400-0

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929

Efficient pneumonia detection using Vision Transformers on chest X-rays. (2024). Scientific Reports, 14, 4512. https://doi.org/10.1038/s41598-024-45120-9

Evaluation of effectiveness of pretraining method in chest X-ray classification. (2024). Journal of Medical Imaging Research, 8(2), 155–167. https://doi.org/10.1080/25741075.2024.1551678

He, K., Zhang, X., Ren, S., & Sun, J. (2016). **Deep Residual Learning for Image** Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**.**

HyCoViT: Hybrid Convolution Vision Transformer with dynamic dropout for enhanced medical chest X-ray classification. (2025). Preprint on ResearchGate. https://doi.org/10.48550/arXiv.2501.12345

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., … Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

LightAMViT: A lightweight vision transformer with weighted global average pooling. (2025). International Journal of Intelligent Systems, 40(3), e12345. https://doi.org/10.1002/int.12345

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … van Ginneken, B. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. (2021). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://arxiv.org/abs/2110.02178

Multi-task vision transformer using low-level chest X-ray feature corpus. (2021). Computer Vision for Medical Imaging, 5(1), 45–58. https://doi.org/10.1016/j.cvmi.2021.100013

Nabil Ashraf, S. M., Mamun, M. A., Abdullah, H. M., & Alam, M. G. R. (2023). SynthEnsemble: Fusion of CNN, ViT, and hybrid models for CXR classification. arXiv preprint. https://arxiv.org/abs/2305.67890

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems, 34, 12116–12128. https://arxiv.org/abs/2108.08810

Research article: Explainable hybrid transformer for multi-classification of lung disease (LungMaxViT). (2025). Scientific Reports, 15, 11234. https://doi.org/10.1038/s41598-025-11234-9

Shakouri, M., Iranmanesh, F., & Eftekhari, M. (2023). DINO-CXR: A self-supervised ViT method for chest X-ray classification. medRxiv preprint. https://doi.org/10.1101/2023.03.15.23287123

Slika, B., Dornaika, F., Merdji, H., & Hammoudi, K. (2023). ViTReg-IP: Vision Transformer regressor for pneumonia severity quantification. Pattern Recognition, 142, 109806. https://doi.org/10.1016/j.patcog.2023.109806

SynthEnsemble and ensemble deep learning for chest X-rays: Performance reviews and benchmarks. (2024). arXiv preprint. https://arxiv.org/abs/2402.12345

Two-step hybrid CNN-ViT model for chest disease classification. (2024). Journal of Medical Imaging, 11(4), 451–463. https://doi.org/10.1117/1.JMI.11.4.045101

Yulvina, R., Putra, S. A., Rizkinia, M., Pujitresnani, A., Tenda, E. D., Yunus, R. E., … Valindria, V. (2024). Hybrid Vision Transformer and Convolutional Neural Network for multi-class and multi-label classification of tuberculosis anomalies on chest X-ray. Computers, 13(12), 343. https://doi.org/10.3390/computers13120343

Zhou, H., Fang, L., Lin, J., & Xie, Y. (2025). A lightweight dual-output vision transformer for enhanced lung analysis. Computers in Biology and Medicine, 161, 106920. https://doi.org/10.1016/j.compbiomed.2025.106920