Multi-Window Vision Transformers (ViT) for X-ray Pneumonia Diagnosis

Abstract: Pneumonia remains a leading cause of child mortality worldwide (Rudan et al., 2008). Accurate diagnosis via chest X-rays is critical yet challenging. This study attempted to replicate and improve a high-performing Vision Transformer (ViT) model (Muchinguri, 2022) for paediatric pneumonia detection using transfer learning from ImageNet-21k. ViT models were fine-tuned on differently windowed images and combined via stacking ensemble classifiers. All trained models showed substantial overfitting, as they exhibit near-perfect training performance but weaker test outcomes (Accuracy of 0,875 / 0.857). Windowing as image preprocessing was hypothesised to improve model performance, but slightly worsened results. The ensemble approach did not mitigate overfitting, suggesting insufficient diversity among base models. Future research should employ adaptive windowing and additional regularisation strategies (dropout, MixUp, stochastic depth). Simpler model architectures such as CNNs may be better suited to explore the effects of windowing on pneumonia detection. Larger, balanced datasets may be needed to effectively utilise ViT capabilities for pneumonia diagnostics.

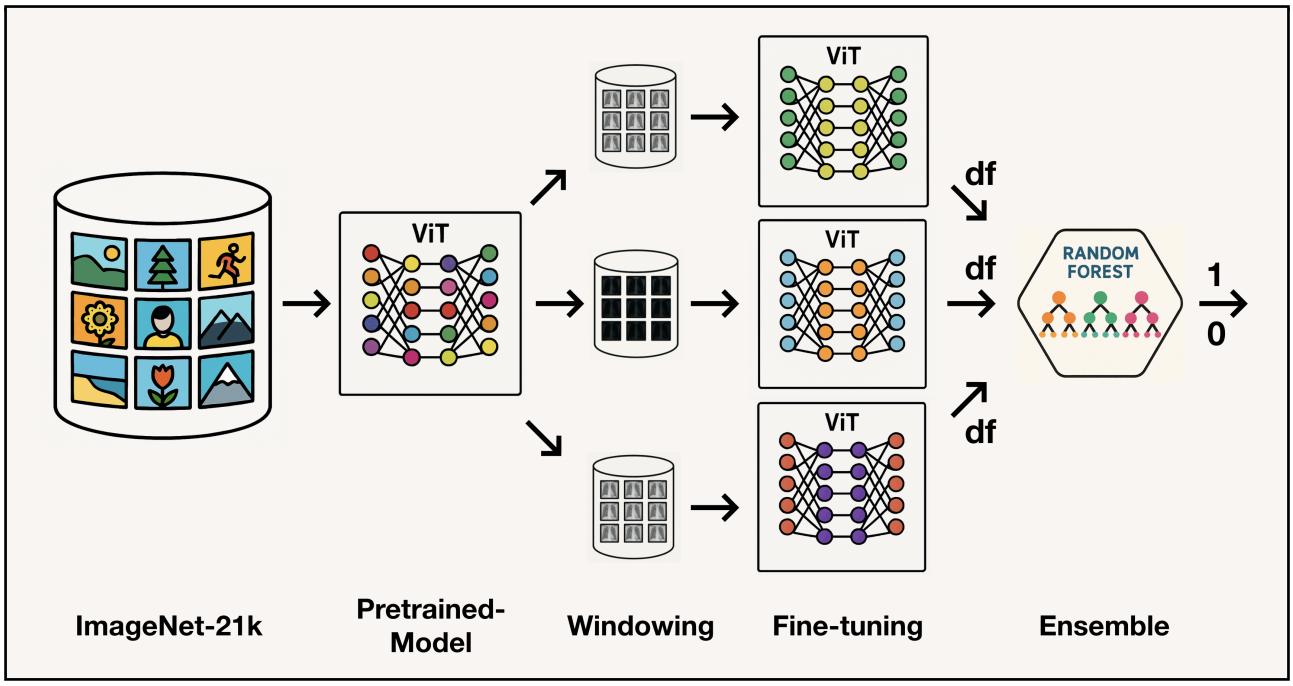

Figure 1. Simplified illustration of this study's transfer learning, data preprocessing, model training, and second stage classifier pipeline. Figure created with ChatGPT 4o (OpenAI, 2025).

1. Introduction

Pneumonia is the single leading cause of childhood mortality worldwide, as the World Health Organisation (WHO) estimates that pneumonia kills about 2 million children under the age of 5 every year (Rudan et al., 2008). Thereby the disease is killing more children than HIV/AIDS, malaria, and measles combined (Adegbola, 2012).

Chest X-rays are still the most effective method for diagnosing pneumonia (Metlay et al., 2007). However, detecting pneumonia in chest radiography can be difficult for radiologists. The appearance of pneumonia in X-ray images is often vague, can overlap with other diagnoses, and can mimic many other benign abnormalities. Accurate and instant pneumonia detection through machine learning (ML) could bring enormous value.

This paper attempts to replicate the results of a high-performing open source ML model for pneumonia detection, a fine-tuned Vision Transformer (ViT) model pre-trained on ImageNet-21k using transfer learning (Muchinguri, 2022; Deng et al., 2009; Wu et

al., 2020). Then this paper attempts to improve the model's performance by further pre-processing the used data set using different windowing parameters. Models trained on differently windowed data sets will then feed into a second stage classifier that makes ensemble predictions.

2. Literature Review

Initial attempts

Wang et al. (2017) introduced the ChestX-ray8 dataset with baseline deep CNNs for disease detection and localisation. They used SGD variants and achieved their highest accuracy for pneumonia detection (0.6333) using a ResNet-50 model architecture. ResNet stands for Residual Network, a deep convolutional neural network architecture (here with 50 layers).

Transfer learning

Transfer learning uses a model's learned weights on one data set as a useful starting point for fine-tuning on different data. Transfer learning is highly effective, especially for tasks with limited data (Weiss et al., 2016, Figure 2). CheXNet (Rajpurkar et al., 2017) employed a DenseNet-121 (121-layer convolutional neural network), pre-trained on ImageNet. From a chest x-ray the model outputs the probability of pneumonia as well as a heat-map of the areas in the image most indicative of pneumonia. The model exceeded radiologist-level accuracy in pneumonia detection (AUC: 0.7680). Kermany et al. (2018) collected and labeled 5,232 chest X-ray images from children, of which 3,883 had pneumonia. Then they adapted transfer learning (e.g., VGG-16) for paediatric pneumonia detection (Figure 2) with an SGD optimiser (LR: 0.001, batch size: 64, ≤100 epochs), and achieved over 92.8% accuracy, sensitivity of 93.2%, specificity of 90.1%, and area under ROC curve of 96.8%. This study uses their data set.

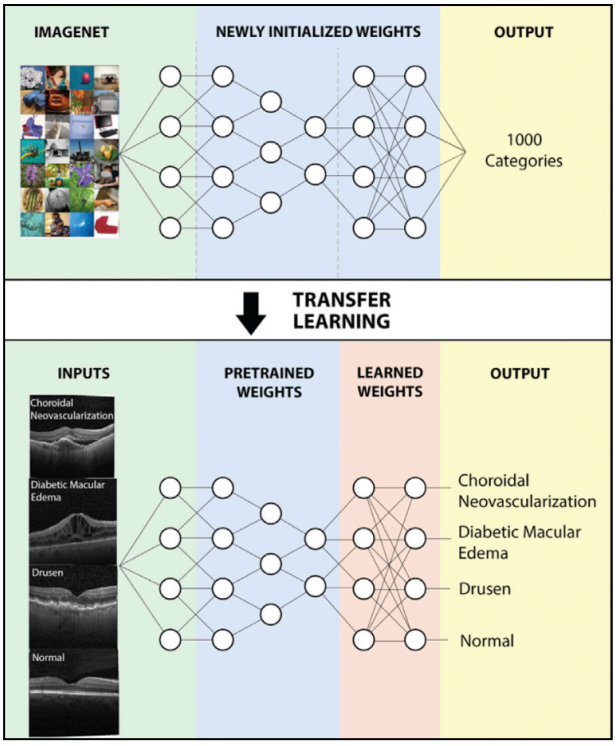

Figure 2. Transfer learning overview (taken from Kermany et al. (2018).

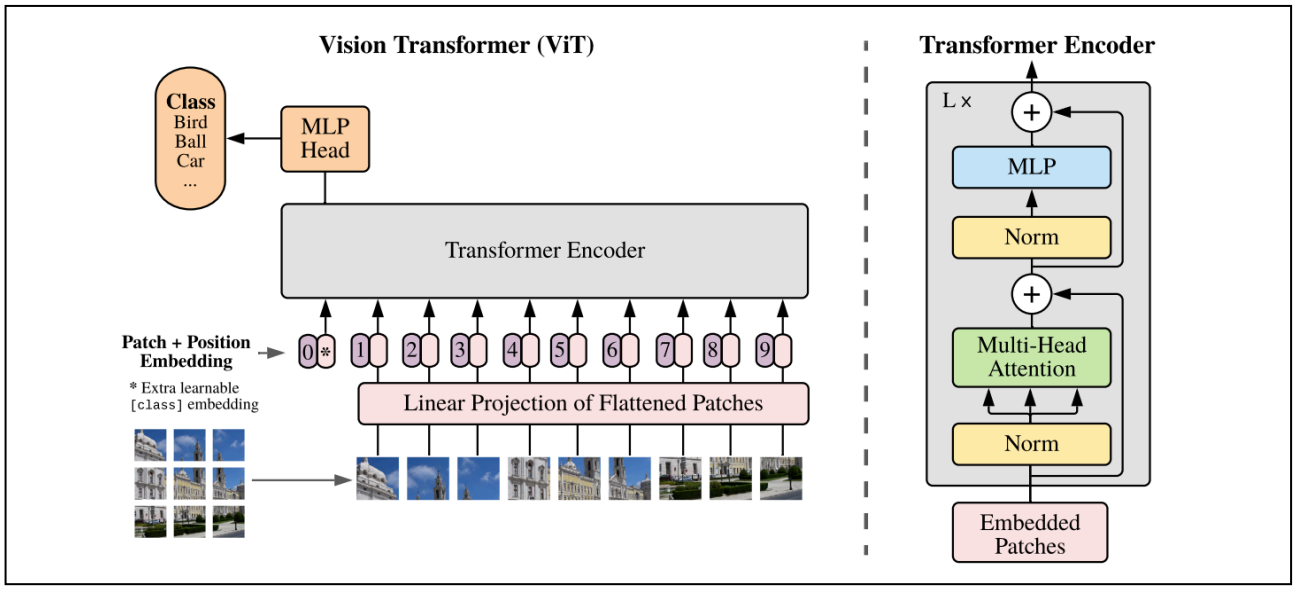

Figure 3. ViT Model overview (taken from Dosovitskiy et al. (2020)). An image is "split into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, the architecture adds an extra learnable "classification token" to the sequence."

Image augmentation
Manickam et al (2021) used a U-Net and image augmentation for pneumonia detection. A U-net is a type of convolutional neural network (CNN) designed for image segmentation tasks. For image augmentation Manickam et al used scaling, cropping, rotation, width change height change, padding, rotation, affine transformation, brightness, contrast, saturation, hue and filling operations on their training set. Thereby they more than tripled the number of images in their training data. They achieved an accuracy of 93.06%, precision of 88.97%, recall of 96.78%, and F1-score of 92.71%.

- Horizontal Flip Randomness =50%
- Vertical Flip Randomness =50%
- Horizontal Shift Up to 12 pixels
- Vertical Shift Up to 12 pixels
- Gaussian Blurring Up to σ=1

Windowing
Radiologists window chest x-ray scans to increase contrast for inspection. For deep learning models however, most chest x-rays are processed without any prior windowing. Wollek et al. (2023) demonstrated that windowing can strongly improve the chest x-ray classification performance of machine learning models. They used a DenseNet-121 pre- trained on ImageNet. They also extended the basic architecture with a windowing layer to learn the optimal windowing parameters. Their learnable windowing shows high improvements for the detection of several lung pathologies, and slight improvements in pneumonia detection (0.731 vs 0.75 AUC).

Vision Transformers (ViT)
A vision transformer (ViT) is a deep learning model that adapts the transformer architecture for language processing to computer vision. Instead of using convolutions (like CNNs), ViTs divide images into small patches and treat each patch as a token. They then process the tokens with a self-attention mechanisms to learn relationships across the entire image (Dosovitskiy et al., 2020; Figure 3). Despite their effectiveness, vision transformers are prone to overfitting due to their large number of parameters and lack of strong spatial inductive biases. Without large datasets, these models sometimes learn noise in the training data rather than generalisable features (Dosovitskiy et al., 2020).

Nicholas Muchinguri in 2022, however achieved very strong results on the relatively small Kermany et al., 2018 data set by fine-tuning a Vision Transformer (ViT) pre-trained on ImageNet-21k (Deng et al., 2009) (Link to open source mode: https://huggingface.co/nickmuchi/vit-finetuned-chest-xray-pneumonia). He achieved a loss of 0.1271 and an accuracy of 0.9551 on the test set. He only specified limited training parameters, and there is limited information about any data-preprocessing, which is why there is value in reproducing and properly documenting his results.

2. Methods

This study aims to incorporate the findings of Manickam et al (2021) on image augmentation, the findings of Wollek et al. (2023) on windowing, as well as the high-performing open-source model by Nicholas Muchinguri to improve pneumonia diagnosis even further. This study uses the paediatric pneumonia data set from Kermany et al. (2018), comprised of 5,232 labeled chest X-ray images from children. Though a ViT might not seem like the appropriate model type for a relatively small data set, Muchinguri's strong results warrant further exploration in this direction.

Data Exploration
First, this study, converted all images to grayscale with pixel intensities between 0 and 255. Then it concatenated all images with and without pneumonia, summed the frequencies of all pixel's intensities, and normalised as well as plotted the results (Figure 4.) The goal was to detect differences in the pixel intensity distribution between images with and without pneumonia. Then it is possible to window exactly the range of pixel intensities with the largest differences. This might exaggerate the differences between classes and improve classification. As can be seen in Figure 4, the largest difference is between 150 and 200 pixel intensities, which is why these parameters have been used for windowing later. Figure 5 shows example original and windowed (150, 200 and 30, 200) images.

The provided validation set only comprised 16 images. That is why an additional 20% of the training set were moved to the validation set, a common share of the training data. Images were grouped by person before moving, to avoid having the images from the same person in both the training and validation set.

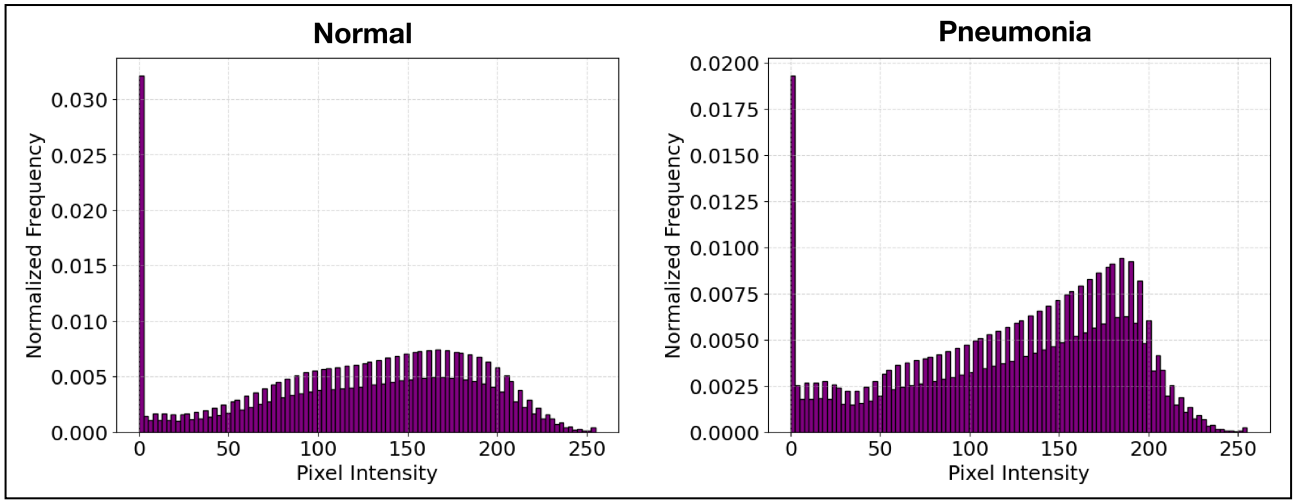

Figure 4. Pixel Intensities for normal vs pneumonia images

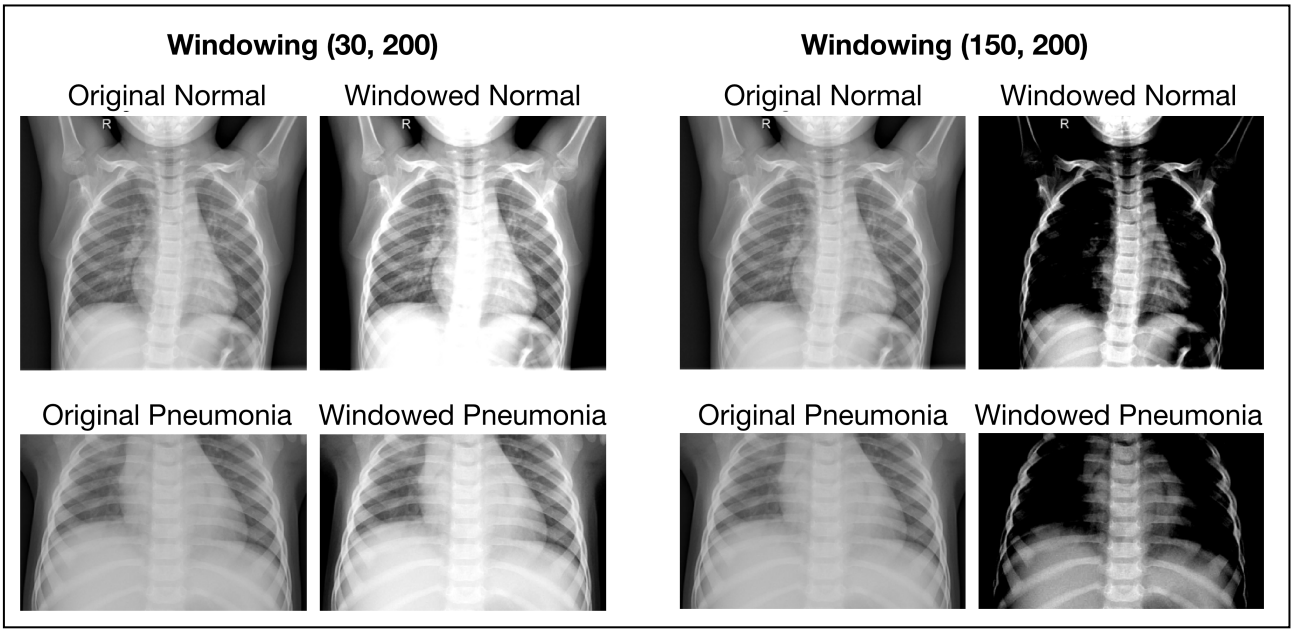

Figure 5. Original and windowed images with different windowing parameters for example

Table 1. Hyper-parameters for fine-tuning google/vit-base-patch16-224-in21k

Hyperparameters	
batch_size	16
num_workers	4
num_train_epochs	10
learning_rate	2E-05
optim	adamw_torch
adam_beta1, 2	0.9, 999
adam_epsilon	1E-08
lr_scheduler_type	linear
seed	42

Hyperparameters	
logging_steps	10
disable_tqdm	FALSE
num_train_epochs	10
eval_strategy	epoch
save_strategy	epoch
load_best_model_at_end	TRUE
metric_for_best_model	eval_loss
greater_is_better	FALSE

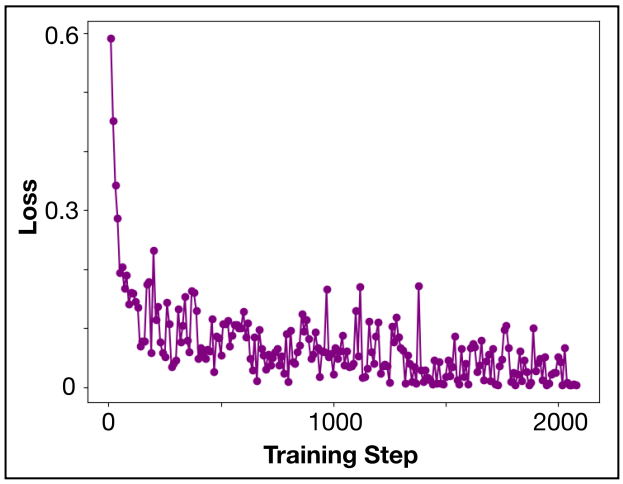

Figure 6. Training loss of fine-tuned model without windowing, with early stopping

Reproduction of Muchinguri's Model

Second, this study aims to validate and replicate Muchinguri's model performance. A Hugging Face (citation) pipeline for the model "vit-finetuned-chest-xray-pneumonia" from Nicholas Muchinguri was created using the python library transformers (citation). The xray_loader function loads in the data set (batch size of 16, num_workers=4). The pipeline was then evaluated on the test set to confirm Muchinguri's performance claims.

Using Muchinguri's training parameters (learning_rate: 2e-05, train_batch_size: 16, eval_batch_size: 8, seed: 42, optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08, lr_scheduler_type: linear, num_epochs: 10) this study then fine-tuned the ViT by Dosovitskiy et al. (2020), which was pre-trained on ImageNet-21k, with 14 million images and 21,000 classes, at a resolution of 224x224 pixels. Thereby it was utilising HuggingFace's ViT architecture (Wolf et al., 2020) and Dosovitskiy et al. (2020) pre-trained weights.

Images were resized to 224x224 pixels, optionally windowed (clamping pixel intensities to a specified range and rescaling), and converted to RGB. To improve performance this study also incorporated image augmentation with the same parameters as Manickam et al (2021), These included random horizontal and vertical flips (probability = 0.5 each), random affine translation up to 12 pixels (~5.4%), and Gaussian blur (sigma: 0.01–1.0). Data were loaded with PyTorch's ImageFolder and DataLoader classes (Paszke, 2019). In an attempt to avoid overfitting the training arguments were amended by early stopping (patience set to 3 epochs), and save best model. Model evaluation metrics included accuracy, precision, recall, F1 score, ROC-AUC, and confusion matrix, calculated after applying PyTorch Softmax probabilities to the raw logits. Training loss progression was visualised (Figure 6). All hyper parameters can be seen in Table 1. The model was trained on the University of St Andrews "cluster kennedy" High Performance Computer (containing 90 standard compute nodes, each containing 168 cores (two 2.25GHz 84-core AMD Broadwell (EPYC 9634 Zen 4 Genoa) processors), between 128GB and 1.5TB of memory and 1.7TB scratch disk. Two nodes are equipped with two NVidia Tesla V100 GPU processors each).

Ensemble model with windowed images

Third, this study attempts to improve pneumonia X-ray classification accuracy by training an ensemble of Vision Transformer (ViT) models on differently windowed images. Three ViT models were fine-tuned: (i) without intensity windowing (section above), (ii) with windowing between pixel intensities 150 and 200 (range of greatest differences), and (iii) with windowing between a conservative range of intensities 30 and 200. Each model was trained independently using the training methods described previously.

Continuous predictions as probabilities (between 0 and 1) of the positive pneumonia class were generated by each model separately on the training, validation, and test sets. It is very important, that these predictions are continuous rather than binary to capture the nuances of the different models trained on different windowing approaches. The continuous prediction probabilities from each model and target values were then saved into a Pandas DataFrame. This DataFrames then served as the training, validation, and test data sets for the second-stage stacking classifier. The study only examined the test set evaluations after the second-stage classifier was trained and tested. A second-stage classifier was chosen over simpler ensemble methods, because it captures nuanced interactions between base models and typically results in superior predictive performance (Wolpert, 1992).

Three classifiers, Logistic Regression, Support Vector Machine (SVM), and Random Forest, were trained on the aggregated predictions using the scikit-learn library (Pedregosa et al. 2001). The Random Forest classifier (n_estimators=50, max_depth=10,

min_samples_split = 5, random_state = 5220, class_weight = 'balanced') was selected as the final meta model and evaluated on the held-out test set. Performance measures include accuracy, precision, recall, F1 Score, confusion matrix, and ROC curve. If the ensemble model were to be employed, the manual work with DataFrames in this study would have to be automated into one pipeline.

Table 2. Evaluation metrics for 3 models on test set

Metric	Model by Muchinguri	Reproduced	Multi-Window Ensemble
Loss	0,182	0,494	N/A
Accuracy	0,953	0,875	0,857
Precision	0,959	0,836	0,816
Recall	0,966	0,995	0,997
F1	**0,963**	**0,909**	**0,897**
AUC	0,991	0,974	N/A
ROC	N/A	N/A	0,896

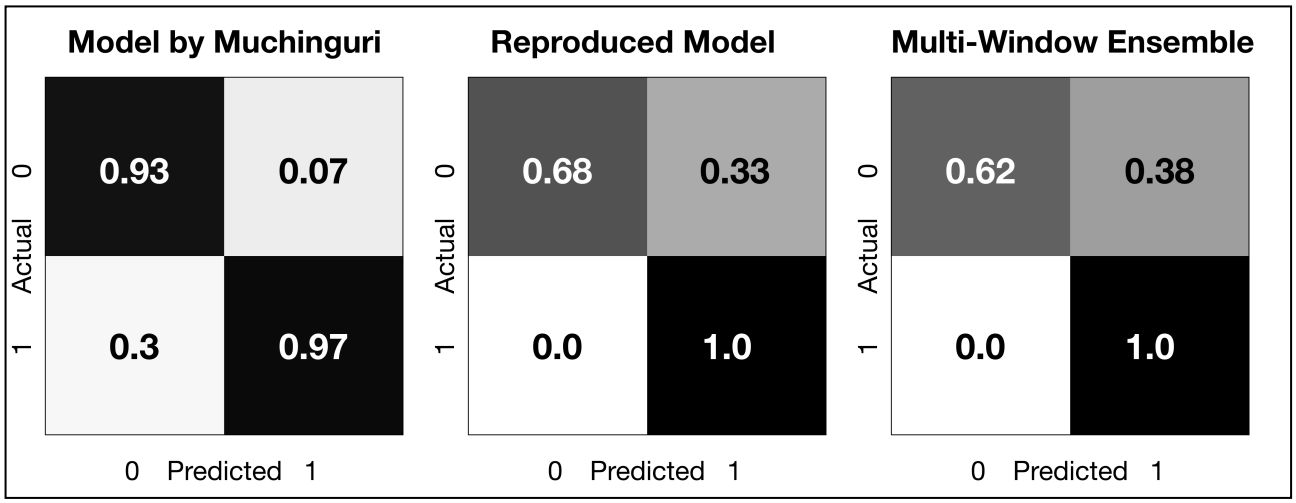

Figure 7. Confusion matrices for 3 models

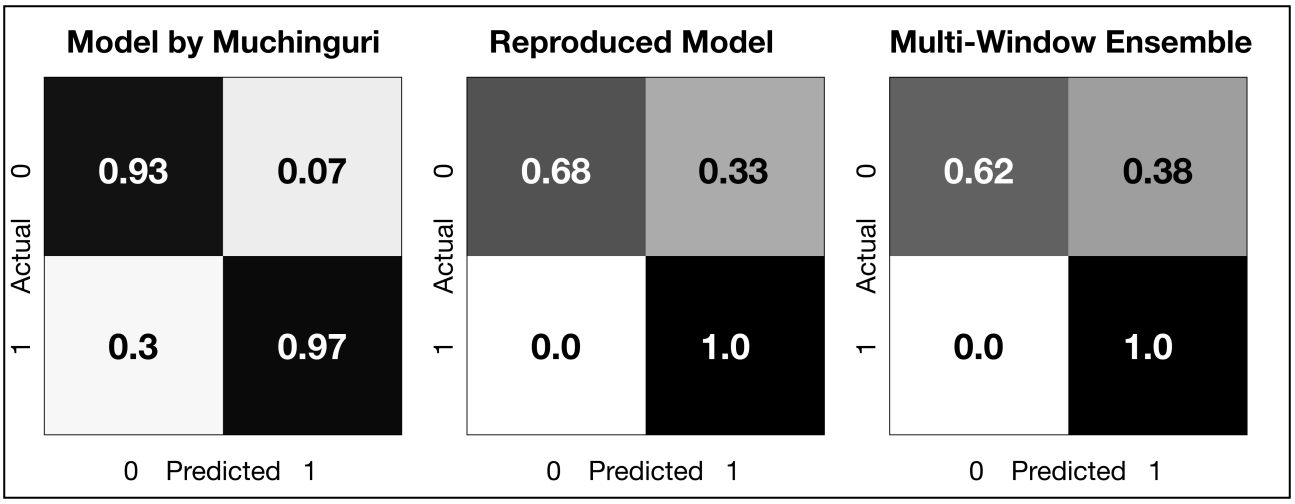

Figure 8. Confusion matrices for final ensemble model on train, validation and test sets

3. Results

Reproduction of Muchinguri's Model
Muchinguri's pneumonia detection model (https://huggingface.co/nickmuchi/vit-finetuned-chest-xray-pneumonia) achieved a similar accuracy as claimed, but a little higher loss on the test set (Table 2). See Figure 7 for Confusion Matrices. This study was not able to replicate these results by fine-tuning a model on the same base model as Muchinguri while using the same training hyper parameters. While the trained model achieved near perfect performance on the training and validation sets, it did not nearly as well perform on the test set, suggesting overfitting (Table 2). While achieving near perfect recall, the model produced a false positive rate of 32.4% (Figure 7, Figure 8).

Ensemble model with windowed images
Similar to the previous section, the two models trained on windowed images as well as the ensemble model performed near perfect on training and validation sets (Figure 8). The ensemble model however, and thus also presumably the base windowed models, performed worse on the test set, suggesting overfitting (Table 2, Figure 7, Figure 8). The ensemble model even performed slightly worse than the base model without windowing. This suggests an even worse performance by base models trained on windowed data as well as further overfitting.

4. Discussion

This study aimed to replicate and improve the performance of a Vision Transformer (ViT) model for pneumonia detection (Muchinguri, 2022) through image windowing and ensemble learning. While successfully confirming the accuracy of the original Muchinguri model, this study encountered challenges in replicating the strong results. An exact replication was difficult as Muchinguri does not provide extensive documentation and information on data pre-processing. The study's fine-tuned ViT model seemed to substantially overfit. This aligns with Vision Transformers large number of features as well as lack of inherent spatial inductive biases and resulting susceptibility to overfitting when trained on smaller datasets (Dosovitskiy et al., 2020). Despite not succeeding in its objective, this study's models still outperformed several past attempts at predicting pneumonia from chest x-rays (Wang et al., 2017; Rajpurkar et al., 2017).

Ineffectiveness of Windowing
Contrary to literature supporting windowing as a way to improve predictive performance on chest X-rays (Wollek et al., 2023), the models trained with windowed data exhibited even slightly greater evidence of overfitting. The ensemble model was designed to leverage complementary information from images with different windowing parameters. Surprisingly it performed slightly worse than the single, non-windowed ViT model. This suggests the chosen windowing parameters (30–200 and 150–200 pixel intensity ranges) did not sufficiently generalise discriminative features across the pneumonia classes. Despite very different windowing parameters, the base models were most likely not diverse enough (Wolpert, 1992). Future research should adaptively learn windowing parameters directly from the data (Wollek et al., 2023) rather than pre-selecting them.

Besides the data set's small size, this study was also limited by the imbalance of the Kermany et al. (2018) dataset. This is despite using class balancing strategies such as balanced class weights in Random Forest.

Avoiding Overfitting
Future research should utilise an array of advanced regularisation and training strategies to mitigate this study's overfitting issues. In additional to the here used standard data augmentation techniques (flipping, rotations), future research should utilise MixUp or CutMix (Zhang et al., 2018; Yun et al., 2019) to diversify expand and the training dataset.

Furthermore, future research should considerer different optimisation strategies involving learning rate scheduling with warm-up phases or gradual decay. Very importantly, regularisation methods including dropout (Srivastava et al., 2014), stochastic depth (Huang et al., 2016), label smoothing (Müller et al., 2019), and weight decay (Loshchilov et al., 2019) should also be incorporated to reduce model complexity and avoid overly confident predictions.

Finally, future research could repeat the study with these methodological refinements on a larger and more balanced datasets such as the ChestX-ray8 dataset (Wang et al., 2017).

Too Complex Model Architecture
Finally, the study's results suggest that the chosen model architecture may be too complex for the problem, research aim as well

as data size. The effects of windowing may have been better explored with simpler model architectures such as CNNs.

5. Conclusion

This study attempted to replicate and enhance a fine-tuned Vision Transformer (ViTs) for pneumonia detection from paediatric chest X-rays. While it successfully confirmed Muchinguri's (2022) reported accuracy, the study's fine-tuned ViT models exhibited substantial overfitting. This may be due ViT's inherent lack of spatial inductive biases as well as dataset limitations. The study's ensemble method using different image windowing parameters slightly worsened test set performance. This indicates limited base model diversity, and suggests that the windowing parameters selected based on pixel intensity distribution differences did not generalise well. Despite these limitations, the achieved accuracy surpasses several historical models.

Future research should employ additional regularisation methods such as dropout, MixUp, stochastic depth. It should also utilise learnable windowing parameters, optimised learning schedules, and test simpler model architectures such as CNNs. Lastly, it could train on a larger and more balanced data set to better explore the effects of windowing on machine learning pneumonia detection.

References

1. Kasban, H., El-Bendary, M. A. M., & Salama, D. H. (2015). A comparative study of medical imaging techniques. *International Journal of Information Science and Intelligent System*, *4*(2), 37-58

2. Dhawan, A. P., D'Alessandro, B., & Fu, X. (2010). Optical imaging modalities for biomedical applications. *IEEE Reviews in Biomedical Engineering*, *3*, 69-92.

3. Ritman, E. L. (2006). Medical x-ray imaging, current status and some future challenges. *Advances in X-ray analysis*, *49*, 1-12.

4. Spahn, M. (2013). X-ray detectors in medical imaging. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 731, 57-63.

5. M. S. Atkins and B. T. Mackiewich, "Fully Automatic Segmentation of the Brain in MRI", *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 98-107, 1998.

6. S. C. Strother, "Evaluating fMRI Preprocessing Pipelines", *IEEE Engineering in Medicine and Biology Magazine*, pp. 27-41, 2006.

7. Rudan, I., Boschi-Pinto, C., Biloglav, Z., Mulholland, K., & Campbell, H. (2008). Epidemiology and etiology of childhood pneumonia. *Bulletin of the world health organization*, *86*, 408-416B.

8. Adegbola, R. A. (2012). Childhood pneumonia as a global health priority and the strategic interest of the Bill & Melinda Gates Foundation. *Clinical infectious diseases*, *54*(suppl_2), S89-S92.

9. Metlay, J. P., Waterer, G. W., Long, A. C., Anzueto, A., Brozek, J., Crothers, K., ... & Whitney, C. G. (2019). Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. *American journal of respiratory and critical care medicine*, *200*(7), e45-e67.

10. OpenAI. (2025). *ChatGPT* 4o (Mar 10 version) [Large language model]. https://chat.openai.com/chat

11. Muchinguri, N. (2022). vit-finetuned-chest-xray-pneumonia. https://huggingface.co/nickmuchi/vit-finetuned-chest-xray-pneumonia

12. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE.

13. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision [Preprint]. arXiv. https://arxiv.org/abs/2006.03677

14. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017, July). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

15. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

16. Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 1-40.

17. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, *172*(5), 1122-1131.

18. Manickam, A., Jiang, J., Zhou, Y., Sagar, A., Soundrapandiyan, R., & Samuel, R. D.

J. (2021). Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures. *Measurement*, *184*, 109953.

19. Wollek, A., Hyska, S., Sabel, B., Ingrisch, M., & Lasser, T. (2023). WindowNet: Learnable Windows for Chest X-ray Classification. *Journal of Imaging*, *9*(12), 270.

20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

21. Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241-259.

22. Baltruschat, I.M., Nickisch, H., Grass, M. et al. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Sci Rep 9, 6381 (2019). https://doi.org/10.1038/s41598-019-42294-8

23. Paszke, A. (2019). Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

24. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020, October). Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.emnlp-demos.6

25. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

26. Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep networks with stochastic depth. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14 (pp. 646-661). Springer International Publishing.

27. Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, *32*.

28. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

30. Bradski, G. (2000). The OpenCV library. Dr. Dobb's Journal of Software Tools, 25(11), 120–123.