A Comparative Study of Swin Transformer and CNN for Pneumonia Detection in Chest X-Rays

Daniel Levkovitz - 205789167, Or Bachar - 205972805
Reichman University
Project GitHub Repository

August 9, 2025

Abstract

This paper presents a direct comparative study of a Swin Transformer and a Convolutional Neural Network (CNN) for the binary classification of pneumonia from chest X-ray images. Pneumonia is one of the leading causes of global morbidity and mortality, making a rapid and accurate diagnosis essential for effective treatment. While chest radiography is a common diagnostic tool, its interpretation can be challenging and subject to variability. This study evaluates two distinct deep learning paradigms—the well-established CNN and the state-of-the-art Swin Transformer—to determine which architecture is better suited for this critical medical application. We implement both models, training them on the same dataset while leveraging advanced techniques such as Focal Loss [1], CutMix augmentation [7], and k-fold cross-validation [5] to ensure robust evaluation. The results demonstrate that the Swin Transformer, augmented with modern training strategies, achieves superior diagnostic performance, highlighting its potential to improve clinical decision making in combating this global health threat.

1. Introduction

Pneumonia remains a leading cause of death worldwide, particularly affecting children and the elderly. The timely and accurate diagnosis of this respiratory infection is paramount for improving patient outcomes. Chest X-rays are the most common imaging modality for this purpose, though their interpretation is often challenging, relying on the availability and expertise of radiologists. This creates a critical need for reliable and accessible automated diagnostic tools, which motivates the exploration of advanced artificial intelligence to augment clinical workflows.

For years, Convolutional Neural Networks (CNNs) have been the standard for medical image analysis. Their architecture, which is inspired by the human visual cortex, excels at learning hierarchical spatial features, making them highly effective for classification and segmentation tasks. However, the recent advent of Vision Transformers (ViTs), adapted from natural language processing, has introduced a new paradigm. Architectures like the Swin Transformer have shown immense promise by using a hierarchical structure and an efficient, shifted-window self-attention mechanism to capture both local and long-range dependencies more effectively than many CNNs [2]. Studies have successfully applied Swin Transformer variants to tasks like chest X-ray classification, often demonstrating superior performance over traditional CNN baselines by better modeling these dependencies [4].

This paper provides a direct comparative analysis of these two architectures for pneumonia detection. We implement and train a well-established CNN model and a state-of-the-art Swin Transformer on the same dataset to rigorously evaluate their diagnostic performance, identify their respective strengths, and analyze the clinical trade-offs of each approach. By combining these models with state-of-the-art training and regularization techniques—including Focal Loss to address class imbalance [1, 6], CutMix augmentation to improve generalization [7, 3], and k-fold cross-validation for robust evaluation [5]—we aim to provide clear insights into the optimal architectural design for this critical medical task.

2. Methodology

2.1. Vision Transformer Approach: Swin Transformer

The standard Vision Transformer's (ViT) use of global self-attention incurs a computational complexity that is quadratic with image size, posing challenges for high-resolution medical imaging. The Swin Transformer addresses this by computing self-attention within local, non-overlapping windows and enabling cross-window connections via a shifted-window mechanism. This hierarchical structure is conceptually similar to that of CNNs and is well-suited for dense prediction tasks.

For this study, we utilized the `swin_base_patch4_window7_224` variant, pre-trained on ImageNet, which processes input images resized to 224×224 pixels. The standard classifier was replaced with a custom sequential head consisting of a LayerNorm layer, a dropout layer (p=0.2), a linear layer that halves the feature dimension, a GELU activation function, another dropout layer (p=0.1), and a final linear output layer for the 'Normal' and 'Pneumonia' classes. This architecture, including the custom head, contains approximately 88 million trainable parameters.

2.2. Convolutional Neural Network (CNN) Approach

To establish a strong convolutional baseline, we implemented a deep CNN configuration identified through a cross-validation–driven architecture search. The final architecture consists of six sequential convolutional blocks, each using a kernel size of 5, Swish activation, and 2×2 max pooling. Batch normalization was disabled, while dropout with a probability of 0.285 was applied to enhance generalization. The number of feature channels starts at 48 and increases by a multiplicative factor of approximately 1.96 with each block. A global adaptive max pooling layer feeds into a fully connected layer with 512 hidden units before the final output logits. This configuration contains approximately 34 million trainable parameters.

2.3. Experimental Design

Both models were trained and evaluated on the same public chest X-ray dataset, where pneumonia cases constitute approximately 74% of the samples. All images were resized to 224×224 pixels. The Swin Transformer was normalized using ImageNet statistics, while the CNN used its own internal per-channel normalization.

Data Augmentation: A strong augmentation strategy was employed for the Swin Transformer to mitigate overfitting, including geometric transformations (ShiftScaleRotate, HorizontalFlip, ElasticTransform), intensity adjustments (RandomBrightnessContrast, RandomGamma), and noise injection (GaussNoise, MotionBlur). Critically, CutMix augmentation was used with a probability of 0.5 to encourage the model to learn from regional features. In contrast, the CNN was trained with lighter augmentations aligned with medical imaging best practices, such as RandomHorizontalFlip, small rotations, and RandomResizedCrop; CutMix was not used.

Training Parameters: The Swin Transformer was trained using the AdamW optimizer (learning rate 2×10^{-4}, weight decay 0.02) with a CosineAnnealingWarmRestarts scheduler and Focal Loss. Addressing the significant class imbalance was critical; initial experiments with conventional Focal Loss parameters (e.g., $\alpha = 0.25, \gamma = 2.0$) resulted in poor performance for the underrepresented 'Normal' class. This motivated a tuning process, which led to final optimized hyperparameters of $\alpha = 0.6$ and $\gamma = 1.2$ to better balance model sensitivity and specificity. The CNN was trained from scratch using AdamW (learning rate 3×10^{-4}, weight decay 1×10^{-4}) with a cosine learning rate schedule and class-weighted Cross-Entropy Loss to mitigate imbalance. Both models used a batch size of 32 and an early stopping protocol, with a patience of 3 for the Swin Transformer and 5 for the CNN.

Evaluation Strategy: The Swin Transformer's performance was evaluated using 5-fold cross-validation on the training set for model selection, followed by a final evaluation on a held-out test set. The CNN underwent a 3-fold cross-validation architecture search, after which the best-performing configuration was retrained and evaluated on the same held-out test set.

3. Results

The Swin Transformer demonstrated superior diagnostic capability, consistently outperforming the optimized CNN baseline across key metrics on the held-out test set.

3.1. Quantitative Performance

The Swin Transformer achieved an accuracy of 86.1%, a weighted F1-score of 0.852, and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.977 (Figure 1). In comparison, the CNN achieved an accuracy of 82.5%, a weighted F1-score of 0.813, and an AUC-ROC of 0.915 (Figure 2). The higher AUC for the Swin Transformer confirms its superior threshold-independent discrimination between the 'Normal' and 'Pneumonia' classes.

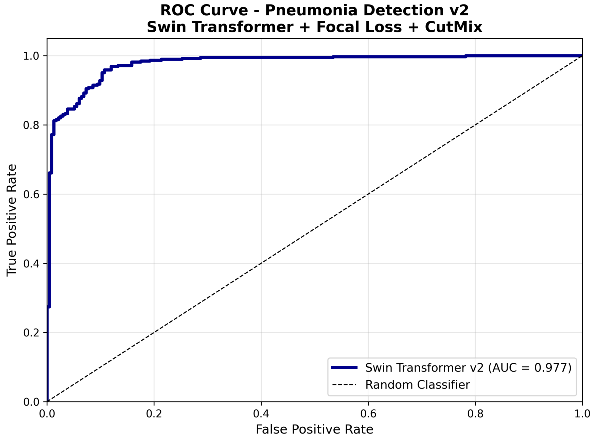

Figure 1: ROC Curve of the Swin Transformer model on the test set, showing an AUC of 0.977.

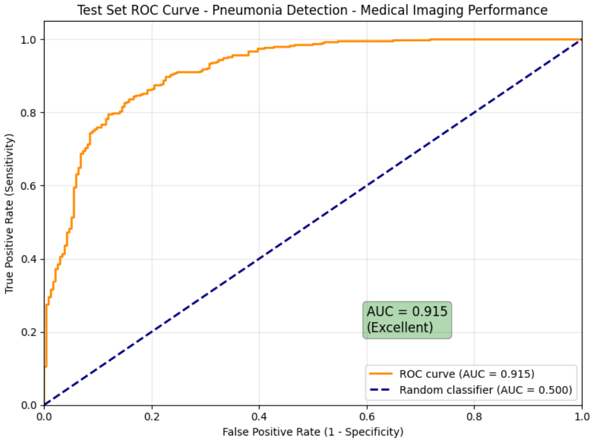

Figure 2: ROC Curve of the CNN model on the test set, showing an AUC of 0.915.

3.2. Per-Class Performance and Clinical Relevance

Analysis of the confusion matrices provides further insight into the clinical utility of each model. The Swin Transformer correctly identified 388 of 390 pneumonia cases, achieving a pneumonia recall (sensitivity) of 99.5% with only 2 false negatives (Figure 3). For the 'Normal' class, it achieved a recall of 63.7% (149 true negatives, 85 false positives). **The model's precision for the pneumonia class was 82.0%, while**

its precision for the normal class was an excellent 98.7%. The CNN, while also emphasizing sensitivity, was less specific; it achieved a pneumonia recall of 97.9% (382 true positives, 8 false negatives) and a 'Normal' class recall of 56.8% (133 true negatives, 101 false positives) (Figure 4).

Clinically, both models are effective at minimizing false negatives for pneumonia, which is critical in a screening context. However, the Swin Transformer not only reduces the number of missed pneumonia cases (2 vs. 8) but also produces fewer false positives for the 'Normal' class (85 vs. 101), suggesting it would lead to fewer unnecessary follow-up procedures. This indicates a better balance between sensitivity and specificity. The precision-recall curves (Figure 5 and Figure 6) further support this, showing the Swin Transformer maintains high precision across a wider range of recall values compared to the CNN.

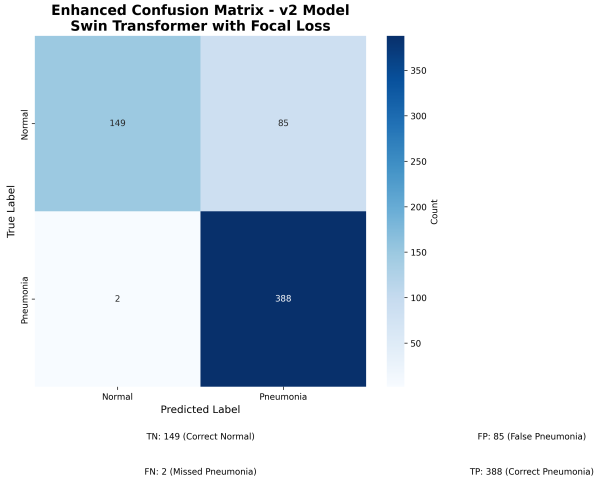

Figure 3: Swin Transformer confusion matrix on the test set.

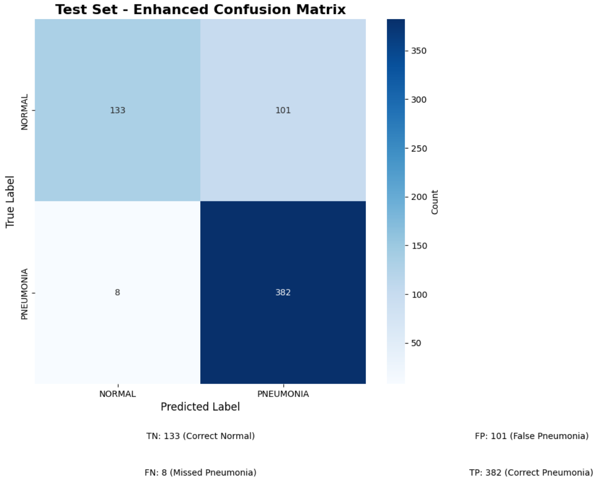

Figure 4: CNN confusion matrix on the test set.

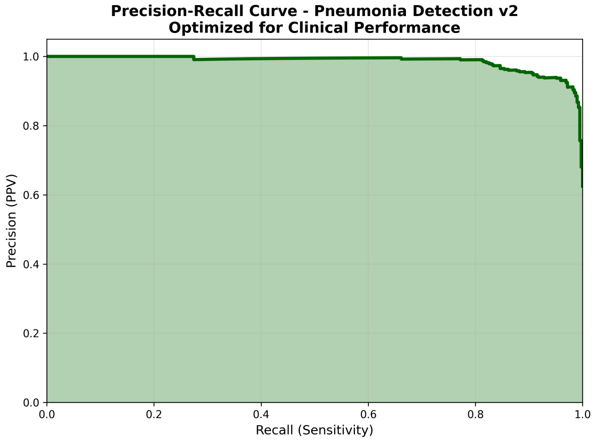

Figure 5: Swin Transformer Precision-Recall curve.

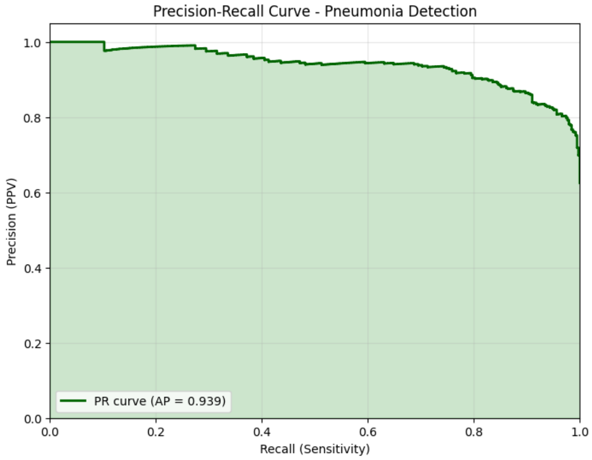

Figure 6: CNN Precision-Recall curve.

3.3. Training Dynamics and Generalization

Both models demonstrated stable convergence during cross-validation. The Swin Transformer achieved a mean validation accuracy of $94.7\% \pm 1.2\%$ across 5 folds. The selected CNN architecture achieved a mean validation accuracy of $96.0\% \pm 1.3\%$ across 3 folds. For both models, the performance gap between cross-validation and the final test set suggests the test set distribution was more challenging.

4. Discussion

This comparative study demonstrates that the Swin Transformer, when combined with Focal Loss, CutMix, and a strong augmentation strategy, provides superior overall performance for pneumonia detection. It achieved an outstanding pneumonia recall of 99.5% and an AUC-ROC of 0.977, indicating excellent and robust discrimination. Its stability during 5-fold cross-validation further confirmed its ability to generalize well.

Both models highlight a classic clinical trade-off: in prioritizing high sensitivity to avoid missing pneumonia cases, both sacrificed some specificity. The CNN baseline, optimized via an extensive architecture search, also proved clinically viable but demonstrated a less favorable balance, with a specificity of 56.8% for the 'Normal' class. While the Swin Transformer also made this trade-off, achieving a specificity of 63.7%, its superior overall metrics indicate it struck a more effective balance between minimizing missed cases and limiting false alarms.

From a computational standpoint, the comparison also reveals interesting trade-offs. The Swin Transformer is a significantly larger model, with approximately 88 million parameters compared to the CNN's 34 million. Despite its size, the benefits of transfer learning from a pre-trained model were evident; it achieved its optimal performance in approximately 37 minutes. The scratch-trained CNN, while smaller, required a similar training time of about 39 minutes to converge, highlighting the efficiency gains that pre-trained weights can provide.

The key takeaway from this experiment is that modern training techniques materially improve performance. Methods that explicitly focus the learning process on difficult examples, such as Focal Loss, and regularization strategies that enhance local feature diversity, like CutMix and strong augmentations, enable a model to achieve better threshold-independent performance and generalization. While cross-validation can produce a strong scratch-trained CNN, the Swin Transformer architecture, coupled with these advanced training methods, generalized more effectively on this dataset.

5. Conclusion

For clinical applications where minimizing missed pneumonia cases is the primary objective, the Swin Transformer is the preferred model. It offers a superior balance of sensitivity and specificity, reducing both false negatives for pneumonia and false positives for normal cases. The optimized CNN serves as a robust baseline and could be a candidate for future work in model ensembling or knowledge distillation. Ultimately, the deployment of either model would require calibrating operating thresholds based on the specific clinical costs associated with false positives versus false negatives.

References

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[3] Raheel Siddiqi and Sameena Javaid. Deep Learning for Pneumonia Detection in Chest X-ray Images: A Comprehensive Survey. *Journal of Imaging*, 10(8):176, 2024.

[4] Sara Taslimi, Saba Taslimi, Narges Fathi, Mahsa Salehi, and Mohammad Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2201.03319*, 2022.

[5] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017.

[6] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Medical Image Analysis*, 75:102239, 2022.

[7] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.