ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

CLASSIFICATION OF LUNG DISEASES IN X-RAY IMAGES USING TRANSFORMER-BASED DEEP LEARNING MODELS

Nyoman Sarasuartha Mahajaya[1], Putu Desiana Wulaning Ayu[2], Roy Rudolf Huizen[3]

[1,2,3]Departement of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia

email: 222012002@stikom-bali.ac.id[1], wulaning_ayu@stikom-bali.ac.id[2], roy@stikom-bali.ac.id[3]

Abstract

This research evaluates the performance of two Transformer models, the Vision Transformer (ViT) and Swin Transformer, in the analysis of thoracic X-ray images. The study's objective is to determine whether Transformer models can enhance diagnostic accuracy for lung diseases, considering challenges such as early symptom variability and similar radiological signs. The dataset includes 21,165 X-ray images, featuring 3,616 COVID-19 cases, 10,192 normal images, 6,012 images of Lung Opacity, and 1,345 pneumonia images. Model development involved tuning hyperparameters such as epoch numbers and optimizer choice. The results indicate that using the AdamW and Adamax optimizers achieves an optimal balance between computational efficiency and accuracy. The Swin Transformer model, using the Adamax optimizer, reached the highest testing accuracy of 96.10% in 33,802.70 seconds, while the Vision Transformer achieved a testing accuracy of 95.10% in 33,503.10 seconds.

Keywords : vision transformer, swin transformer, classification, deep learning, chest x-ray

Received: 27-06-2024 | **Revised:** 30-07-2024 | **Accepted:** 05-08-2024
DOI: https://doi.org/10.23887/janapati.v13i3.81425

INTRODUCTION

Lung diseases are a major global health issue, encompassing conditions such as pneumonia, lung opacity, chronic bronchitis, emphysema, lung cancer, and chronic obstructive pulmonary disease (COPD) [1], [2]. Identifying and diagnosing lung diseases often presents several challenges. The primary difficulties faced by physicians and paramedics stem from the variation in symptoms and clinical signs, which are frequently nonspecific, especially in the early stages of the disease [3]. Some pulmonary conditions, including interstitial fibrosis, exhibit similar radiological signs, complicating accurate diagnosis based on thoracic X-ray images alone. For instance, both pneumonia and COVID-19 can show infiltrate signs on X-ray images, but require different diagnostic approaches and management [4].

One solution to these challenges is the use of artificial intelligence technology, particularly image processing with deep learning techniques such as convolutional neural networks (CNNs). By leveraging the sophistication of these algorithms, more accurate and rapid analysis of X-ray images can be achieved, enabling more efficient detection and classification of lung diseases [5].

Previous research in the detection and classification of lung diseases using medical imaging has shown significant advancements. Many studies have employed deep learning techniques, notably convolutional neural networks (CNNs), with one highlighting progress in using deep learning algorithms for early disease detection from image data [6]. However, despite their benefits, some studies also revealed that CNNs have architectural limitations, such as limited vision due to their sliding window algorithm, which hinders the ability to view an entire image at once [7].

In 2020, a Google scientist named Dosovitskiy et al. [8] introduced an artificial intelligence model called Vision Transformer (ViT), followed by Liu et al. [9] with the Swin Transformer, aimed at addressing these CNN issues. Several studies have successfully applied these AI models to medical images such as computed tomography (CT) scans and X-ray images, claiming performance improvements in their detection processes with Transformer models using TransUnet and SwinUnet algorithms [10]. Other research has also claimed increased accuracy with Transformer models for MRI images, specifically for osteosarcoma [11]. In the field of remote sensing, the use of Swin Transformer enhanced with a Dynamic High-Pass Preservation module showed significant improvements in pansharpening techniques, producing images with finer details and more

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

complete spectral information [12]. Research Peng et al [13] introduced modifications to the Swin Transformer that focus on spatial feature extraction for spectral analysis. This model successfully improved accuracy in object classification tasks by leveraging spectral and spatial information from images. Further research using MRI data demonstrated that the Swin Transformer could speed up MRI scanning processes. This technique reduces processing time without compromising image quality, offering an efficient solution to the performance time issues of traditional MRI [14].

For autonomous control, the use of the Vision Transformer in autonomous vehicle systems has successfully processed and visually understood the surrounding environment, implementing functions such as navigation and obstacle avoidance with high accuracy [15]. In medical image analysis, this technology aids in more effectively identifying and segmenting affected areas, demonstrating improved accuracy in disease classification [16]. The Vision Transformer also effectively reduces the time and resources needed for diagnosing COVID-19, with significantly improved accuracy compared to traditional CNN models [17]. Gender classification based on facial images using Vision Transformer shows good robustness against variations in pose and facial expressions [18].

Given the substantial potential of Transformer models in medical image processing, this study aims to conduct an in-depth evaluation of the capabilities of two Transformer models, Vision Transformer and Swin Transformer, in classifying lung diseases via X-ray images. This research contributes by evaluating these two Transformer models for classifying lung diseases from X-ray images. The research stages include: data preprocessing to standardize images to consistent dimensions, followed by model training processes using various hyperparameter settings such as number of epochs and the use of optimizers like Adam, AdamW, Adamax, Nadam, Lamb, RMSprop, or SGD to explore their impact on model effectiveness in enhancing classification accuracy. After the training process, performance evaluation is conducted by measuring metrics such as accuracy, precision, recall, and F1-score, and conducting a comparative analysis of computational efficiency between both models to measure and compare the computational time of each model. This ensures that the assessment of both models is not only from the diagnostic accuracy perspective but also in terms of operational efficiency. The results of this study demonstrate the significant potential of Transformer models in enhancing the diagnostic accuracy of lung diseases, offering new insights into the application of artificial intelligence technology in the medical field.

METHOD

This research flow is designed to analyze the lung disease classification process using the Vision Transformer and Swin Transformer models, as depicted in Figure 1. The study adopts a structured methodology to evaluate the effectiveness of both models and also analyzes the efficacy of various optimizers in enhancing the accuracy of these models in classifying lung diseases using X-ray images. The stages involved include data preprocessing, model performance evaluation using metrics such as accuracy, precision, recall, and F1-score, and considering the training time efficiency.

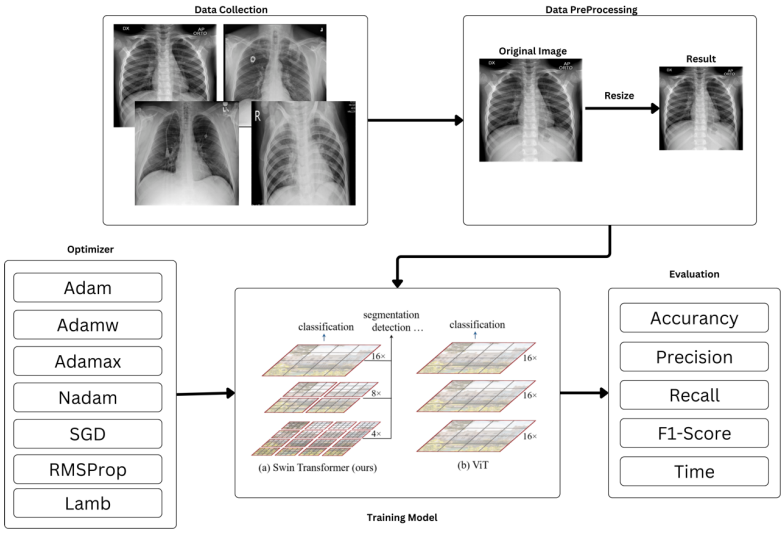

Figure 1. Research Flow

A. Dataset

This study utilizes a dataset from Kaggle [19] comprising a total of 21,165 chest X-ray images, including 3,616 COVID-19 positive case images, 10,192 normal images, 6,012 images of Lung Opacity, and 1,345 pneumonia images. As depicted in Figure 2, there are four panels each displaying chest X-ray images with different labels: Covid, Normal, Lung Opacity, and Pneumonia. Images labeled Covid illustrate typical features of COVID-19 infection, including the presence of hazy areas known as ground-glass opacities, which usually appear in both lungs with an irregular pattern. The panel labeled Normal shows a radiograph of the lungs without signs of opacity or disease, with clear lung structures and normal vascular patterns, free of any hazy or dense areas. The panel labeled Lung Opacity displays dark areas or opacities within the lungs, indicating the presence of denser material than air, such as fluid, cells, or tissue.

These conditions can be caused by various factors, including infection, inflammation, or cancer. Meanwhile, images labeled with Pneumonia display whiter or denser areas, indicating inflammation in the lungs, often characteristic of pneumonia caused by the accumulation of fluid or pus in the lung alveoli.

B. Pre-Processing Data

This process involves resizing the images to a standard size of 224 x 224 pixels to ensure uniform input into the transformer models, adapting methods proposed by Alexey Dosovitskiy et al. [8] and Ze Liu et al.

This resizing aims to standardize the image dimensions used, creating a consistent format that aligns with the tested parameters in transformer model research.

All preprocessed data is then divided into two sets: a training set and a testing set. The training set is used to train the Vision Transformer and Swin Transformer models, while the testing set is used to evaluate the models' performance in classifying lung diseases during the training process. The data split is randomly done with a 70% proportion for training and 30% for testing. We did not create a separate validation set, instead, we utilized the testing set as an evaluation set to validate the model's performance iteratively. This approach ensures that the models' accuracy is measured on data that they have not seen during training, providing a robust evaluation of their performance.

C. Vision Transformer

Vision Transformer (ViT) is a deep learning model that processes images by dividing them into fixed-size sections, which are then treated as tokens for analysis. The structure of the Vision Transformer consists of three main parts: patch and position embedding, transformer encoder, and classification head. Initially, the incoming image is split into several patches.

As shown in Figure 3, the image is divided into small, equally sized pieces. Each patch is transformed into a vector through a flattening process and is given a linear projection to convert it into a format suitable for integration into the transformer encoder.

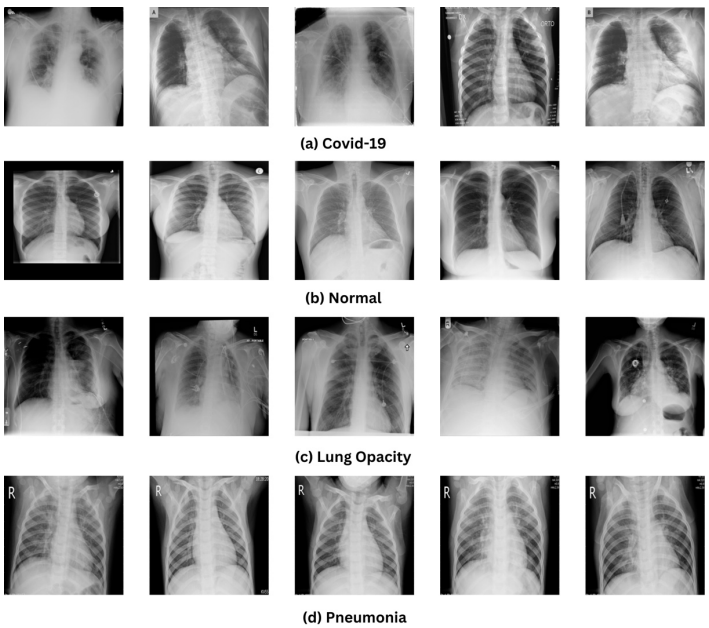

Figure 2. Example of a lung x-ray image

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

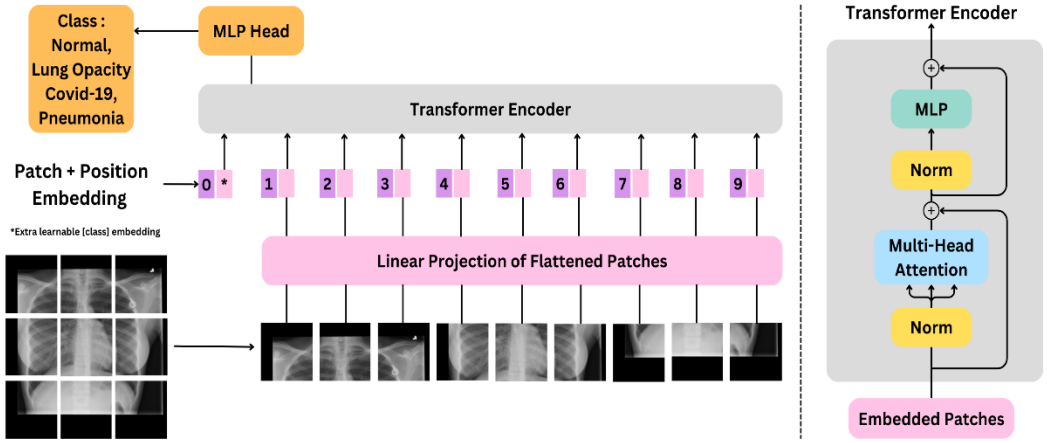

Figure 3. Vision Transformer architecture - The illustration inspired by [8]

Each patch vector is also given positional encoding to inform the model about the relative position of each patch within the image. An additional class token is included to serve as a placeholder for the global representation of the image, which is later used for classification purposes.

The next step is feature extraction, performed by the component called the encoder in the transformer. This encoder consists of several blocks, each with two main parts: Multi-Head Self-Attention and MLP (Multi-Layer Perceptron). Multi-Head Self-Attention allows the model to simultaneously observe and compare all image patches, aiding the model in understanding the relationships between different parts of the image and identifying important features spread throughout the image. The second part, the MLP, is a neural network that enriches the model's processing by adding complexity to the information obtained from the attention mechanism. Each block in the encoder is also equipped with a normalization layer, which helps to stabilize the model's learning by normalizing the output distribution.

After the encoder processes all the patches, the extracted information is further processed by the MLP Head. This MLP Head consists of one or more dense neural network layers that utilize non-linear functions to transform the information into a global image representation, based on the classification token. This global representation is then used to generate the final prediction in the form of a probability distribution that indicates the likelihood of the image belonging to a predefined category during training. This prediction is key to the model's final classification decision, determining which category is most likely to match the object in the image based on previous learning.

This allows the model to capture relationships between different parts of the image through self-attention. The use of self-attention enables the Vision Transformer to effectively extract image features and perform tasks such as image classification, object detection, and semantic segmentation[16]–[18], [20].

D. Swin Transformer

The Swin Transformer is a neural network model that leverages transformer technology for image analysis, focusing on understanding the hierarchical relationships between different parts of the image. This model is more efficient than traditional transformers because it adopts a shifted window technique that reduces computational complexity. This approach allows it to dynamically adjust the focus on various parts of the image, enhancing its ability to process large images more efficiently while maintaining high accuracy in tasks such as object detection and semantic segmentation.

The Swin Transformer architecture is divided into three main parts. First, the image is broken down into small blocks for easier analysis. Each stage of the architecture increases the number of channels while reducing spatial resolution, allowing the model to extract features at various levels of abstraction. Second, this process is carried out through several stages that progressively increase the complexity of the information processed.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

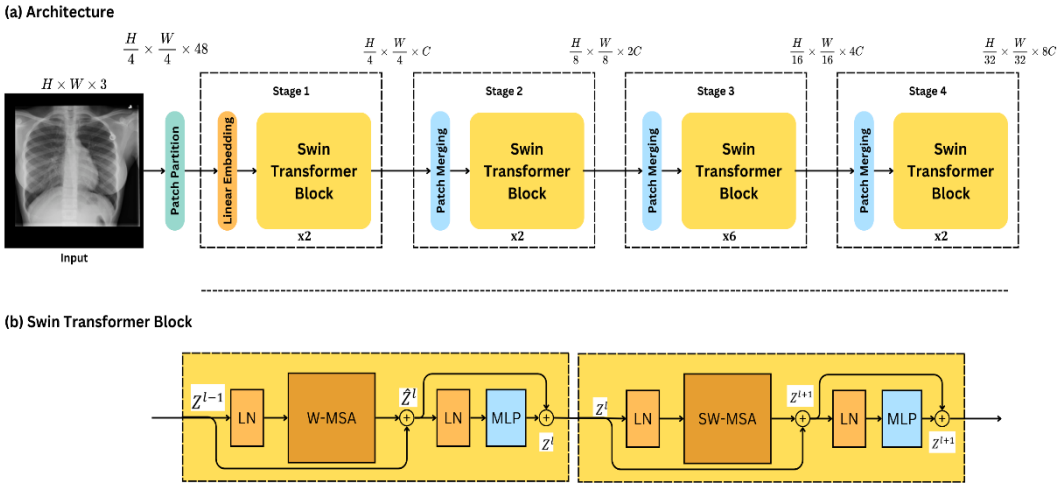

Figure 4. Swin Transformer architecture - The illustration inspired by [9]

Third, two consecutive Swin Transformer blocks utilize Multi-Layer Perceptrons, Layer Normalization, and Multi-head Self-Attention (including the Shifted Window variant) to effectively process information. This approach enables a deeper and more holistic understanding of the image, making the Swin Transformer suitable for various applications in computer vision, including classification, object detection, and semantic segmentation. The Swin Transformer achieves sophisticated results that surpass previous models such as ViT, DeiT, and ResNe(X)t in terms of accuracy while maintaining similar computational efficiency. Its performance in tasks like COCO object detection, ADE20K semantic segmentation, and ImageNet-1K image classification demonstrates its effectiveness in handling complex visual tasks. The success of the Swin Transformer highlights the potential of integrated modeling for computer vision and language, which could be beneficial for both computer vision and natural language processing domains[12], [13], [21].

E. Hyperparameter Tuning

Selecting the right parameters for a developed model is crucial to improving classification performance. Hyperparameter tuning plays a critical role in machine learning and deep learning, as the chosen hyperparameters greatly influence the model's effectiveness. In this study, hyperparameter tuning methods are employed to determine the parameters of the optimizer used. An optimizer is an algorithm or method in artificial intelligence that plays a crucial role in adjusting parameters such as weights and biases, aiming to reduce the loss function or enhance production efficiency. This facilitates changes in weight values and adjusts the learning rate in neural networks so that losses can be minimized[22]. The parameters to be compared during hyperparameter tuning include Adam, AdamW, Adamax, Nadam, SGD, RMSprop, and Lamb.

F. Model Evaluation

To evaluate the Vision Transformer and Swin Transformer models, performance parameters such as Accuracy, Recall, Precision, and F1-score are determined. These parameters are calculated using a confusion matrix created for each model. Accuracy is calculated to determine the percentage of correct predictions. Precision is calculated to determine the probability of positive classification. Specificity determines the percentage of negative classifications correctly predicted from all parameters. Unlike specificity, recall determines the percentage of positive classes correctly predicted. The F1-score is used to determine the balance between specificity and recall. The parameters are expressed in the following equations:

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

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

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

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

$$F1-Score = 2\left(\frac{Precision\ x\ Recall}{Precision+Recall}\right) \quad (5)$$

RESULT AND DISCUSSION

This study utilizes four testing scenarios, which are as follows:

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

A. Base Model Result

In this first scenario, testing was conducted on the proposed model as shown in Figure 5 and Figure 6. The testing was performed using the Adam optimizer with a learning rate of 0.0001, 10 epochs, and a batch size of 32. In this first scenario, during the training process, the Vision Transformer model achieved a training accuracy of 0.991 and a testing accuracy of 0.942, while the Swin Transformer model achieved a training accuracy of 0.985 and a testing accuracy of 0.952.

The graph above shows the performance of the model trained using the Vision Transformer architecture with the Adam optimizer, evaluating metrics from epoch 1 to 10. It can be observed that the train accuracy consistently increases, reaching a high value by epoch 10, indicating that the model is learning effectively from the training data. In contrast, testing accuracy peaks at epoch 2 before experiencing a significant drop at epoch 6 and then stabilizing. This may indicate variations in the testing data the model encounters or potential overfitting as the number of epochs increases.

Metrics for Precision, Recall, and F1-Score show a similar trend to the testing accuracy, achieving higher values in the early epochs and then declining at epoch 6 before stabilizing again. This decline might be due to variations in more complex data or anomalies that the model encounters in the validation batches.

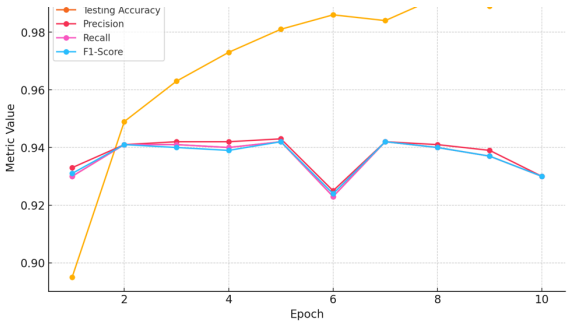

Figure 5. Vision Transformer with Base Model Result

Figure 6 presents the initial results from testing the Swin Transformer model. Train accuracy shows consistent improvement, indicating that the model continues to learn from the training data more effectively with each epoch. At the same time, testing accuracy peaks at Epoch 3 with a value of approximately 0.952 before experiencing a slight decline and then stabilizing around 0.95. This suggests potential overfitting, where the model is very well-optimized on the training data but less generalized to new data. Precision, Recall, and F1-Score, which measure the model's accuracy and success in correctly classifying data, follow a similar trend, reaching their highest values at Epoch 3 and showing slight variations in subsequent epochs. The consistency between Precision, Recall, and F1-Score highlights that the model demonstrates stable performance across various evaluation aspects.

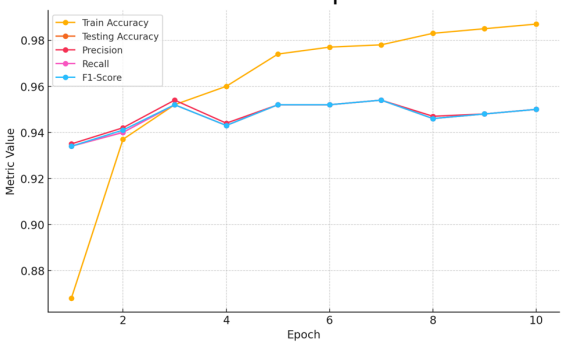

Figure 6. Swin Transformer with Base Model Results

B. Testing Results with Hyperparameter Tuning

In the second scenario, parameter optimization was performed on the proposed model using hyperparameter tuning. The model training was conducted with a learning rate of 0.0001 and a batch size of 32.

Table 1 shows the performance testing results of the Vision Transformer (ViT) architecture-based model using different optimizers over 50 epochs.

Table 1. Vision Transformer Model Hyperparameter Results

Optimizer	Max Train Acc	Max Testing Acc	Max Precision	Max Recall	Max F1-Score	Time (second)
Adam	0.999	0.947	0.9	0.947	0.947	32635.58
Adamw	0.999	0.948	0.9	0.948	0.947	33072.99
Adamax	1	0.951	0.9	0.951	0.951	33503.10
Nadam	0.998	0.948	0.9	0.948	0.948	33084.32
SGD	0.925	0.905	0.9	0.905	0.905	33514.04
RMSProp	0.997	0.942	0.9	0.942	0.942	33827.18
Lamb	1	0.945	0.9	0.945	0.945	34496.97

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

The results of the Vision Transformer (ViT) model, tested with various optimizers over 50 epochs, showed that nearly all optimizers achieved very high training accuracy, with some even reaching a perfect score (1.000). This indicates the optimizers' effectiveness in learning from the training data. However, this also suggests potential overfitting, especially for optimizers that achieved perfect values. In terms of testing accuracy, the model with the Adamax optimizer stood out, achieving the highest value (0.951), indicating superior ability in generalizing learning to new data.

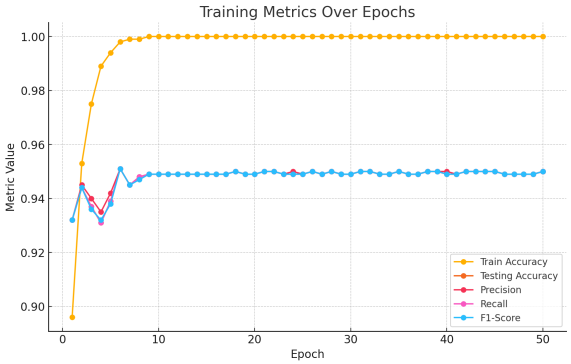

Figure 7. Vision Transformer Model Results with Adamax Optimizer

The choice of optimizer not only affects the model's performance in terms of accuracy and other metrics but can also impact computational time efficiency, with Table 1 showing that the Adam optimizer had the most efficient total computational time compared to other optimizers.

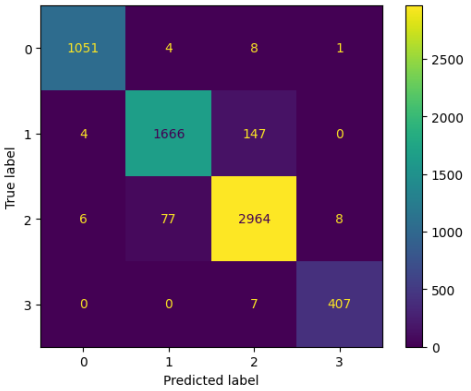

Figure 8. Confusion Matrix of Vision Transformer Model with Adamax Optimizer

Overall, Precision and Recall move in alignment with testing accuracy, confirming that the model not only correctly classifies most classes but also effectively identifies relevant positive cases. From this analysis, it is evident that the choice of optimizer significantly impacts the model's ability to avoid overfitting while maintaining high generalization capability.

Table 2 shows the performance testing results of the Swin Transformer architecture-based model using different optimizers.

Table 2. Swin Transformer Model Hyperparameter Results

Optimizer	Max Train Acc	Max Testing Acc	Max Precision	Max Recall	Max F1-Score	Time (second)
Adam	0.998	0.956	0.957	0.956	0.956	35532.04
Adamw	0.998	0.955	0.956	0.955	0.955	32766.43
Adamax	1	0.961	0.961	0.961	0.961	33802.70
Nadam	0.997	0.954	0.955	0.954	0.954	34514.58
SGD	0.927	0.929	0.929	0.929	0.928	33592.71
RMSProp	0.997	0.950	0.951	0.950	0.950	34108.11
Lamb	0.999	0.954	0.954	0.954	0.953	36039.29

According to Table 2, it can be concluded that the Adamax optimizer is a highly effective option for achieving a balance in metric performance. The Swin-224 model with the Adamax optimizer demonstrated outstanding performance during training, as seen in the graphs of Figures 10 and 11 related to the confusion matrix. The rapid achievement and stability of maximum training accuracy indicate the model's ability to efficiently learn from the training data.

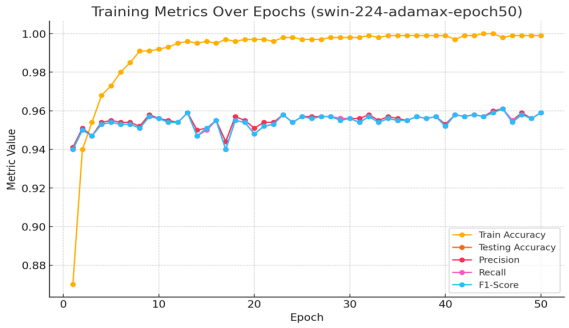

Figure 9. Swin Transformer Model Results with Adamax Optimizer

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

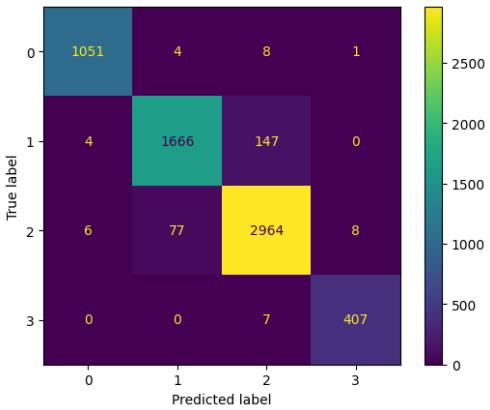

Figure 10. Confusion Matrix of Swin Transformer Model with Adamax Optimizer

Consistent validation metrics show the model's strong generalization capability, which is crucial for real-world applications. Despite some minor fluctuations, the overall performance remains stable and high. The Adamax optimizer has proven effective in achieving and maintaining high model performance with the best time efficiency among other optimizers, at 33,802.70 seconds, making it an excellent choice for training the Swin Transformer model.

C. Testing Results on Swin Transformer Model with Adamax per Class

Based on the analysis from figures 11, 12, and 13, the Swin Transformer model with the Adamax optimizer shows excellent performance in classifying images into four classes: COVID, Lung_Opacity, Normal, and Viral Pneumonia. The confusion matrix in figure 11 shows that the model has very high accuracy for the COVID, Normal, and Viral Pneumonia classes, with accuracies of 0.984, 0.970, and 0.979, respectively. However, the accuracy for the Lung_Opacity class is slightly lower at 0.925, indicating that this class is more challenging to classify correctly.

The class accuracy graph in figure 12 shows that the accuracy for the COVID class remains high and stable around 0.98 throughout the epochs. The accuracy for the Normal class is also very high, almost always above 0.95 after the first few epochs. On the other hand, the accuracy for the Lung_Opacity class varies and tends to be lower compared to the other classes, fluctuating between 0.80 and 0.95. The accuracy for the Viral Pneumonia class is also high, though showing some minor fluctuations, remaining around 0.93 to 0.98.

The performance metrics graph in figure 13 shows that the training accuracy consistently increases and reaches nearly 1.0 by epoch 30, indicating that the model is very effective in learning from the training data. The testing accuracy stabilizes around 0.95 after epoch 10 and reaches its highest accuracy at epoch 18, indicating that the model generalizes well to the testing data. Precision, recall, and F1 score are also consistent and stable around 0.95 after epoch 10, demonstrating a good balance between precision and recall. Overall, despite some difficulty in correctly classifying the Lung_Opacity class, the model's performance is very strong with high accuracy in all other classes.

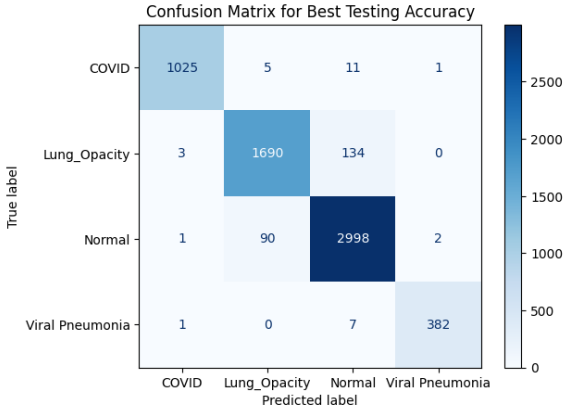

Figure 11. Confusion Matrix of Swin Transformer with Adamax Optimizer per Class

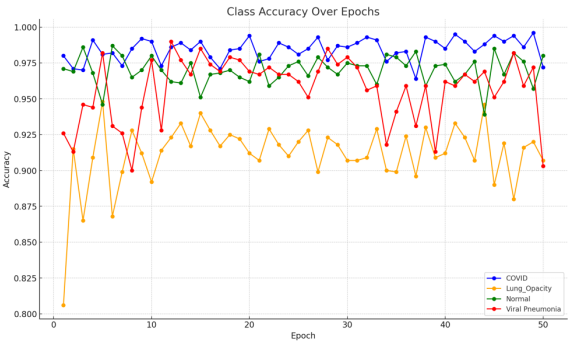

Figure 12. Testing Accuracy per class (COVID, Lung_Opacity, Normal, Viral Pneumonia).

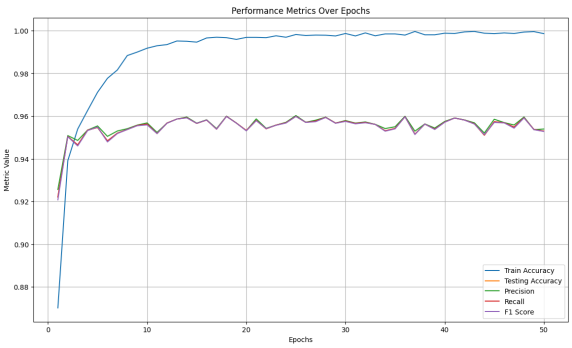

Figure 13. Performance Metrics Swin Transformer Model with Adamax per Class

D. Comparison of Proposed Model Testing Results with Previous Research

In this third scenario, the proposed model is compared with previous research studies, as shown in Table 3.

Table 3. Results of Proposed Model Testing Compared with Previous Research

Model	Class	Optimizer	Train Accuracy	Testing Accuracy (%)	F1-Score (%)
CNN_Model [23]	Normal and Covid-19	Adam	99.08	96.71	97
MobileNetv2 [23]	Normal and Covid-19	Adam	99.93	95.73	96
ResNet50 [23]	Normal and Covid-19	Adam	98.50	91.54	91
CNN_Model [24]	Covid-19, Normal and Pneumonia	Adam	-	95.8	95.79
InceptionV3 [25]	Normal, Covid-19, Lung Opacity, Pneumonia	Adam	-	73.32	-
ResNet50v1 [25]	Normal, Covid-19, Lung Opacity, Pneumonia	Adam	-	92.38	-
ResNet50v1 [25]	Normal, Covid-19, Lung Opacity, Pneumonia	SGD	-	92.22	-
Purpose Method (ViT)	Normal, Covid-19, Lung Opacity, Pneumonia	Adamax	100	95.1	95.1
Purpose Method (Swin)	Normal, Covid-19, Lung Opacity, Pneumonia	Adamax	100	96.1	96.1

E. Discussion

The testing results for both transformer models, Vision Transformer and Swin Transformer, show significant performance variations depending on the hyperparameter tuning applied. Overall, the choice of optimizer has a significant impact on the model's time efficiency and accuracy. AdamW and Adamax stand out as the best options for both models, offering an excellent balance between performance and computational time efficiency.

Training accuracy is frequently higher than testing accuracy, a phenomenon attributed to the tuning of model hyperparameters specifically for the training dataset during training. This specific tuning can lead to overfitting, where the model performs exceptionally well on training data but less effectively on new, unseen test data. For example, while the Swin Transformer achieved a testing accuracy of 96.1% using the Adamax optimizer, it is crucial to also evaluate the test data accuracy to ensure the model's generalization capability. In this study, hyperparameter tuning methods were employed

using various optimizers including Adam, AdamW, Adamax, and others to determine their impact on model performance metrics such as accuracy, precision, recall, and F1-score.

The ViT model generally shows variations in computation time and performance depending on the optimizer used. The Adam optimizer shows the fastest total computation time of 32,635.58 seconds, with a max train accuracy of 99.9% and testing accuracy of 94.7%. The AdamW optimizer records a total computation time of 33,072.99 seconds, with a train accuracy of 99.9% and testing accuracy of 94.8%. The Nadam optimizer has a total computation time of 33,084.32 seconds, with a max train accuracy of 99.8% and testing accuracy of 94.8%. The Adamax optimizer records a slightly higher total computation time of 33,503.10 seconds but achieves a max train accuracy of 100% and testing accuracy of 95.1%.

The Swin model generally shows variations in computation time and performance depending on the optimizer used. The AdamW optimizer shows a total computation time of 32,766.43 seconds, with a maximum train accuracy of 99.8% and testing accuracy of 95.5%. The SGD optimizer records a total computation time of 33,592.71 seconds, with a maximum train accuracy of 92.7% and testing accuracy of 92.9%. The Adamax optimizer records a slightly higher total computation time of 33,802.70 seconds but achieves a maximum train accuracy of 100% and testing accuracy of 96.1%.

In the Swin model, AdamW and Adamax provide the best balance between computation time and accuracy, along with excellent performance metrics. Adam and Nadam are also effective choices, offering very competitive performance.

Although SGD is more efficient in computation time, it shows lower performance compared to other optimizers, making it less ideal for achieving maximum accuracy. RMSprop and LAMB perform well but with higher total computation times.

Pointing the issue regarding the unbalanced dataset used in the experiment. We acknowledge that an unbalanced dataset can lead to biased prediction results, often favoring a dominant class. To address this, we have taken several measures in this research:

We have **added the experimental** results and the best performance evaluation of the **Swim Transformer for each class**. The accuracy for the Covid class is 0.984, for the Normal class is 0.97, and for the Viral Pneumonia class is 0.979. The overall accuracy of the Swim Transformer model is 0.961 or 96.1%. These results indicate that there is no significant difference between the per-class accuracy and the total accuracy, suggesting that the issue of unbalanced data in this dataset is within acceptable limits. For future research, we plan to analyze the problem of unbalanced datasets using imbalanced model approaches such as SMOTE. **Performance Metrics**: in this study employed a variety of performance metrics, including precision, recall, and F1-score, alongside accuracy. These metrics offer a more thorough assessment of the model's performance and assist in understanding its behavior across different classes. **Hyperparameter Tuning**: in this study, performed extensive tuning of hyperparameters and used optimizers such as Adamax and AdamW. These optimizers have proven effective in achieving more balanced performance across classes, despite the dataset imbalance.

Testing Data Comparison: we evaluated the model's performance on test data to ascertain the actual impact of the unbalanced dataset on the model's ability to generalize. This comparison is vital for a comprehensive assessment of the bias introduced by the unbalanced dataset.

Three methods above as strategies to mitigate bias and enhance the reliability and accuracy of our model's predictions.

CONCLUSION

In this study, we evaluated the performance and computational efficiency of the Swin and ViT models using various optimizers: Adam, AdamW, Adamax, Nadam, LAMB, SGD, and RMSprop. The analysis shows that the AdamW and Adamax optimizers consistently provide an optimal balance between computation time and performance metrics such as training accuracy, testing accuracy, precision, recall, and F1-score.

For the Swin model, the AdamW optimizer recorded a total computation time of 32,766.43 seconds with a maximum testing accuracy of 95.5%, while Adamax required 33,802.70 seconds with a maximum testing accuracy of 96.1%. The ViT model showed similar results, where AdamW recorded a total computation time of 33,072.99 seconds with a maximum testing accuracy of 94.8%, and Adamax recorded a computation time of 33,503.10 seconds with a maximum testing accuracy of 95.1%.

The SGD and RMSprop optimizers, although efficient in computation time, showed lower performance in terms of accuracy, particularly in the ViT model. The LAMB optimizer, despite requiring higher computation time, still provided good metric performance but was less efficient compared to AdamW and Adamax.

Overall, the Swin model with the Adamax optimizer achieved the best results in this study. This model demonstrated an optimal combination of high testing accuracy (96.1%) and efficient computation time (33,802.70 seconds). The Adamax optimizer on the Swin model provided strong performance metrics across all aspects (Precision, Recall, F1-score), making it the best choice based on the analyzed data.

ACKNOWLEDGMENT

I would like to extend my sincere gratitude to my advisor and everyone who contributed to the completion of this journal publication as part of my studies in the Magister Program, Department of Magister Information System, Institut Teknologi dan Bisnis STIKOM Bali.

REFERENCES

[1] T. W. Ferkol and D. E. Schraufnagel, "The global burden of respiratory disease," *Ann. Am. Thorac. Soc.*, vol. 11, no. 3, pp. 404–406, 2014, doi: 10.1513/ANNALSATS.201311-405PS.

[2] J. P. Wisnivesky and J. P. de-Torres, "The Global Burden of Pulmonary Diseases: Most Prevalent Problems and Opportunities for Improvement," *Ann. Glob. Heal.*, vol. 85, no. 1, 2019, doi: 10.5334/AOGH.2411.

[3] L. B. Tolle, "Challenges in the Diagnosis and Management of Patients with Fibrosing Interstitial Lung Disease," *Case Rep. Pulmonol.*, vol. 2022, 2022, doi: 10.1155/2022/9942432.

[4] C. Xu, L. Li, and W. Wang, "Challenges in Advanced Lung Cancer Diagnosis During the COVID-19 Pandemic," *Technol. Cancer Res. Treat.*, vol. 21, 2021, doi: 10.1177/15330338211050764.

[5] Y. Tang *et al.*, "Automated abnormality classi fi cation of chest radiographs using deep convolutional neural networks," *npj Digit. Med.*, doi: 10.1038/s41746-020-0273-z.

[6] N. Adithyaram, "Early Detection of Lung Disease Using Deep Learning Algorithms on Image Data," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 7, pp. 466–469, 2023, doi: 10.22214/ijraset.2023.53802.

[7] A. W. Salehi *et al.*, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," *Sustain.*, vol. 15, no. 7, 2023, doi: 10.3390/su15075930.

[8] A. Dosovitskiy *et al.*, "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.

[9] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[10] D. Nam and A. Pak, "Overview of Transformer-Based Models for Medical Image Segmentation," *Sci. J. Astana IT Univ.*, pp. 64–75, 2023, doi: 10.37943/13bkbf2003.

[11] K. He, F. Gou, and J. Wu, "Image segmentation technology based on transformer in medical decision-making system," *IET Image Process.*, vol. 17, no. 10, pp. 3040–3054, 2023, doi: https://doi.org/10.1049/ipr2.12854.

[12] W. Li, "A Swin Transformer with Dynamic High-Pass Preservation for Remote Sensing Image Pansharpening," 2023.

[13] Y. Peng, J. Ren, J. Wang, and M. Shi, "Spectral-Swin Transformer with Spatial Feature Extraction Enhancement for Hyperspectral Image Classification," pp. 1–19, 2023.

[14] J. Huang *et al.*, "Swin transformer for fast MRI," *Neurocomputing*, vol. 493, pp. 281–304, 2022, doi: 10.1016/j.neucom.2022.04.051.

[15] I. Sonata, Y. Heryadi, A. Wibowo, and W. Budiharto, "End-to-End Steering Angle Prediction for Autonomous Car Using Vision Transformer," vol. 17, no. 2, pp. 221–234, 2023.

[16] Y. Zhang, J. Wang, and J. M. Gorriz, "Deep Learning and Vision Transformer for Medical Image Analysis," pp. 9–12, 2023.

[17] J. A. Figo, N. Yudistira, and A. W. Widodo, "Deteksi Covid-19 dari Citra X-ray menggunakan Vision Transformer," vol. 7, no. 3, pp. 1116–1125, 2023.

[18] G. G. Tahyudin, "Klasifikasi Gender Berdasarkan Citra Wajah Menggunakan Vision Transformer," vol. 10, no. 2, pp. 1808–1823, 2023.

[19] "CHEST X-RAY DATABASE." https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database (accessed Feb. 01, 2024).

[20] Y. Zhang and W. H. Id, "Vision Transformer with hierarchical structure and windows shifting for person re-identification," pp. 1–16, 2023, doi: 10.1371/journal.pone.0287979.

[21] S. Aburass and O. Dorgham, "Performance Evaluation of Swin Vision Transformer Model Using Gradient

Accumulation Optimization Technique," pp. 56–64, 2023, doi: 10.1007/978-3-031-47448-4_5.

[22] P. D. Wulaning Ayu and G. A. Pradipta, "U-Net Tuning Hyperparameter for Segmentation in Amniotic Fluid Ultrasonography Image," *2022 4th Int. Conf. Cybern. Intell. Syst. ICORIS 2022*, no. June, 2022, doi: 10.1109/ICORIS56080.2022.10031294.

[23] Tamil, "Biomedical Signal Processing and Control Deep learning-based approach for detecting COVID-19 in chest X-rays," *Biomed. Signal Process. Control*, vol. 78, no. January, p. 103977, 2022, [Online]. Available: https://doi.org/10.1016/j.bspc.2022.103977

[24] B. T. Magar, K. Shrestha, and M. A. Rahman, "CNN-based Clinical Diagnosis and Decision Support System for Chest X-ray," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1305, no. 1, p. 012027, 2024, doi: 10.1088/1757-899x/1305/1/012027.

[25] A. W. Setiawan, "Perbandingan Arsitektur Convolutional Neural Network Pada Klasifikasi Pneumonia, COVID-19, Lung Opacity, dan Normal Menggunakan Citra Sinar-X Thoraks," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 7, pp. 1563–1570, 2022, doi: 10.25126/jtiik.2022976742.