



Benchmarking Transformer Architectures for Chest X-ray Classification

Joshua Pinem¹, Widi Astuti², Adiwijaya³

¹Informatics Study Program, School of Computing, Telkom University, Bandung, Indonesia

¹jopisan@student.telkomuniversity.ac.id, ²widiwdu@telkomuniversity.ac.id, ³adiwijaya@telkomuniversity.ac.id

Abstract

Lung diseases remain a major global health concern, necessitating accurate and timely diagnoses. Chest X-ray (CXR) imaging is widely used but challenging to interpret due to overlapping radiographic features and subjective variability among radiologists. Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have shown promise but are limited in capturing global spatial dependencies. Vision Transformers (ViTs) overcome this limitation through self-attention, making them increasingly attractive for medical image analysis. This study systematically evaluates 13 Transformer-based architectures across three CXR datasets with distinct tasks: Pneumonia (3-class: Normal, Bacterial, Viral), COVID-QU-Ex (3-class: Normal, Non-COVID Pneumonia, COVID-19), and Tuberculosis (2-class: Normal, Tuberculosis). All models were trained using a unified setup with consistent preprocessing, augmentation, and evaluation protocols. To improve robustness, a soft voting ensemble of the top five models was also implemented. Results demonstrate that Transformer-based models provide highly competitive performance. On the Pneumonia dataset, the ensemble achieved an accuracy of 0.8743 and F1-score of 0.8615, surpassing several single models such as DeiT-Base (F1 = 0.8725). On COVID-QU-Ex, the ensemble soft voting obtained 0.9593 accuracy and 0.9582 F1-score, effectively balancing precision and recall. On Tuberculosis, ViT-B/16 and MobileViT-S achieved perfect performance (F1 = 1.0), likely influenced by dataset imbalance. These findings highlight the clinical potential of Transformer-based models, particularly when combined through ensembles, for robust and accurate CXR classification.

Keywords: Chest X-ray; classification; deep learning; lung diseases; transformer; vision transformer; ViT

How to Cite: J. Pinem, W. Astuti, and Adiwijaya, "Benchmarking Transformer Architectures for Chest X-ray Classification", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 1, pp. 62 - 70, Feb. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i1.7132>

Received: August 29, 2025

Accepted: February 6, 2026

Available Online: February 16, 2026

This is an open-access article under the [CC BY 4.0 License](#)
Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Lung diseases, particularly lower respiratory tract infections, remain a major global health challenge with high morbidity and mortality. In 2019 alone, an estimated 488.9 million new cases were reported worldwide, causing about 2.4 million deaths [1]. Pneumonia, the most common type, is especially severe in young children and accounted for 14% of under-five deaths globally, or about 740,180 cases in 2019 [2]. These figures highlight the need for accurate, timely diagnosis to improve outcomes and reduce preventable deaths.

Chest X-ray imaging is a commonly used method in the diagnosis of pneumonia because it can reveal important signs, such as increased opacity and consolidation in the lungs [3]. However, interpreting radiographic images can be challenging, as the manifestations of pneumonia

can resemble or overlap with other lung diseases [3]. In addition, variations in Chest X-ray imaging results can lead to subjective assessments among radiologists, resulting in differences in diagnosis. This process also takes time, with a delay between scanning and diagnosis that can reach 1-2 days [4]. Such delays can have a significant impact on patient management and safety. These challenges highlight the need for automated tools that can support radiologists in achieving faster and more reliable diagnosis.

In this context, Artificial Intelligence (AI) and Machine Learning (ML) provide promising solutions for developing objective and accurate pneumonia detection systems [5]. Rapid progress in deep learning has greatly enhanced medical image analysis capabilities [6]. One of the most widely used approaches is the Convolutional Neural Network (CNN) architecture [7], which has proven effective in identifying pathological

lung conditions, as demonstrated by the CheXNet model [8]. However, CNNs depend on local convolution filters, limiting their ability to capture global spatial relationships and making them less optimal for understanding the overall image context [9].

Transformers are increasingly applied in medical image analysis to address CNN limitations. Their self-attention mechanism enables global context modeling and better handling of complex chest structures [10], and they have shown superior performance in pneumonia detection from X-rays [11]. However, most studies focus on single Transformer variants, and comprehensive comparisons across architectures of differing complexity, efficiency, and representational capacity remain limited [12].

Previous studies on chest X-ray (CXR) classification have widely applied deep learning methods, with CNN-based models showing strong performance in terms of accuracy and efficiency. For instance Sanida and Dasygenis [13] achieved 98.56% accuracy with a lightweight CNN, Reshan et al. [14] found MobileNet most effective for pneumonia detection (94.23% on 5,856 images; 93.75% on 112,120 images). Guddati [15] implemented a lightweight DCNN on FPGA with 96.39% accuracy. Despite strong performance, CNNs struggle to capture global spatial relationships, motivating exploration of alternatives like Vision Transformers, which leverage self-attention for better global context understanding.

To further address the limitations of CNNs, Vision Transformer (ViT) has been increasingly explored in CXR analysis. Elrefaei et al. [16] reported ViT-132 with 92.56% accuracy compared to DenseNet-169 with 92.31%, while Pal et al. [17] showed ViT achieving 95.87% accuracy and 0.93 AUC, higher than VGG-19 with 94.71% and 0.94. Liang et al. [18] reduced MAE to 3.64 using self-supervised ViT compared to 5.70 with ResNet-50, and Singh et al. [19] reached 97.61% accuracy and 0.96 AUC for binary pneumonia classification. Chen et al. [4] found fine-tuned ViT most effective, with 95.79% accuracy for four-class, 99.57% for three-class, and consistently strong results in binary tasks.

Hybrid architecture has shown strong results in CXR classification. Mustapha et al. [9] combined CNN and Swin Transformer, achieving 98.72% accuracy and F1-score, outperforming pure CNNs. Wang et al. [20] introduced PneuNet (ResNet18 + ViT), reaching 95.16% accuracy for three-class, 99.32% for binary, and 90.03% for four-class tasks. Ko et al. [21] evaluated Transformer variants (ViT-B/16, FastViT, CrossViT) with different optimizers and datasets, finding ViT-B/16 with RAdam best on balanced data (95.87%), while FastViT with Nadam excelled on unbalanced data.

Padmavathi and Ganesan [22] optimized a ViT for CXR disease severity classification using metaheuristic

algorithms, where Grey Wolf Optimizer tuned hyperparameters and Particle Swarm Optimizer selected features, yielding higher accuracy but at high computational cost. Pal et al. [23] further showed that an ensemble of ResNet50, VGG-19, and ViT outperformed a single CNN for COVID-19 CXR classification (95.14% multi-class, 92.46% binary). Similarly, Oltu et al. [24] integrated DenseNet201 with ViT, achieving 97.87% accuracy and 96.23% F1-score, reinforcing the effectiveness of hybrid CNN-ViT approaches in medical image classification.

Selvano et al. [25] showed that DINO ViT-S/16 outperformed CNNs in multi-class lung disease classification across different dataset settings with F1-scores above 93%. Kadry et al. [26] found that patch size selection significantly impacts ViT accuracy for COVID-19 detection. Yang et al. [27] introduced CovidViT (ViT-B/16), which outperformed all baselines with 98% accuracy and 0.998 AUC, highlighting ViT's effectiveness in capturing global features for CXR disease detection. Overall, Vision Transformers and CNN-ViT hybrids achieve accuracy and F1-score above 90% in CXR tasks, highlighting their potential over conventional CNNs.

Previous studies have compared Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for medical image classification, particularly pneumonia detection from chest X-rays. Overall, ViTs tend to outperform CNNs in accuracy and global context modeling. Several enhancements have been explored, including CNN-ViT hybrid architectures, ensemble learning, and hyperparameter optimization. However, systematic comparisons across Transformer architectures with varying complexities remain limited, obscuring the trade-offs between accuracy, efficiency, and model complexity.

To address this gap, this study aims to systematically benchmark 13 Transformer architectures for pneumonia detection on chest X-ray (CXR) images using standardized datasets, preprocessing procedures, and evaluation metrics under a unified experimental setting. The outcome of this research is expected to provide practical guidance for researchers and healthcare practitioners in selecting appropriate Transformer models for medical image classification tasks.

The main contributions of this study include: (1) providing comprehensive benchmarking across diverse Transformer architectures, (2) analyzing the strengths and weaknesses of each model based on key evaluation metrics, (3) identifying the most suitable Vision Transformer (ViT) architectures for clinical application by considering accuracy, efficiency, and model complexity, and (4) evaluating the effectiveness of a soft voting ensemble constructed from the top five performing models.

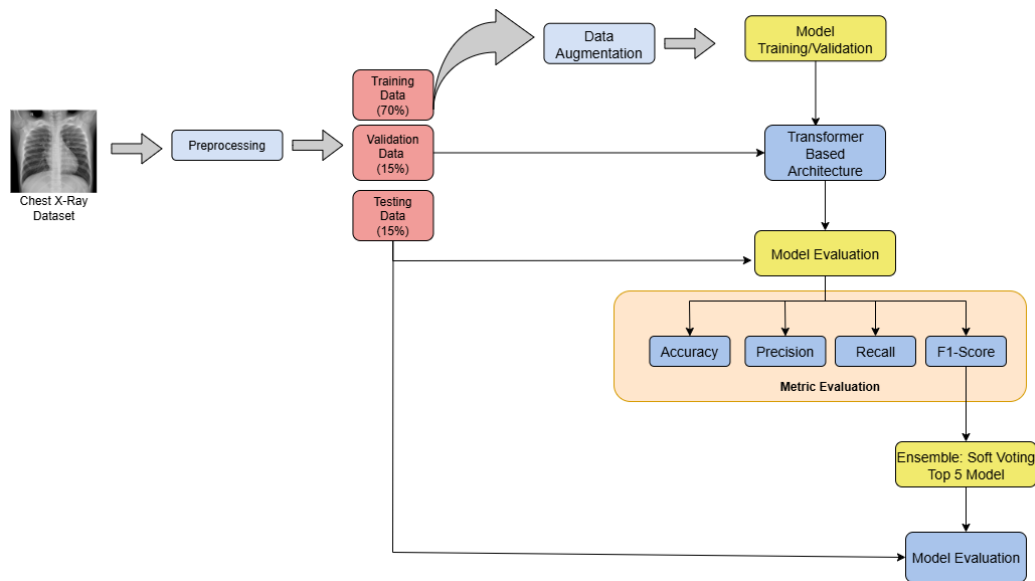


Figure 1. Overview of the benchmarking Transformer Architectures for Chest X-ray Classification

2. Methods

2.1 Dataset Description and Preprocessing

The datasets utilized in this study include three widely used chest X-ray collections. The first is the Chest X-Ray Images (Pneumonia) dataset by Kermany et al. [28], which consists of two categories: normal and pneumonia. The second dataset is COVID-QU-Ex [29], providing chest X-ray images classified into three categories: COVID-19, non-COVID pneumonia, and normal. The third dataset is the Tuberculosis Chest X-Ray Database [30], which comprises two categories: tuberculosis (TBC) and normal. To ensure a comprehensive evaluation of model performance, each dataset was divided into training, validation, and testing subsets. To ensure a comprehensive evaluation of model performance, each dataset was divided into training, validation, and testing subsets. Figure 2 shows representative examples from each class, highlighting differences in chest X-ray appearances between healthy lungs and those with disease. In this study, the Pneumonia dataset was treated as a three-class task (Normal, Viral Pneumonia, Bacterial Pneumonia), the COVID-QU-Ex dataset as a three-class task (COVID-19, Pneumonia, Normal), and the Tuberculosis dataset as a binary task (Tuberculosis vs. Normal). The distribution of images across each class and subset is shown in Table 2.

Prior to model training, all images underwent a standardized preprocessing pipeline. Each image was resized to a fixed resolution of 224×224 pixels, converted into tensor format, and normalized using ImageNet mean and standard deviation values to ensure compatibility with pre-trained Transformer architectures. Data augmentation was applied exclusively to the training set to improve model generalization and mitigate overfitting. The augmentation techniques included random horizontal flipping ($p = 0.3$), random rotation ($\pm 10^\circ$), and color

jitter adjustments in brightness and contrast (0.2). Validation and testing images were not augmented and only underwent resizing and normalization to preserve their original distribution for unbiased evaluation. In this study, each dataset was treated independently to evaluate model performance across different diagnostic scenarios. Specifically, models were trained and tested separately on the Chest X-Ray Pneumonia dataset, the COVID-QU-Ex dataset, and the Tuberculosis Chest X-ray dataset. This design allows for a fair assessment of generalization capability across distinct clinical tasks without merging the datasets into a single multi-class problem.

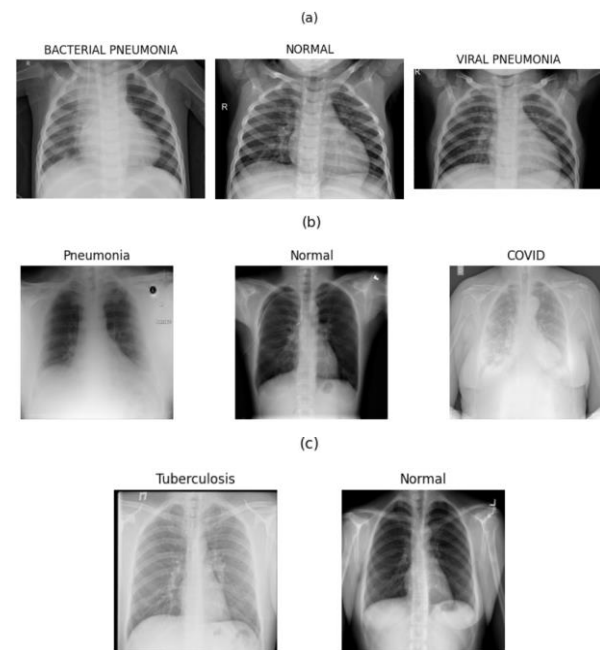


Figure 2. Representative chest X-ray images from the three datasets used in this study: (a) Pneumonia Chest X-ray (Normal, Viral Pneumonia, Bacterial Pneumonia), (b) COVID-QU-Ex (COVID-19, Pneumonia, Normal), and (c) TB Chest X-ray (Tuberculosis, Normal)

Table 2. Distribution of Images Across Datasets

Dataset	Class	Number of Images
Pneumonia	Bacterial	1,583
	Pneumonia	
Chest X-ray	Normal	2,780
	Viral Pneumonia	1,493
COVID-QU-Ex	COVID	11,956
	NORMAL	10,701
	Pneumonia	11,263
Tuberculosis	Tuberculosis	700
Chest X-ray	Normal	3,500

2.2 Model Architecture s

This study benchmarks 13 Transformer-based architectures, including DeiT-Tiny, DeiT-Small, DeiT-Base, ViT-S/16, ViT-B/16, ViT-L/16, Swin-T, Swin-S, PiT-Ti, ConViT, MobileViT-S, MobileViT-XS, and CrossViT. All images were resized to a fixed resolution, converted into tensor format, and normalized using ImageNet statistics. Data augmentation was applied only to the training set, while validation and test sets were strictly resized and normalized to ensure unbiased evaluation. To improve robustness, and to emphasize the most clinically relevant architectures, predictions from the five models with the highest F1-scores were further combined using a soft voting ensemble.

The overall workflow is summarized in Figure 1, showing the sequential steps from dataset preparation, preprocessing and augmentation, model training, evaluation, to ensemble integration. The chosen architecture represents a range of design complexities, from lightweight models such as MobileViT and PiT optimized for efficiency, to large-scale variants like ViT-L/16 that emphasize representational capacity. In addition, hybrid variants such as Swin, ConViT, and CrossViT were included to capture hierarchical and locality-aware features. Transformers were selected as they are inherently capable of modeling long-range global dependencies in medical images. This diversity ensures a fair and comprehensive benchmarking. Table 1 presents a comparison of the number of parameters across Transformer-based architectures. The results show that MobileViT-XS contains the lowest number of parameters, reflecting its relatively compact structure, whereas ViT-L/16 has the highest parameter count, indicating a significantly larger model size.

Table 1. Model architectures and number of parameters

Model	Parameters
DeiT-Tiny	5,717,416
DeiT-Small	22,050,664
DeiT-Base	86,567,656
ViT-S/16	22,050,664
ViT-B/16	86,567,656
ViT-L/16	304,326,632
Swin-T	28,288,354
Swin-S	49,606,258
PiT-Ti	4,847,272
ConViT	27,777,322
MobileViT-S	5,578,632
MobileViT-XS	2,317,848
CrossViT-Ti	7,553,296

This difference illustrates the variation in architectural complexity among Transformer models, where smaller models emphasize efficiency, while larger models are designed to maximize representational capacity.

2.3 Model Training and Evaluation

All Vision Transformer architectures were trained under a unified setup to ensure fair comparison. Training was conducted for 20 epochs with a batch size of 32, using AdamW (learning rate 0.0001, weight decay 0.01), cosine annealing as the scheduler, and early stopping with a patience of five epochs. The complete configuration is summarized in Table 3.

Table 3. Training configuration of all Vision Transformer architectures

Parameter	Setting
Input image size	224 × 224
Batch size	32
Number of epochs	20
Optimizer	AdamW
Learning rate	0.0001
Weight decay	0.01
Learning rate scheduler	Cosine annealing
Early stopping	Patience 5 epochs
Data augmentation	Resize, RandomHorizontalFlip (p=0.3), RandomRotation (10°), ColorJitter (brightness=0.2, contrast=0.2)
Train/Val/Test	70%/15%/15%
Loss function	Cross-Entropy-Loss
Evaluation metrics	Accuracy, precision, recall, F1-score

The learning rate of 0.0001 was chosen following common practice for fine-tuning Transformer-based architectures on chest x-ray dataset [31], [32], and preliminary experiments indicated that it yielded stable convergence.

The dataset was split into training, validation, and testing sets using stratified sampling to maintain class balance. Training images underwent uniform augmentation, including resizing, random horizontal flips, rotations, and color jitter.

For all datasets, stratified splitting was applied to maintain class distribution across training, validation, and test sets. No specific imbalance handling techniques were applied during training. This choice was made to evaluate the architectures to highlight potential sensitivity to imbalance.

Model evaluation was conducted using accuracy, class-wise recall dan precision, and F1-score. Furthermore, soft voting ensembles were created for all trained models as well as for the top five models based on F1-score to investigate the potential improvements in predictive performance through model aggregation. The formulas for these evaluation metrics are presented as Equations 1-4.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

Equations 1-4 present the formulas for Accuracy, Precision, Recall, and F1-Score. In these equations, TP refers to True Positives, TN to True Negatives, FP to False Positives, and FN to False Negatives. Precision measures the proportion of correctly predicted positive cases, Recall emphasizes sensitivity in identifying actual positive cases, and F1-Score balances both Precision and Recall.

3. Results and Discussions

3.1 Performance Comparison of Models

Model performance evaluation was conducted under identical experimental conditions to maintain fair comparison. A total of 13 transformer-based models were compared, and the 5 models with the highest F1-

scores were selected. These five best models were then combined using the soft voting ensemble method to improve overall prediction accuracy. The evaluation metrics included accuracy, precision, recall, and F1-score, chosen to provide a comprehensive assessment of model performance, both in terms of classification accuracy and the balance between positive and negative errors. Testing was conducted independently on three datasets: Pneumonia Chest X-ray (Bacterial Pneumonia, Viral Pneumonia, Normal), COVID-QU-Ex (COVID, Normal, Pneumonia), and Tuberculosis Chest X-ray (Tuberculosis, Normal). The evaluation results for each dataset are shown in Tables 4-6 and Figure 3.

The evaluation results show that on the Pneumonia Chest X-ray, COVID QU Ex, and Tuberculosis Chest X-ray datasets, the transformer based models achieved highly competitive classification performance. On the Pneumonia dataset, the soft voting ensemble reached 0.8743 accuracy and 0.8615 F1 score, exceeding several single models. The DeiT Base model recorded the highest F1 score of 0.8725.

Table 4. Performance Comparison on the Pneumonia Chest X-ray Dataset

Model	Metric Evaluation						Accuracy	F1-Score
	Bacterial	Precision Normal	Viral	Bacterial	Recall Normal	Viral		
DeiT-Tiny	0.8285	0.9826	0.7810	0.91578	0.9537	0.6534	0.8621	0.8498
DeiT-Small	0.8502	0.9625	0.7473	0.8815	0.9705	0.6881	0.8597	0.8495
DeiT-Base	0.8625	0.9709	0.8044	0.9078	0.9831	0.7128	0.8817	0.8725
ViT-S/16	0.8545	0.9870	0.7461	0.8815	0.9579	0.7277	0.8658	0.8589
ViT-B/16	0.8135	0.9746	0.7176	0.8842	0.9705	0.6039	0.84024	0.8253
ViT-L/16	0.8205	0.9744	0.7884	0.9263	0.9621	0.6089	0.8585	0.8418
Swin-T	0.8337	0.9504	0.7371	0.8842	0.9663	0.6386	0.8475	0.8336
Swin-S	0.855	0.9957	0.7860	0.9	0.9747	0.7277	0.8792	0.8726
PiT-Ti	0.8033	0.9465	0.7090	0.871	0.9663	0.5792	0.8268	0.8099
ConViT	0.86	0.9543	0.7688	0.8894	0.9663	0.7079	0.867	0.8573
MobileViT-S	0.8947	0.9663	0.6881	0.8436	0.9787	0.7637	0.8646	0.8549
MobileViT-XS	0.847	0.957	0.702	0.8429	0.9702	0.6995	0.8439	0.8368
CrossViT-Ti	0.8636	0.9827	0.7604	0.9	0.9579	0.7227	0.8731	0.8642
Soft Voting Ensemble Top 5 Model	0.847	0.9629	0.8121	0.9184	0.9831	0.6633	0.8743	0.8615

Table 5. Performance Comparison on the COVID-QU-Ex Dataset

Model	Metric Evaluation						Accuracy	F1-Score
	COVID	Precision Normal	Pneumonia	COVID	Recall Normal	Pneumonia		
DeiT-Tiny	0.9915	0.9106	0.9392	0.9866	0.9395	0.9159	0.9483	0.9471
DeiT-Small	0.9932	0.9034	0.9368	0.9821	0.9389	0.9129	0.9455	0.9444
DeiT-Base	0.991	0.9146	0.9359	0.991	0.9345	0.9165	0.9485	0.9472
ViT-S/16	0.991	0.9192	0.9276	0.9899	0.9214	0.9265	0.9473	0.9459
ViT-B/16	0.9795	0.9042	0.9365	0.9877	0.9239	0.9088	0.9414	0.94
ViT-L/16	0.9871	0.8999	0.9548	0.9882	0.9526	0.9005	0.9479	0.9467
Swin-T	0.9938	0.9225	0.9475	0.9916	0.9426	0.9301	0.9557	0.9546
Swin-S	0.979	0.9281	0.9287	0.9921	0.909	0.9336	0.9465	0.945
PiT-Ti	0.9732	0.9043	0.9417	0.9938	0.9246	0.9005	0.941	0.9395
ConViT	0.9883	0.9134	0.9367	0.9921	0.9271	0.9194	0.9475	0.9461
MobileViT-S	0.9932	0.9096	0.9387	0.9816	0.9345	0.9259	0.9483	0.9472
MobileViT-XS	0.991	0.9152	0.941	0.9866	0.9352	0.9259	0.9502	0.9491
CrossViT-Ti	0.9866	0.9033	0.9498	0.98773	0.9439	0.9082	0.9475	0.9463
Soft Voting Ensemble Top 5 Model	0.9944	0.9232	0.9572	0.9944	0.952	0.9289	0.9593	0.9582

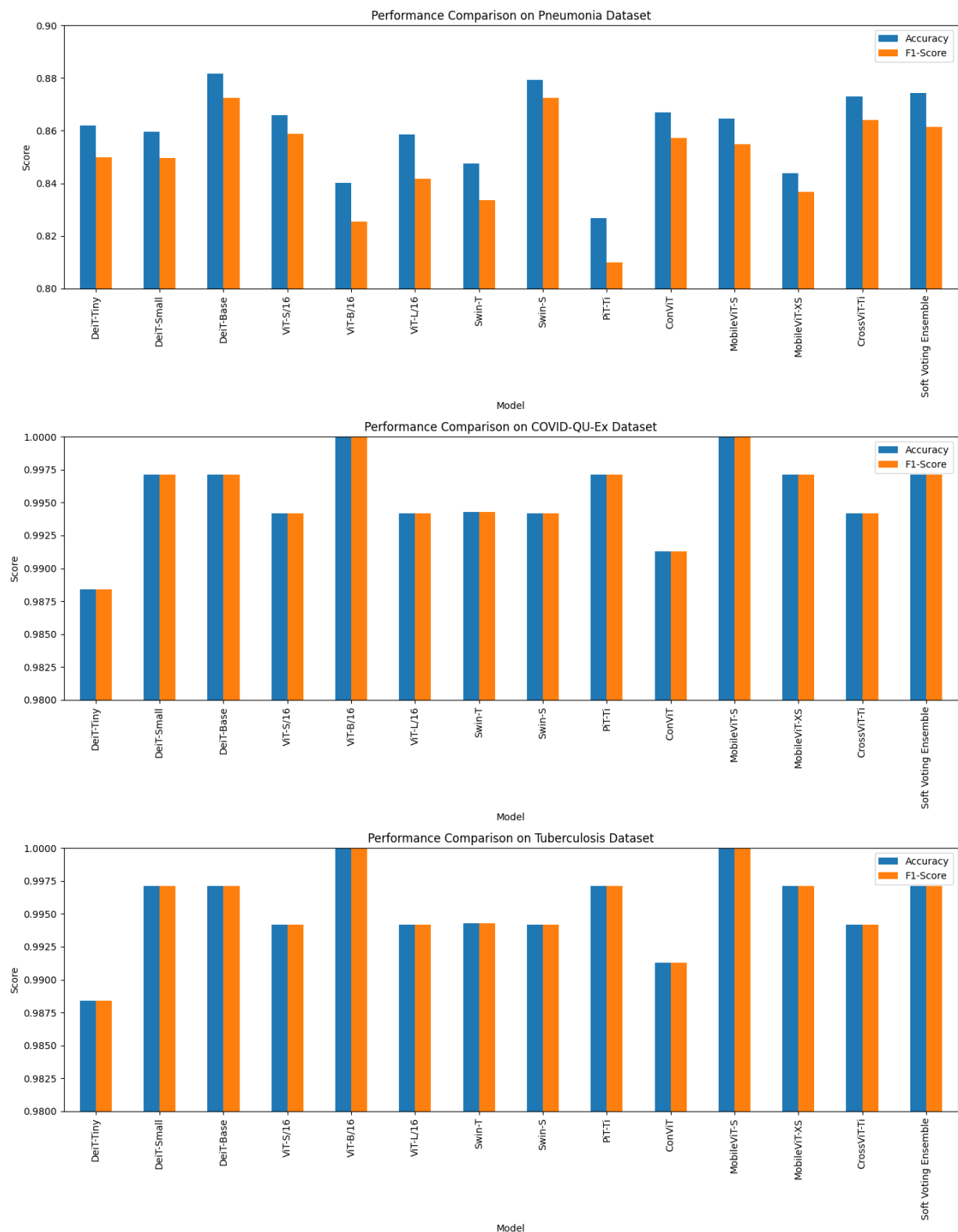


Figure 3. Comparison of model performance on Pneumonia, COVID-QU-Ex, and Tuberculosis datasets in terms of Accuracy and F1-Score.

On the COVID-QU-Ex dataset, a soft voting ensemble of the five best models achieved an accuracy of 0.9593 and an F1-score of 0.9582. Other models on COVID-QU-Ex also tended to have competitive performance. For example, the Swin-S model showed the highest precision for the Normal class (0.9281) and Pneumonia class (0.9336), while ViT-L showed the best recall for

the Normal class (0.9281). The ensemble leveraged the strengths of individual models to achieve an optimal balance between precision and recall, especially on the COVID-QU-Ex dataset, which has a more balanced distribution.

In the Tuberculosis Chest X-ray dataset, many models show near-perfect performance, with some models such as ViT-B/16 and MobileViT-S achieving an F1-score of 1. This may be due to the unbalanced distribution of the

dataset, where the number of Tuberculosis images (700) is far fewer than Normal images (3500). The model is able to classify tuberculosis with perfect precision and recall.

Table 6. Performance Comparison on the Tuberculosis Chest X-ray Dataset

Model	Precision		Recall		Accuracy	F1-Score
	COVID	Normal	COVID	Normal		
DeiT-Tiny	0.9902	0.9943	0.9714	0.998	0.9936	0.9884
DeiT-Small	1	0.998	0.9904	1	0.9984	0.9971
DeiT-Base	1	0.998	0.9904	1	0.9984	0.9971
ViT-S/16	0.9904	0.998	0.9904	0.998	0.9968	0.9942
ViT-B/16	1	1	1	1	1	1
ViT-L/16	0.9904	0.998	0.9904	0.998	0.9968	0.9942
Swin-T	0.9813	1	1	0.9961	0.9968	0.9943
Swin-S	1	0.9962	0.9809	1	0.9968	0.9942
PiT-Ti	1	0.998	0.9904	1	0.9984	0.9971
ConViT	1	0.9943	0.9714	1	0.9952	0.99132
MobileViT-S	1	1	1	1	1	1
MobileViT-XS	1	0.998	0.9904	1	0.9984	0.9971
CrossViT-Ti	1	0.9962	0.9809	1	0.9968	0.9942
Soft Voting Ensemble	1	0.998	0.9904	1	0.9984	0.9971
Top 5 Model						

3.2 Discussions

This study conducted a comprehensive evaluation of various transformer-based architectures for classification tasks on three chest X-ray medical image datasets, namely Pneumonia Chest X-ray, COVID-QU-Ex, and Tuberculosis. The results shown in Tables 4, 5, and 6 indicate that model performance is greatly influenced by the characteristics and complexity of the dataset.

In the Tuberculosis dataset, which is a binary classification task with significant class imbalance, almost all models show outstanding performance. The ViT-B/16 and MobileViT-S models achieved a score of 1 for all metrics. This shows that the distinguishing features between the Normal and Tuberculosis classes are very clear and have been learned effectively.

On the COVID-QU-Ex dataset, the performance of Transformer-based models is quite competitive. On this dataset, the top 5 soft voting ensemble models consistently outperform all single models, with a highest accuracy of 0.9593 and a highest F1-score of 0.9582. The success of this ensemble model highlights the benefits of combining the predictive power of several best models.

The Pneumonia chest X-ray dataset presents challenges in terms of model performance. Performance on this dataset was lower than on the other two, reflecting the greater difficulty of distinguishing bacterial from viral pneumonia. The main challenge in this dataset is the relatively low recall of the Viral class for all models, indicating significant visual feature overlap between Viral Pneumonia and Bacterial Pneumonia, which poses a challenge for the tested models.

Despite performance differences across datasets, Transformer models consistently showed strong capability in capturing global image context. The

findings of this study are in line with previous research highlighting the effectiveness of Vision Transformers for chest X-ray classification. Chen et al. [4], for example, reported that a fine-tuned Vision Transformer achieved 95.79% accuracy in four-class classification and up to 99.57% in three-class scenarios. The performance observed in this study is comparably strong across several Transformer architectures, reinforcing the capability of self-attention mechanisms to capture global contextual features in CXR images. However, differences in performance levels may stem from variations in dataset composition, preprocessing pipelines, and model configuration. Unlike Chen et al. [4], which focused on optimizing a single ViT model, this study provides a broader benchmarking perspective by systematically evaluating multiple Transformer architectures with different complexity levels.

Overall, this study shows that Transformer-based models tend to be competitive in Chest X-ray classification tasks. The ensemble approach with soft voting methods provides a significant improvement in performance. This proves that the ensemble strategy is an effective method for improving model robustness in Chest X-ray classification tasks.

Beyond accuracy, computational complexity is also critical for clinical deployment. Table 1 shows that parameter counts vary widely across architectures. Notably, larger models such as ViT-L/16 do not always deliver superior performance. For example, PiT-Ti and MobileViT-S achieved comparable or even better F1-scores while being far more efficient. This finding highlights that excessive model complexity can lead to diminishing returns, and that lightweight architecture may offer a more favorable balance between accuracy and efficiency for real-world medical applications. Unlike object-detection models that focus on localizing lesions, this study concentrates on whole-image

classification of chest X-rays, which is more aligned with the global diagnostic tasks considered.

3.3 Limitations

This study has several limitations. First, the Tuberculosis dataset exhibits an imbalanced class distribution, which may lead to inflated model performance and a potential risk of overfitting. Second, the datasets employed, including the Tuberculosis Chest X-ray and Pneumonia Chest X-ray datasets, contain a relatively small number of images, whereas the Vision Transformer model typically requires a large volume of data for effective training. Third, this study did not perform an in-depth analysis of model interpretability. Finally, only publicly available datasets were utilized, highlighting the need for further validation using images from external sources.

4. Conclusions

This study conducted a comparative analysis of 13 Transformer-based architecture models for Chest X-ray classification tasks on three different datasets, namely Pneumonia Chest X-ray, COVID-QU-Ex, and Tuberculosis. Each model was trained and evaluated using uniform settings, including preprocessing, augmentation, and consistent evaluation metrics to ensure a fair comparison. This study also applied an ensemble approach using the soft voting method on the five models with the highest F1-scores.

Based on the experimental results, this study provides several key contributions. First, it offers a systematic benchmarking of 13 Transformer architectures across three chest X-ray datasets under a unified experimental setting. Second, the study demonstrates that soft voting ensemble of top-performing models consistently improves classification performance compared to single architecture. Third, the findings reveal that lightweight Transformer variants can achieve competitive performance while maintaining computational efficiency, highlighting their suitability for real-world clinical deployment.

Transformer-based models demonstrate consistently strong performance across the Pneumonia Chest X-ray, COVID-QU-Ex, and Tuberculosis Chest X-ray datasets. On the Pneumonia dataset, a soft voting ensemble of top models achieved an accuracy of 0.8743 and an F1-score of 0.8615, outperforming individual models such as DeiT-Base, which achieved an F1-score of 0.8725. For COVID-QU-Ex, the ensemble of the best five models reached an accuracy of 0.9593 and an F1-score of 0.9582, effectively balancing precision and recall. In the Tuberculosis dataset, several models, including ViT-B/16 and MobileViT-S, achieved perfect performance with F1-scores of 1, likely due to the dataset's class imbalance, where Tuberculosis images are much fewer than Normal images. These results indicate that transformer-based models, particularly when combined in ensembles, provide robust and

highly effective classification across diverse chest X-ray datasets.

For future work, it is recommended to implement more targeted augmentation strategies, particularly for minority classes, to address data imbalance. Furthermore, validation on external datasets should be conducted to evaluate the model's generalization and robustness. Beyond performance metrics, investigating model explainability is also crucial to enhance interpretability and build trust in the model's predictions.

Acknowledgement

Authors would like to thank Telkom University and Kemdiktisaintek for financial supporting in this research contract number 041/LIT07/PPM-LIT/2025).

References

- [1] S. Safiri *et al.*, "Global burden of lower respiratory infections during the last three decades," *Front Public Health*, vol. 10, Jan. 2023, doi: [10.3389/fpubh.2022.1028525](https://doi.org/10.3389/fpubh.2022.1028525).
- [2] World Health Organization (WHO), "Pneumonia in Children." Accessed: Jul. 11, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/pneumonia>
- [3] A. Moussaid *et al.*, "On the Implementation of a Post-Pandemic Deep Learning Algorithm Based on a Hybrid CT-Scan/X-ray Images Classification Applied to Pneumonia Categories," 2023, doi: [10.3390/healthcare11050662](https://doi.org/10.3390/healthcare11050662).
- [4] T. Chen *et al.*, "A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images," *Healthcare Analytics*, vol. 5, Jun. 2024, doi: [10.1016/j.health.2024.100332](https://doi.org/10.1016/j.health.2024.100332).
- [5] M. Ketfi, M. Belahcene, and S. Bourennane, "Transfer Learning Fusion and Stacked Auto-encoders for Viral Lung Disease Classification," *New Gener Comput*, Nov. 2024, doi: [10.1007/s00354-024-00247-4](https://doi.org/10.1007/s00354-024-00247-4).
- [6] D. Shen, G. Wu, and H. Il Suk, "Deep Learning in Medical Image Analysis," *Annu Rev Biomed Eng*, vol. 19, pp. 221–248, Jun. 2017, doi: [10.1146/annurev-bioeng-071516-044442](https://doi.org/10.1146/annurev-bioeng-071516-044442).
- [7] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," May 2019, doi: [10.1007/s10916-018-1088-1](https://doi.org/10.1007/s10916-018-1088-1).
- [8] P. Rajpurkar *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," no. Computer Vision and Pattern Recognition, Dec. 2017, doi: <https://doi.org/10.48550/arXiv.1711.05225>
- [9] B. Mustapha, Y. Zhou, C. Shan, and Z. Xiao, "Enhanced Pneumonia Detection in Chest x-rays using Hybrid Convolutional and Vision Transformer Networks," *Curr Med Imaging Rev*, vol. 21, Jan. 2025, doi: [10.2174/0115734056326685250101113959](https://doi.org/10.2174/0115734056326685250101113959).
- [10] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: <http://arxiv.org/abs/2010.11929>
- [11] R. Fan and S. Bu, "Transfer-Learning-Based Approach for the Diagnosis of Lung Diseases from Chest X-ray Images," *Entropy*, vol. 24, no. 3, Mar. 2022, doi: [10.3390/e24030313](https://doi.org/10.3390/e24030313).
- [12] K. Kawadkar, "Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Medical Image Classification," Jul. 2025, [Online]. Available: <http://arxiv.org/abs/2507.21156>
- [13] T. Sanida and M. Dasygenis, "A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system," *Applied Intelligence*, vol. 54, no. 6, pp. 4756–4780, Mar. 2024, doi: [10.1007/s10489-024-05420-2](https://doi.org/10.1007/s10489-024-05420-2).

- [14] M. S. Al Reshan *et al.*, "Detection of Pneumonia from Chest X-ray Images Utilizing MobileNet Model," *Healthcare (Switzerland)*, vol. 11, no. 11, Jun. 2023, doi: [10.3390/healthcare11111561](https://doi.org/10.3390/healthcare11111561).
- [15] P. Guddati, S. Dash, and R. K. Tripathy, "FPGA Implementation of the Proposed DCNN Model for Detection of Tuberculosis and Pneumonia Using CXR Images," *IEEE Embed Syst Lett*, 2024, doi: [10.1109/LES.2024.3370833](https://doi.org/10.1109/LES.2024.3370833).
- [16] A. Elrefaei, A. Hossam are, F. A. Mostafa, L. A. Elrefaei, M. M. Fouda, and A. Hossam, "Enhancing Diagnostic Accuracy With Ensemble Techniques: Detecting Covid-19 And Pneumonia On Chest X-Ray Images," 2024, doi: <https://doi.org/10.5455/jjcit.71-1716799666>
- [17] M. Pal *et al.*, "A comparative analysis of the binary and multiclass classified chest X-ray images of pneumonia and COVID-19 with ML and DL models," *Open Medicine (Poland)*, vol. 20, no. 1, Jan. 2025, doi: [10.1515/med-2024-1110](https://doi.org/10.1515/med-2024-1110).
- [18] Z. Liang, Z. Xue, S. Rajaraman, and S. Antani, "Automated quantification of SARS-CoV-2 pneumonia with large vision model knowledge adaptation," *New Microbes New Infect*, vol. 62, Dec. 2024, doi: [10.1016/j.nmni.2024.101457](https://doi.org/10.1016/j.nmni.2024.101457).
- [19] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: [10.1038/s41598-024-52703-2](https://doi.org/10.1038/s41598-024-52703-2).
- [20] T. Wang *et al.*, "PneuNet: deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using Vision Transformer," *Med Biol Eng Comput*, vol. 61, no. 6, pp. 1395–1408, Jun. 2023, doi: [10.1007/s11517-022-02746-2](https://doi.org/10.1007/s11517-022-02746-2).
- [21] J. Ko, S. Park, and H. G. Woo, "Optimization of vision transformer-based detection of lung diseases from chest X-ray images," *BMC Med Inform Decis Mak*, vol. 24, no. 1, Dec. 2024, doi: [10.1186/s12911-024-02591-3](https://doi.org/10.1186/s12911-024-02591-3).
- [22] V. Padmavathi and K. Ganesan, "Metaheuristic optimizers integrated with vision transformer model for severity detection and classification via multimodal COVID-19 images," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: [10.1038/s41598-025-98593-w](https://doi.org/10.1038/s41598-025-98593-w).
- [23] M. Pal, S. Parija, and G. Panda, "An effective ensemble approach for classification of chest X-ray images having symptoms of COVID: A precautionary measure for the COVID-19 subvariants," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 8, Jun. 2024, doi: [10.1016/j.prime.2024.100547](https://doi.org/10.1016/j.prime.2024.100547).
- [24] B. Oltu, S. Güney, S. E. Yuksel, and B. Dengiz, "Automated classification of chest X-rays: a deep learning approach with attention mechanisms," *BMC Med Imaging*, vol. 25, no. 1, Dec. 2025, doi: [10.1186/s12880-025-01604-5](https://doi.org/10.1186/s12880-025-01604-5).
- [25] E. Selvano, A. Y. Paulindino, G. N. Elwirehardja, and B. Pardamean, "Evaluating Self-Supervised Pre-Trained Vision Transformer On Imbalanced Data For Lung Disease Classification," *ICIC Express Letters, Part B: Applications*, vol. 15, no. 1, pp. 83–89, Jan. 2024, doi: [10.24507/icicelb.15.01.83](https://doi.org/10.24507/icicelb.15.01.83).
- [26] S. Kadry *et al.*, "COVID19 detection in chest x-ray using vision-transformer with different patch dimensions," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 3438–3446. doi: [10.1016/j.procs.2024.04.324](https://doi.org/10.1016/j.procs.2024.04.324).
- [27] H. Yang, L. Wang, Y. Xu, and X. Liu, "CovidViT: a novel neural network with self-attention mechanism to detect Covid-19 through X-ray images," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 3, pp. 973–987, Mar. 2023, doi: [10.1007/s13042-022-01676-7](https://doi.org/10.1007/s13042-022-01676-7).
- [28] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Jan. 2019, [Online]. Available: <http://arxiv.org/abs/1706.09516>
- [29] A. M. Tahir *et al.*, "COVID-19 infection localization and severity grading from chest X-ray images," *Comput Biol Med*, vol. 139, Dec. 2021, doi: [10.1016/j.compbiomed.2021.105002](https://doi.org/10.1016/j.compbiomed.2021.105002).
- [30] T. Rahman *et al.*, "Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020, doi: [10.1109/ACCESS.2020.3031384](https://doi.org/10.1109/ACCESS.2020.3031384).
- [31] R. Yulvina *et al.*, "Hybrid Vision Transformer and Convolutional Neural Network for Multi-Class and Multi-Label Classification of Tuberculosis Anomalies on Chest X-Ray," *Computers*, vol. 13, no. 12, p. 343, Dec. 2024, doi: [10.3390/computers13120343](https://doi.org/10.3390/computers13120343).
- [32] Y. Hadhoud *et al.*, "From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images," *Diagnostics*, vol. 14, no. 23, Dec. 2024, doi: [10.3390/diagnostics14232754](https://doi.org/10.3390/diagnostics14232754).
- [33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," May 2016, [Online]. Available: <http://arxiv.org/abs/1506.02640>