Comparative Study of Transformer Architectures for Pneumonia Detection in Chest X-Rays

Anandha Ragaven Ravi

August 1, 2025

1 Introduction

Since pneumonia continues to be a major global source of morbidity and death, prompt and precise diagnosis is essential. For automated medical picture categorization, recent developments in deep learning—in particular, Transformer-based models—offer a promising path. Five cutting-edge Vision Transformer models are compared in this study and used to chest X-ray images in order to identify pneumonia.Building on previous research in medical image classification with convolutional neural networks (CNNs), this study switches to Transformer-based methods that have demonstrated potential in the text and natural image domains. Finding the Transformer design that works best for clinical pneumonia screening is our goal.

2 Methods

This work made use of the Chest X-Ray Pneumonia dataset from Kaggle, which included labeled chest radiograph pictures of either healthy controls or patients with a diagnosis of pneumonia. To guarantee equitable representation, we downsampled the dominant class in order to balance the dataset. To comply with the input specifications of Transformer-based models, all photos were normalized and scaled to 224 by 224 pixels. Using PyTorch and Hugging Face Transformers, we refined five pre-trained Transformer models on this dataset: Vision Transformer (ViT), Swin Transformer, DeiT, PoolFormer, and MobileViT. For binary classification (pneumonia vs. normal), each model was trained under supervision, with early termination determined by validation loss. using F1 Score, AUC, Accuracy, Precision, and Recall. Confusion matrices, precision-recall graphs, t-SNE embeddings, loss curves, and ROC curves were among the visualizations. Using early stopping and optimum learning rates, the same amount of epochs were used to train each model.

3 Results and Discussion

MobileViT outperformed the other three models in the majority of criteria, with an F1 Score of 0.901, Accuracy of 86.4%, and AUC of 0.973. Excellent recall (0.995 and 1.000, respectively) was demonstrated by ViT and PoolFormer, but their lesser accuracy resulted in lower F1 scores. Compared to MobileViT, Swin Transformer had a slightly lower F1 but the highest AUC (0.977).These findings demonstrate the trade-offs that exist in clinical settings between model complexity, memory, and accuracy. Lightweight models, like MobileViT, are appropriate for real-time applications in healthcare settings with limited resources because they provide a good compromise between performance and deployability.

4 Impact

This work shows that lightweight Vision Transformer designs for the identification of pneumonia in chest X-rays are both feasible and effective. In addition to offering a repeatable standard for Transformer-based medical imaging, it delivers insightful information on clinically relevant model selection techniques.

Model	Accuracy	Precision	Recall	F1 Score	AUC
MobileViT	0.863782	0.826552	0.989744	0.900817	0.973099
ViT Base	0.851000	0.810000	0.994900	0.893000	0.965700
PoolFormer	0.847756	0.804124	1.000000	0.891429	0.972315
Swin Transformer	0.844551	0.803313	0.994872	0.888889	0.976742
DeiT	0.833333	0.790650	0.997436	0.882086	0.970321

Table 1: Best Transformer Models by F1 Score and AUC:

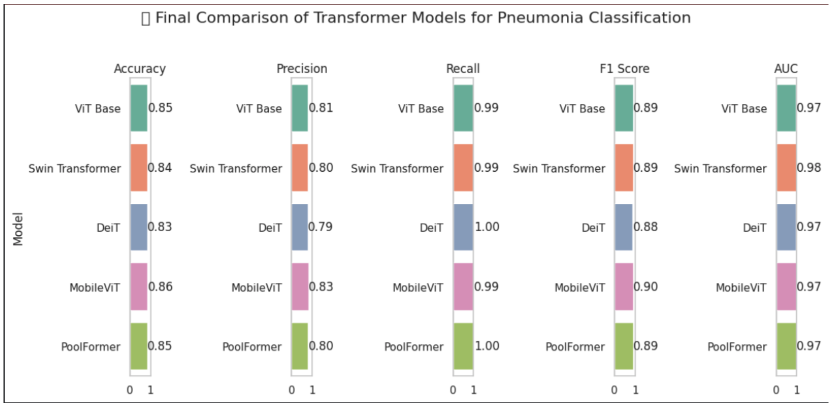

Figure 1: Comparision metrics of Transformer Models for Pneumonia Classification .

5 Acknowledgements

We thank the Hugging Face and PyTorch communities' open-source contributors as well as the Semcer Center for Healthcare Innovation for their assistance. The Stevens Institute of Technology's institutional computing resources provided some help for this endeavor. This study did not receive any outside support.

6 References

[1] Vaswani et al., 2017. ”Attention is All You Need.” [2] Dosovitskiy et al., 2020. ”An Image is Worth 16x16 Words.” [3] Liu et al., 2021. ”Swin Transformer: Hierarchical Vision Transformer.” [4] Mehta and Rastegari, 2022. ”MobileViT: Light-weight Vision Transformer.” [5] Kermany et al., 2018. ”Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.”