Pneumonia Detection Using Vision Transformers and Thresholding Methods

Olivier Makuch **Olivia Mirijello** **Alyssa Oleas** **Pedram Peiroasfia**

Abstract

This project explores how Vision Transformers (ViTs) can be used to detect pneumonia from chest X-ray images, as an alternative to the traditional Convolutional Neural Networks (CNNs). Although effective, CNNs require extensive data and struggle with capturing certain image dependencies. ViTs, using transformer architecture, address this by modeling long-range pixel dependencies without convolutional steps. The goal of this project is to improve pneumonia detection, which is particularly important in regions with limited resources. By integrating ViTs with thresholding techniques, we expect to improve classification performance beyond the benchmarks set by existing CNN models, particularly in datasets with limited size and diversity. This approach is intended not only to enhance diagnostic accuracy but also to contribute significantly to reducing pneumonia-related mortality through improved medical imaging technologies.

1 Introduction

In today's medical imaging field, spotting and sorting pneumonia from chest X-rays is crucial but tough, with big impacts on health care. The challenge in nailing down pneumonia from these images calls for advanced computational models that beat old methods in accuracy, trust, and speed. Given pneumonia's huge effect, especially where medical resources are scarce, our model uses Vision Transformers (ViT) with basic thresholding methods aiming for top-notch accuracy in detecting pneumonia. This could be a key progress in using deep learning for medical imaging.

Traditional medical imaging, especially in detecting pneumonia, has mainly relied on Convolutional Neural Networks (CNNs). CNNs excel in extracting layered features from images, greatly advancing classification and detection. However, they have downsides, like needing lots of data for training and struggling to capture image dependencies. Vision Transformers (ViTs) represent a major change, moving away from CNNs. ViTs use the transformer structure to treat images as pixel sequences. This approach lets ViTs model long-range dependencies directly, bypassing convolution, for a more dynamic image representation. ViTs have started beating CNNs in many benchmarks, suggesting they could reshape medical image analysis standards.

Our project taps into the transformative power of ViTs as a new method for detecting pneumonia in chest X-rays. By moving past CNNs' limitations, ViTs open up better possibilities for classification, even with small or diverse datasets. We are doing a detailed comparison, looking at pure ViTs and one mixed with thresholding methods to fully gauge their impact. Our goal is to not just meet, but exceed the benchmarks set by previous CNN models, like those by Stephen et al. [1]. We are aiming to shift the detection and classification of pneumonia paradigm, potentially lowering pneumonia-related deaths, particularly in underserved areas.

The paper is structured into six main sections. Section 2, the Literature Review, starts with a discussion on CNNs and then explores Transformer models. Section 3, Materials and Methods, is divided into detailed sub-sections covering everything from dataset acquisition to performance

Preprint. Under review.

metrics. Section 4, Model Implementation, shows how both CNNs and ViTs are used, comparing pure ViTs to those combined with thresholding techniques. Section 5 presents the empirical results of these tests in Results. The paper wraps up with Section 6, offering a summary and looking at future research avenues.

2 Literature Review

2.1 Convolutional Neural Networks

In this section, we provide an in-depth review of relevant literature on using deep learning for pneumonia detection and classification from chest X-ray (CXR) images. We evaluate the effectiveness of various CNN architectures and their performance metrics. Specifically, six key studies are chosen for thorough analysis, each presenting unique methods and results. Our review seeks to identify progress, challenges, and areas needing further study in medical image analysis via deep learning. Stephen et al. [1] introduced a CNN with four convolution layers, trained from scratch on a pediatric CXR dataset. The data split included 2134 images for validation and expanded the training set through augmentation techniques. They achieved 93.7% accuracy on the validation set but lacked extensive validation on other metrics. Siddiqi [2] developed an 18-layer customized DCNN trained on a pediatric CXR dataset, achieving 94.3% accuracy, with a high sensitivity of 99% but lower specificity at 86%.

Verma et al. [3] introduced a specialized CNN for classifying lung nodules in CXR images, achieving 99.01% accuracy, but with scant experimental details. Omar and Babalik [4] proposed a customized CNN for pneumonia detection from CXRs, noting an accuracy of 87.65%, but not disclosing other performance metrics. Bhatt et al. [5] developed a cost-effective nine-layer CNN for pneumonia classification from CXRs, reaching 96.18% accuracy, 98.16% sensitivity, and 91.29% specificity, yet their method diverged from standard data splits and showed limited specificity. Wu et al. [6] utilized an adaptive median filter CNN combined with Random Forest (ACNN–RF) for detecting pneumonia in CXRs, achieving 96.9% accuracy, with precision and recall of 90% and 95%, respectively, though the precision was notably low. Sarkar et al. [7] implemented a preprocessing technique followed by a deep learning model, achieving 98.82% accuracy and an AUC of 0.99726 on CXRs, although they too deviated from conventional data division protocols. Rajaraman et al. [8] enhanced CNN transparency by providing visual explanations of model predictions and activations, improving accuracy using the VGG16 architecture. Abiyev and Ma'aitah [9] explored various neural network approaches for pathology detection from chest X-rays, including customized CNNs, competitive neural networks (CpNNs) with unsupervised learning, and backpropagation neural networks (BPNNs) with supervised learning. CpNNs showed faster convergence compared to CNNs, but CNNs maintained higher accuracy (92.4%) than CpNN (80.04%), BPNN (89.57%), VGG16 (86.00%), VGG19 (92.00%), and CNN with GIST (92.00%). However, there remains a need for further improvement in accuracy, and additional performance metrics like sensitivity and specificity were not discussed.

2.2 Transformers

In the rapidly advancing domain of medical imaging, the adoption of sophisticated deep learning methods, especially Vision Transformers (ViTs), has introduced innovative approaches for diagnosing significant diseases like pneumonia and COVID-19 from chest X-rays. This literature review compiles results from recent studies utilizing ViT models, highlighting their enhanced effectiveness and potential benefits over conventional Convolutional Neural Networks (CNNs) in medical diagnostics. Matsoukas et al. [10] challenge the long-standing predominance of CNNs in automated medical diagnosis, presenting ViTs as a formidable alternative. Through extensive testing across various medical image datasets, they find that ViTs, particularly when pretrained on ImageNet and further honed with self-supervised learning techniques, not only rival but occasionally surpass CNNs. Their results indicate a movement towards embracing ViTs in medical imaging, credited to their proficiency in capturing long-range dependencies and providing greater explainability via attention mechanism.

The study by Angara and Thirunagaru [11] explores Vision Transformers (ViTs) for COVID-19 diagnosis using chest X-rays, underscoring their success in image recognition tasks. It compares transformer-based models—ViT, Swin Transformer, MaxViT, and PVT v2—highlighting their efficacy with detection accuracies reaching 99.5%. This analysis is based on the comprehensive

COVIDx CXR-3 dataset, demonstrating ViTs' transformative impact in medical diagnostics. The study shows transformer models' superiority over conventional CNNs and the benefits of transfer learning for peak performance. It presents strong evidence of ViTs' potential to enhance medical image analysis, emphasizing their scalability and ability to discern complex patterns essential for accurate diagnosis. The study by Li et al. [12] marks a critical advancement in medical imaging research, emphasizing the superior capabilities of Transformer models, particularly Vision Transformers (ViTs), compared to traditional Convolutional Neural Networks (CNNs). The review highlights Transformers' ability to detect long-range dependencies and process images in patch sequences, providing a comprehensive perspective vital for pinpointing anomalies in CT scans, MRI, and X-rays. The paper not only demonstrates the versatility and effectiveness of Transformers in medical diagnostics, but also suggests future research avenues to further enhance their diagnostic precision and clinical value, indicating a substantial potential impact on patient care and treatment outcomes.

Usman et al. [13] examines the application of Vision Transformers (ViTs) in detecting pneumonia and other conditions from chest X-rays through transfer learning. The study utilizes a standard ViT model initially trained on ImageNet, then fine-tuned using data from the CheXpert and Pediatric Pneumonia datasets. This enhancement bolsters the model's ability to identify subtle chest anomalies due to ViTs' proficiency in managing long-range dependencies. Although fine-tuning shows limited overall performance gains, the research suggests that domain adaptation and other transfer learning techniques could enhance medical image classification. Broadening the scope of fine-tuning datasets may improve diagnostic accuracy, efficiency, and clarity, potentially revolutionizing patient care with more advanced diagnostic tools. These findings underscore the transformative potential of ViTs in medical imaging, indicating that transitioning from CNNs to ViTs could significantly benefit diagnosis processes, including for pneumonia and COVID-19, thus enhancing patient outcomes.

3 Materials and Methods

3.1 Dataset

The dataset used in this study comes from Kaggle's "Chest X-Ray Images (Pneumonia)" collection by Paul Mooney, featuring 5,863 X-ray images split into pneumonia and normal categories. These images are distributed across train, validation, and test sets to facilitate structured model training and evaluation. Analysis revealed that the validation set, containing only 16 images divided equally between the two categories, was too small for effective validation. To address this, we redistributed the images by merging and then evenly splitting the test and validation sets. This restructuring ensures a more balanced dataset, enhancing the reliability of the machine learning models we will discuss in the following sections.

3.2 Data Analysis, Preprocessing, and Augmentation

To counteract the challenges of a limited and heterogeneous dataset, we implemented a comprehensive strategy involving data preprocessing and augmentation to enhance model performance and reduce overfitting risks. Given the small size of our dataset, we expanded it artificially through augmentation techniques to avoid overfitting. The upcoming section will outline the preprocessing measures applied, following the guidelines recommended by Stephen et al. [1]. This section will also include a succinct data analysis to provide further insights into the improvements and modifications made to the dataset.

Preprocessing medical images for CNN modeling is essential to reduce overfitting and enhance model generalizability, thereby increasing robustness compared to unprocessed images. Data augmentation techniques such as rescaling, rotation, width and height shifts, shearing, and random zooming are critical for artificially enhancing both the quality and quantity of image data. These methods introduce diversity, decreasing the model's dependency on repetitive image patterns. Rescaling adjusts pixel values to fall within a specific range, which is crucial for optimizing network parameters. For grayscale X-ray images, normalizing pixel values to the 0 to 1 range by dividing by 255 is standard. Rotation, shifts, shearing, and zooming alter image characteristics, preventing the model from overly relying on centered images and offering varied perspectives of the modeled features. This not only diversifies the training data but also better represents real-world variability in chest X-rays, as detailed in Table 1.

Method	Setting
Rescale	1/255
Rotation range	40
Width shift	0.2
Height shift	0.2
Shear range	0.2
Zoom range	0.2
Horizontal Flip	True

Table 1: Image augmentation settings used for CNN Model, in accordance with Stephen et al. [1] best results.

3.3 Insights into Identifying Pneumonia Without Machine Learning

The enhancement of medical images through thresholding techniques allows for a more explicit representation of radiographic features, potentially aiding in the visual diagnosis of conditions such as pneumonia. By applying various thresholding methods to the chest X-ray images, we sought to explore the extent to which these transformations could reveal the asymmetrical patterns typical of pneumonia that are usually not present in normal, symmetrical lung presentations.

Thresholding serves to isolate and emphasize areas of differing intensities within the X-ray images. In the case of pneumonia, thresholding can make the asymmetry caused by the disease more apparent even with the eyes. Such visual contrasts are less pronounced in standard imaging without these computational techniques. Our empirical testing spanned a range of thresholding approaches applied to the original images, including binary thresholding, inverse binary thresholding, truncate thresholding, thresholding to zero, and inverse thresholding to zero [14], with varying threshold values (100, 120, 140, 160, and 180). Each method and threshold level offered a unique perspective on the imaging data. Thresholding is an image processing method that alters pixel intensities based on a predefined threshold value. This allows for the distinction between regions of interest and the background, which is especially useful in medical diagnostics for highlighting abnormal patterns. The details on the thresholding approaches is located in Appendix B.

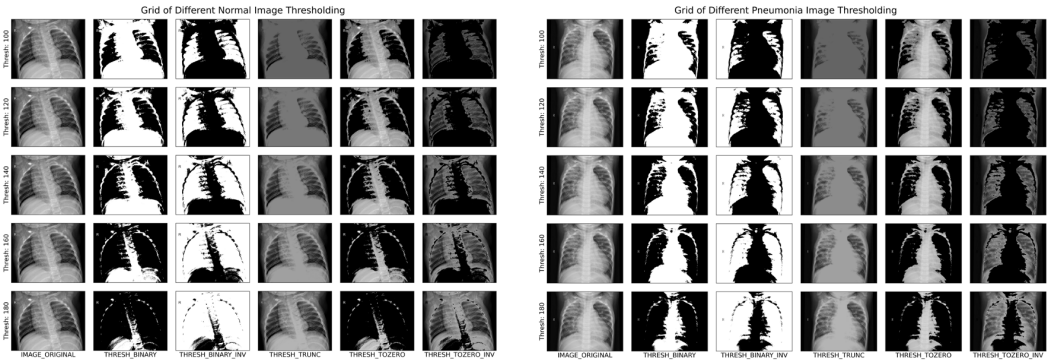

Figure 1: Two grids depicting normal chest X-ray images on the right and pneumonia chest X-ray images on the right. These X-rays are processed using different thresholding techniques to illustrate he effects of each method on the visualization of standard chest anatomy.

As depicted in Figure 1 illustrating normal lungs and pneumonia lungs, particularly in the binary thresholding image with a threshold value of 100, a noticeable asymmetry is evident in images indicating pneumonia, which is absent in those representing normal lungs. This dissimilarity tends to diminish with higher threshold values. Moreover, employing a small threshold value results in significant information loss, rendering the image devoid of meaningful content. Hence, determining an appropriate threshold technique and value is crucial to effectively convey information.

3.4 Performance Measures

In detecting pneumonia from chest X-rays using medical imaging models, we focus on several performance measures: accuracy, recall, precision, F1-score, confusion matrix, and Matthew's Correlation Coefficient. Of these, accuracy and recall are particularly critical due to the nature of medical diagnostics. High accuracy is vital as it shows the model's ability to correctly classify both pneumonia cases and healthy instances, crucial for avoiding misdiagnoses that could lead to unnecessary or missed treatments. High recall is equally important to ensure the model identifies almost all positive pneumonia cases, minimizing the risk of overlooking a diagnosis with severe potential consequences for patient care. These metrics are emphasized to ensure that models are both highly accurate and effective at detecting nearly all instances of pneumonia, considering the serious risks of the disease if undiagnosed and untreated.

Matthew's Correlation Coefficient (MCC) is a reliable metric for assessing binary classification tasks, ranging from -1 to +1, where each value indicates the quality of predictions. MCC uniquely evaluates all elements of the confusion matrix to provide a comprehensive measure of performance. The formula for MCC is:

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

This metric is particularly useful in medical scenarios where the prevalence of the disease is low compared to healthy cases, as it offers a deeper insight into a model's accuracy. MCC, along with other metrics, forms a robust framework for evaluating diagnostic models in chest X-ray imaging, ensuring their accuracy and reliability—essential for aiding clinical decision-making and enhancing patient care outcomes.

4 Model Implementation

4.1 Convolutional Neural Networks

To assess our results, we initially implemented the CNN model proposed by Stephen et al. [1] (Figure 2) to establish a baseline for performance evaluation. The choice to use their model was influenced by their methodology of merging the validation and test sets, rather than using a separate test set. This method risks overfitting to the validation set, which could lead to sub-optimal performance in real-world applications.

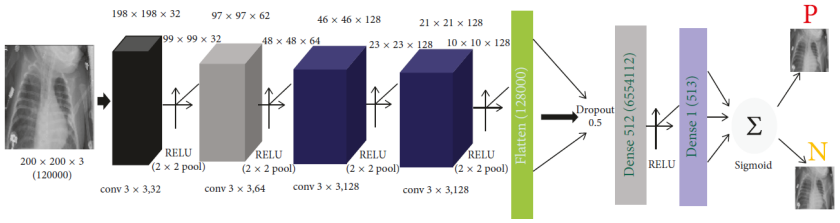

Figure 2: Proposed CNN architecture by Stephen et al. [1]

The CNN architecture consists of four convolutional layers, each paired with max-pooling layers, and uses Rectified Linear Unit (ReLU) activation functions in between. Each convolutional layer employs 3x3 filters: the first layer has 32 filters, the second 64, and the third and fourth layers each have 128 filters. The max-pooling layers operate with a 2x2 window. Post convolution, the data is flattened into a 1D feature vector and goes through a dropout layer with a 0.5 rate to prevent overfitting. This is followed by two dense layers, with a ReLU activation function between them, and the model outputs classification results through a Sigmoid function. It's important to note that the model proposed by Stephen et al. [1] was implemented in two distinct ways. First, it was trained and tested on a balanced version of the validation and test sets, each containing 320 images from both classes. Second, it was evaluated using the combined validation and test set as described in the original study. Contrary to the expected outcomes discussed in the paper, the model's performance in both scenarios was significantly lower than the results reported.

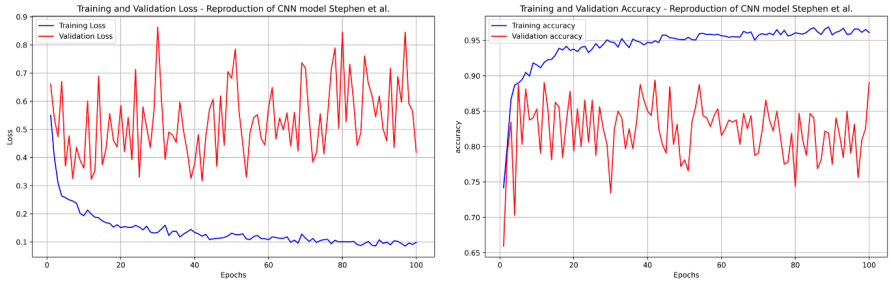

Figure 3: Accuracy and Loss functions from CNN model implemented using the architecture and parameters of Stephen et al. [1].

This performance gap is highlighted by the loss-epoch and accuracy-epoch curves from the first implementation, which used separate validation and test sets. The results (shown in Figure 4) differ notably from those claimed in the study, raising concerns about the reproducibility of the findings. Furthermore, as shown in the curves (refer to Figure 3), the model shows considerable variance and bias, suggesting that it has not yet reached a stable state of training. This indicates a clear need for extensive hyperparameter tuning to effectively address these issues and enhance the model's performance.

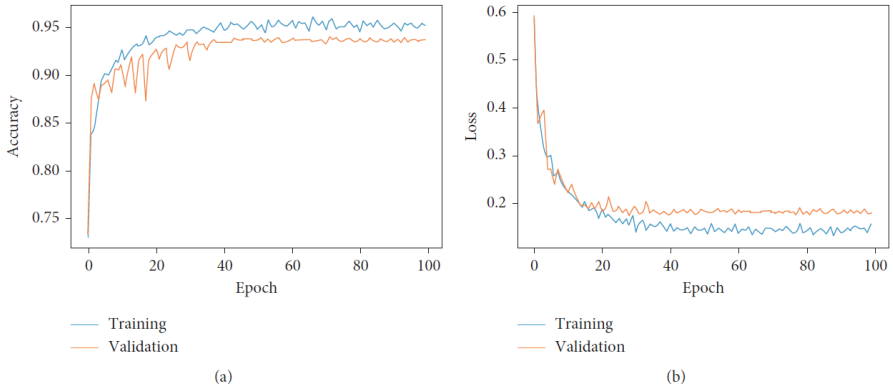

Figure 4: Accuracy and Loss functions present by Stephen et al. [1] using 200x200x3 images.

Consequently, we decided to use our reproduced results as a new benchmark for evaluating our model (Table 2), effectively setting aside the paper's reported performance metrics in favor of our findings from the updated dataset.

	Training		**Validation**	
Metrics	Presented in Paper	Reproduction	Presented in Paper	Reproduction
Accuracy	95.31%	93.80%	93.73%	90.30%
F1-Score	-	95.70%	-	92.50%
Recall	-	92.40%	-	95.50%
Precision	-	99.30%	-	89.60%
MCC	-	85.50%	-	79.30%

Table 2: Performance metrics comparison between original results presented in the paper and subsequent reproduction efforts for both training and validation phases, highlighting discrepancies in Accuracy, F1-Score, Recall, Precision, and MCC.

4.2 Vision Transformers

Transformers have become the benchmark in natural language processing (NLP) due to their innovative self-attention mechanisms. Their success extends beyond NLP into image classification and

object detection, setting new performance standards. A key feature of transformers, particularly noted in NLP, is their exceptional capability for transfer learning across various tasks. In image classification, the transformer architecture typically utilizes only the encoder portion of its design. To prepare inputs for this encoder, embeddings are generated from patches of the image. Positional encodings are then added to these patch embeddings to preserve the sequence of patches. After this setup, the positional encoded patch embeddings are processed through the encoder.

4.2.1 Transformer Encoder

In the Vision Transformer's encoder block, radiographic scans are segmented into 16x16 pixel patches. These patches are then encoded into a feature matrix, X, to which positional encodings are added. The introduction of positional encoding aims to maintain the spatial integrity of the radiographic images, ensuring that the model recognizes the original layout of the scans. To capture the interrelations among these patches, a self-attention mechanism is employed. This mechanism operates through three distinct embeddings: Query (Q), Key (K), and Value (V). These embeddings facilitate the model's ability to weigh the importance of different patches relative to one another, thereby enabling a nuanced understanding of the composite image.

$$Query(Q) = X \times W_q$$
$$Key(K) = X \times W_k$$
$$Value(V) = X \times W_v$$

The Vision Transformer (ViT) encoder utilizes matrices W_q, W_k, and W_v to project patch features into the embeddings for Query(Q), Key (K), and Value (V). It operates through self-attention, first by calculating the similarity between patch embeddings. This is done by computing the dot product of Q and K, enabling the encoder to understand the relationships between image patches effectively.

$$Q \times K^T$$

To ensure stability in the gradient during training, the similarity scores derived from the dot product of Q and K are adjusted by dividing them by the square root of their dimension. This adjustment mitigates the potential for large multiplication results to disrupt the learning process.

$$\frac{QK^T}{\sqrt{d_k}}$$

A softmax layer transforms the calculated similarity scores between Q and K into a probability distribution. This transformation enables the model to decisively focus on specific patches, enhancing its ability to distinguish relevant features within the image.

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

The step involving attention-based feature weighting is designed to adjust the importance of the chest image embeddings, V, utilizing the self-attention scores generated in the preceding step. This process ensures that the model prioritizes the most informative parts of the image for further analysis.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

As delineated in equation above, attention scores are computed through a process that includes matrix multiplication and concatenation. This procedure underscores the mathematical operations that underpin the self-attention mechanism within the Vision Transformer.

4.2.2 Multihead Attention

In the multi-headed attention framework, each "head" operates independently, capturing unique data aspects. This diversity allows for a more comprehensive representation. For multi-headed attention, the query, key, and value embeddings are split into N vectors, each processed by its own self-attention mechanism. Each head processes information independently, contributing to a holistic understanding. The outputs from all heads are then concatenated into one vector and passed through a final linear layer, enriching the model's ability to discern different features within the data.

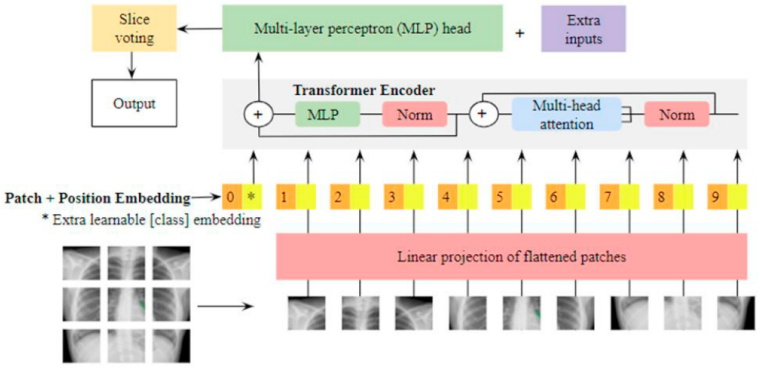

Figure 5: Architecture of ViT, implementation on chest x-rays (Uparkar et al. [15])

4.2.3 Transfer Learning Implementation

After unsatisfactory results with the CNN model, we shifted to using the Vision Transformer (ViT) model, specifically the vit-base-patch16-224-in21k variant developed by Dosovitskiy et al. [16]. This model, pre-trained on a large dataset, is known for its diverse visual representations and has shown excellent performance in various computer vision tasks, as highlighted by Henry et al. [17], Maurício et al. [18], and Uparkar et al. [15].

We adopted a conservative fine-tuning approach, freezing all layers except the output layer, which was fine-tuned to our diagnostic needs. This strategy maintains the robust features learned from ImageNet-21K while adapting the model to recognize pneumonia-specific features effectively. The output layer was fine-tuned over a limited number of epochs to prevent overfitting and improve generalization on new X-ray images. Additionally, considering the unique characteristics of pneumonia in X-ray images, we applied various thresholding techniques as detailed in section 3.3, enhancing feature visibility for accurate diagnosis. Our comprehensive approach tested 25 models using thresholded images versus models trained on original images, aiming to exploit the benefits of thresholding to better identify critical image attributes. This method blends the discriminative capabilities of vision transformers with specific adaptations for pneumonia detection.

5 Results

Table 3 depicts a thorough performance analysis comparing our conventional Vision Transformer (ViT) model with the ViT utilizing `TRUNC` thresholding at a value of 140 (ViT `TRUNC 140`), alongside our benchmark Convolutional Neural Network (CNN) model.

	Test Set		
Metrics	**CNN**	**Conventional ViT**	**ViT Thresholds**
Accuracy	88.80%	97.81%	98.12%
F1-Score	91.20%	100.00%	98.49%
Recall	94.00%	96.60%	98.49%
Precision	88.60%	98.27%	98.49%
MCC	75.90%	95.40%	98.01%

Table 3: Comparative Analysis of Machine Learning Model Performance

Performance metrics such as accuracy, F1 score, recall, precision, and the Matthews correlation coefficient were employed to gauge the effectiveness of each model. The training process of the conventional ViT, as shown in Figure 6, is fairly a steady learning process that is terminated at epoch 12. The best model is the model obtained in epoch 12. Moreover, for the ViT used with thresholding methods, the optimal model selected stems from epoch 16, exhibiting the lowest loss compared to others. It appears that the model tends to overfit beyond this epoch.

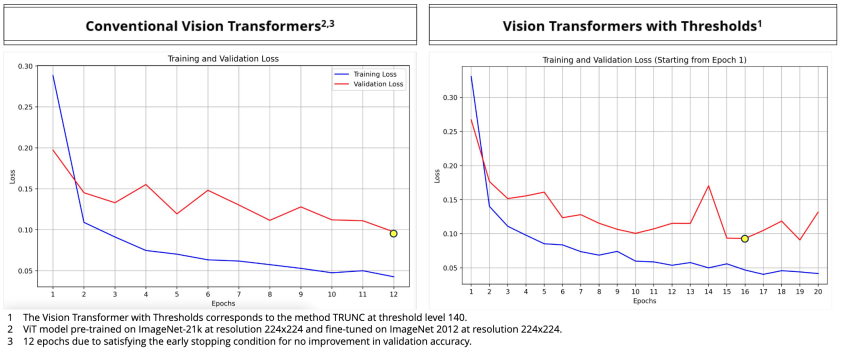

Figure 6: Training and validation loss trends for Conventional Vision Transformers (left) which stop at epoch 12 due to early stopping, and Vision Transformers with Thresholds (right) trained for 20 epochs, highlighting fluctuations with thresholding.

The ViT `TRUNC 140` model showed significant improvement across all metrics compared to both the standard ViT and the benchmark CNN models. This highlights the effectiveness of the thresholding preprocessing step in fine-tuning ViT models for detecting pneumonia in chest X-ray images. Figure 7 displays the confusion matrices for our CNN, conventional ViT, and ViT `TRUNC 140` models. These matrices underline the predictive accuracy of each model, with the ViT `TRUNC 140` model exhibiting the best performance. It recorded only three false positives and three false negatives, demonstrating a robust balance between sensitivity and specificity in detecting pneumonia from chest X-rays.

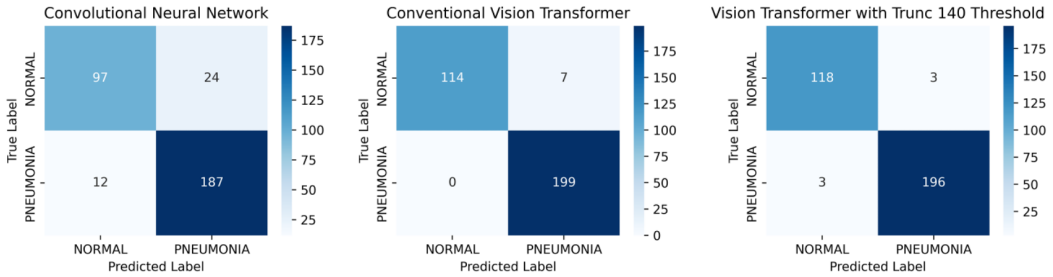

Figure 7: Confusion matrices comparing the performance of three different image classification models on pneumonia detection on the test set.

6 Conclusion and Future Work

In conclusion, our research successfully introduced and compared two models: a standard Vision Transformer (ViT) and an enhanced ViT with thresholding methods. Both models achieved our primary objective of surpassing the CNN benchmark set by Stephen et al. [1] in pneumonia detection from chest X-ray images. Implementing simple thresholding techniques within the ViT framework validated our hypothesis that performance could be further improved. This innovative approach not only proved effectiveness, but also reaffirmed the potential of thresholding enhancements in complex image analysis tasks. However, it is important to note that our model's performance could potentially have been further improved through meticulous hyperparameter tuning. This avenue remained partially unexplored due to constraints in computational resources, specifically the lack of GPU availability, which limited our ability to conduct extensive experiments. Additionally, exploring more sophisticated thresholding methods, such as Otsu's Binarization, could have contributed to even more significant improvements in our models' accuracy and robustness.

Future research into Convolutional Vision Transformers (CViTs) could greatly improve medical image analysis by combining CNNs' detailed processing with ViTs' comprehensive attention mechanisms. This could advance deep learning's role in diagnostics and expand its application to more medical conditions, potentially enhancing patient care, especially in resource-limited areas.

7 Appendix

7.1 Appendix A

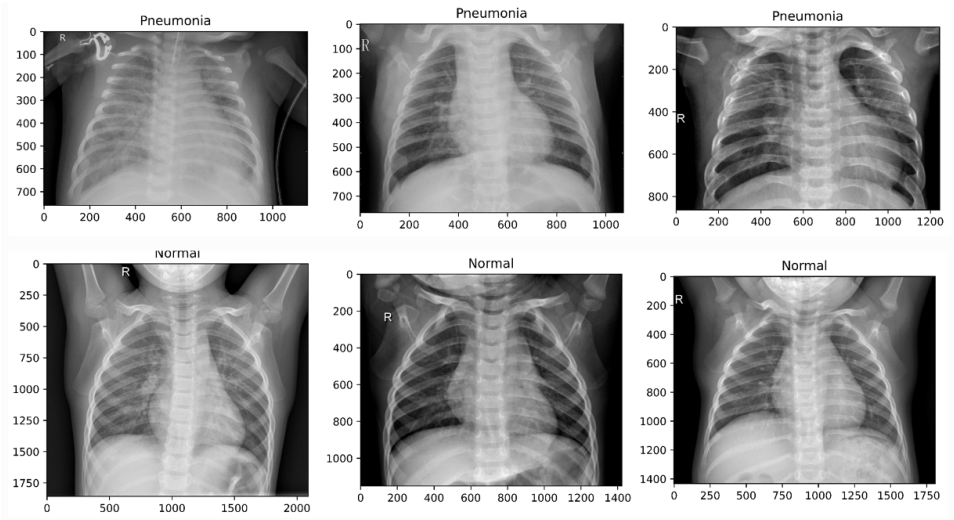

Figure 8: Series of chest X-ray images depicting various presentations of normal and pneumonia lungs.

7.2 Appendix B

The mathematical expressions for the thresholding methods applied are described below. `src` is a source 8-bit single-channel image and `dst` is the destination image of the same size and the same type as src.

1. `THRESH_BINARY`: This method sets all pixels in the original image 'src' that have intensities above a certain threshold 'thresh' to a maximum value 'maxval', making these pixels completely white. Pixels with intensities below the threshold are set to zero, turning them black.
$$\text{dst}(x,y) = \begin{cases} \text{maxval} & \text{if } \text{src}(x,y) > \text{thresh} \\ 0 & \text{otherwise} \end{cases}$$
2. `THRESH_BINARY_INV`: Inverse to the `THRESH_BINARY`, this method inverts the pixel intensities' assignment, offering a negative of the original image and highlighting different aspects of the X-ray.
$$\text{dst}(x,y) = \begin{cases} 0 & \text{if } \text{src}(x,y) > \text{thresh}, \\ \text{maxval} & \text{otherwise} \end{cases}$$
3. `THRESH_TRUNC`: This method caps the pixel values at the threshold. Pixels above the threshold are set to the threshold value itself, preserving the details up to a certain intensity without altering the lower intensities.
$$\text{dst}(x,y) = \begin{cases} \text{threshold} & \text{if } \text{src}(x,y) > \text{thresh}, \\ \text{src}(x,y) & \text{otherwise} \end{cases}$$
4. `THRESH_TOZERO`: It retains the pixel values that are above the threshold unchanged, while setting all others to zero, potentially enhancing the visibility of structures of interest.
$$\text{dst}(x,y) = \begin{cases} \text{src}(x,y) & \text{if } \text{src}(x,y) > \text{thresh}, \\ 0 & \text{otherwise} \end{cases}$$
5. `THRESH_TOZERO_INV`: Opposite to `THRESH_TOZERO`, this method keeps the lower-intensity pixels unchanged, providing a means to focus on variations in the image that may be subdued.
$$\text{dst}(x,y) = \begin{cases} 0 & \text{if } \text{src}(x,y) > \text{thresh}, \\ \text{src}(x,y) & \text{otherwise} \end{cases}$$

References

[1] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong. An efficient deep learning approach to pneumonia classification in healthcare, 2019.

[2] R. Siddiqi. Automated pneumonia diagnosis using a customized sequential convolutional neural network, 2019.

[3] D. Verma, C. Bose, N. Tufchi, K. Pant, V. Tripathi, and A. Thapliyal. An efficient framework for identification of tuberculosis and pneumonia in chest x-ray images using neural network, 2020.

[4] H. S. Omar and A. Babalik. Detection of pneumonia from x-ray images using convolutional neural network, 2019.

[5] . Bhatt, Y. Sudhanshusinh, and N. S. Jignesh. Convolutional neural network based chest x-ray image classification for pneumonia diagnosis, 2020.

[6] H. Wu, X. Pengjie, Z. Huiyi, L. Daiyi, and C. Ming. Predict pneumonia with chest x-ray images based on convolutional deep neural learning networks, 2020.

[7] R. Sarkar, H. Animesh, S. Koulick, and G. Preetam. A novel method for pneumonia diagnosis from chest x-ray images using deep residual learning with separable convolutional networks, 2020.

[8] S. Rajaraman, C. Sema, T. George, and A. Sameer. Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs, 2019.

[9] R. H. Abiyev and M. K. S. Ma'aitah. Deep convolutional neural networks for chest diseases detection, 2018.

[10] C. Matsoukas, J. Haslum, M. Soderberg, and K. Smith. Is it time to replace cnns with transformers for medical images?, 2021.

[11] S. Angara and S. Thirunagaru. Study of vision transformers for covid-19 detection from chest x-rays, 2023. URL `https://doi.org/10.48550/arXiv.2307.09402`.

[12] J. Li, J. Chen, Y. Tang, C. Wang, B. Landman, and S. Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives, 2022. URL `https://doi.org/10.1016/j.media.2022.102762`.

[13] M. Usman, T. Zia, and A. Tariq. Analyzing transfer learning of vision transformers for interpreting chest radiography, 2022. URL `https://doi.org/10.1007/s10278-022-00666-z`.

[14] Open Source Computer Vision. Miscellaneous image transformations. OpenCV, 2024. [Online]. Available: `https://docs.opencv.org/3.4/d7/d1b/group__imgproc__misc.html`.

[15] O. Uparkar, J. Bharti, R. K. Pateriya, R. K. Gupta, and A. Sharma. Vision transformer outperforms deep convolutional neural network-based model in classifying x-ray images, 2023.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[17] E. U. Henry, O. Emebob, and C. A. Omonhinmin. Vision transformers in medical imaging: A review, 2022.

[18] J. Maurício, I. Domingues, and J. Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review, 2023.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database, 2009.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need, 2017.