Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Classification of Pneumonia Using CNN and Vision Transformer

Ma'dan Shomsomi[1], Widhaksa Triawan[2], Purwadi[3*]

[1,2]Informatic, Faculty of Computer Science, Amikom Purwokerto University
[3]Master of Computer Science, Faculty of Computer Science, Amikom Purwokerto University
madanshomsomi2@gmail.com[1], widhaksatriawan7@gmail.com[2], purwadi@amikompurwokerto.ac.id[3*]

Abstract

Pneumonia remains one of the leading causes of mortality among children worldwide. This study aims to evaluate the performance of two deep learning architectures, Convolutional Neural Network (CNN) and Vision Transformer (ViT), for pneumonia classification using chest X-ray images. Four training scenarios were examined, consisting of MobileNetV2 baseline, MobileNetV2 fine-tuned, ViT baseline, and ViT fine-tuned models. The dataset was obtained from the Chest X-Ray Images (Pneumonia) collection and was processed through augmentation and preprocessing to produce a balanced set of 9,000 images. Baseline models were trained using a feature extraction approach, while fine-tuning was conducted by selectively unfreezing internal layers. Experimental results show that all models achieved accuracy above 95%. The MobileNetV2 baseline reached 97.63%, while its fine-tuned counterpart did not yield further improvement, achieving 97.41%. In contrast, the Vision Transformer demonstrated substantial performance gains, where partial fine-tuning produced the highest accuracy of 98.59% with an f1-score of 0.99. These findings indicate that ViT with targeted fine-tuning is more effective in capturing global representations within X-ray images, making it a strong candidate for computer-aided pneumonia detection systems supported by artificial intelligence.

Keywords: *Deep Learning; MobileNetV2; Pneumonia; Vision Transformer; X-Ray Imaging*

1. Introduction

Pneumonia is one of the most common respiratory tract infections in the world and remains a leading cause of death, especially among children. According to the 2021 Global Burden of Disease (GBD) report, there are approximately 344 million new cases of lower respiratory tract infections (LRTI) with an incidence rate of around 4,350 cases per 100,000 population. This condition causes 2.18 million deaths or 27.7 deaths per 100,000 population, with approximately 502,000 deaths occurring in children under five years of age, and more than half (approximately 254,000) in countries with low socio-demographic indices [1][2]. Data from the United Nations Children's Fund (UNICEF) in 2024 also reports that every 43 seconds, one child under five dies from pneumonia, with a total of 700,000 deaths each year, including 190,000 newborns [3]. These high figures underscore the urgency of developing a faster, more accurate, and reliable early detection system.

Advances in artificial intelligence (AI) have encouraged the use of medical image processing to support the diagnosis of pneumonia more efficiently [4]. One of the most widely used paradigms is Convolutional Neural Network (CNN) due to its ability to extract visual patterns hierarchically without requiring manual feature engineering [5][6]. This study used the MobileNetV2 CNN model, which is designed for mobile devices to produce small models through the use of separable convolution to reduce the computational load without sacrificing performance, and is known as a lightweight and efficient architecture [7][8][9]. MobileNetV2 has been widely used in various medical applications due to its good performance even when run on devices with limited resources [9][10][11].

Although effective, CNN architectures such as MobileNetV2 have limitations in capturing global context and long-range spatial relationships in images[7]. Vision Transformer (ViT) was developed to overcome these limitations through a self-attention mechanism capable of modeling global dependencies in visual data [12][13][14]. ViT has demonstrated competitive performance in various chest radiography classification tasks, including pneumonia detection, with an advantage in stronger global representation capabilities compared to CNN [15]. The study by Angara et al. [16], for example, shows that the combination of CNN and ViT in an ensemble architecture achieves 94.87% accuracy on the Kaggle pediatric pneumonia dataset, outperforming single models.

In addition to the advantages of each architecture, the development of deep learning-based pneumonia classification systems still faces challenges, such as limited high-quality data, class imbalance, and the need for adequate model interpretability for clinical adoption [6]. Various studies have evaluated fine-tuning techniques on CNN models, but the scope of comparison is generally still limited. Most studies

only compare the performance between CNN architectures, such as MobileNet, EfficientNet, and NASNetMobile [17][18].Meanwhile, several other studies focus on Transformer variants. For example, Mahajaya et al. [19] compared ViT (95.10% accuracy) with Swin Transformer (96.10% accuracy). To date, there has been no comprehensive analysis that directly compares MobileNetV2 and Vision Transformer in a unified experimental scheme with consistent baseline and fine-tuning configurations.

This study was designed to provide a more structured comparative analysis between four deep learning model configurations MobileNetV2 baseline, MobileNetV2 fine-tuning, ViT baseline, and ViT fine-tuning to assess the influence of architecture and training strategies, from feature extraction to partial unfreezing, on feature representation quality and pneumonia classification performance in X-ray images. The relevance of this study is supported by the fundamental differences between CNNs, which excel at local pattern extraction, and Vision Transformers, which are capable of capturing global context through self-attention. The results of this study are expected to provide empirical and conceptual contributions to the selection of appropriate models for pneumonia detection systems, particularly in healthcare environments with computational limitations, as well as to enrich the understanding of representation characteristics between architectures. Although this study is still limited to a single dataset without cross-source validation or clinical trials, the findings remain an important foundation for further research and the application of more efficient and adaptive medical detection technology in medium-scale healthcare facilities

2. Research Method

This study systematically applies several methodological stages to achieve its main objective: to present a comparative analysis of deep learning models for X-ray-based pneumonia classification. The main focus of the study is to compare two fundamental architectures: Convolutional Neural Network (CNN), represented by MobileNetV2, and Vision Transformer (ViT).
This approach was designed to evaluate the effectiveness of four different training scenarios: MobileNetV2 baseline (feature extraction), MobileNetV2 fine-tuned, ViT baseline (feature extraction), and ViT fine-tuned. These four scenarios were tested on a binary classification task to distinguish X-ray images into two classes: Pneumonia and Normal. The overall research methodology includes data collection, data pre-processing and augmentation, architecture configuration, training of the four scenarios, and comparative performance evaluation. The complete flow of the research stages is illustrated in Figure 1 below.

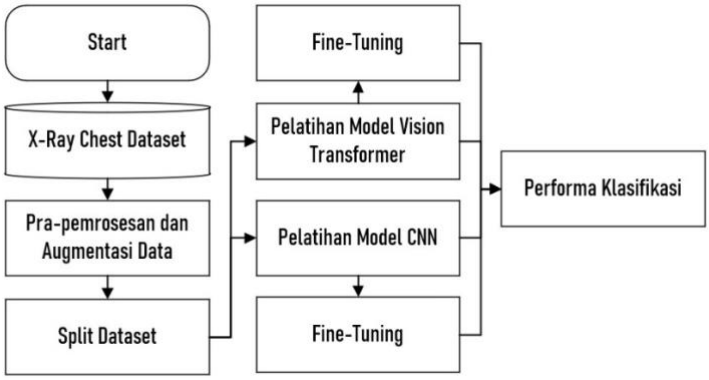

Fig. 1: Research Stages

2.1. Dataset

This study uses a publicly available dataset of chest X-ray images titled Chest X-ray Images (Pneumonia) on Kaggle, which originates from pediatric clinical data and has been widely used in deep learning-based pneumonia classification studies. This dataset was first introduced and validated in a study by Kermany et al. (2018), which demonstrated the effectiveness of the transfer learning approach in distinguishing between normal conditions and pneumonia using chest X-ray images [20].

Fig. 2: Sample Dataset X-Ray

The dataset is divided into three subsets, namely train, validation, and test, with two main labels: NORMAL and PNEUMONIA. The initial data distribution is imbalanced, where the training data contains 1341 NORMAL images and 3875 PNEUMONIA images, while the validation data only contains 8 images for each class, and the test data consists of 234 NORMAL images and 390 PNEUMONIA images. This imbalance has the potential to cause model bias in the training process, so a data augmentation strategy was carried out on the minority

class (NORMAL) until the number of examples in both classes became more proportional, with the aim of improving generalization capabilities and avoiding prediction domination by the majority class.

2.2. Pre-processing and Augmentation

To address class distribution imbalance and improve model generalization capabilities, this study applied a data augmentation process during the preprocessing stage. After combining the initial datasets, each class (NORMAL and PNEUMONIA) was balanced to reach 4,500 images each, resulting in a total of 9,000 images after augmentation. Augmentation was performed using ImageDataGenerator with controlled geometric transformations relevant to medical images, including rotation up to 15°, horizontal and vertical shifts of 10%, 10% shear, 10% zoom, horizontal flipping, and pixel filling using the nearest interpolation method to preserve the pathological visual characteristics of X-ray images. All images then went through a standard preprocessing stage, namely resizing to 224×224 pixels to be compatible with the model input, and normalizing the pixel intensity values by scaling them to the range [0,1] by dividing the pixel values by 255.0, in order to accelerate model convergence and stabilize the training process.

After the augmentation and preprocessing stages, the dataset was randomly divided using a 70:15:15 ratio into training, validation, and testing subsets, with random shuffling to prevent sampling bias. The partition results show a balanced class distribution in each subset, with 3,150 images per class in the training data (6,300 total), 675 images per class in the validation data (1,350 total), and 675 images per class in the testing data (1,350 total). This proportional distribution is designed to ensure that the model does not experience the dominance of a particular class during the learning and evaluation process, as well as to guarantee the reliability of the model performance analysis in each experimental scenario.

2.3. Global Seed Configuration

To ensure the reproducibility of experiments and consistency of model training results, this study applies global seed settings to all modules that have the potential to generate random processes. In CNN-based experiments (TensorFlow), the seed is set through the random, numpy, and tensorflow libraries using a constant value of 42, so that processes such as weight initialization, dataset randomization, and batch formation will produce identical sequences in each execution. Meanwhile, in Vision Transformer-based experiments (PyTorch), the seed is applied to the random, numpy, and torch modules, including specific GPU settings via torch.cuda.manual_seed_all().

Additionally, the flag torch.backends.cudnn.deterministic is enabled and torch.backends.cudnn.benchmark is disabled to stabilize computational behavior on the GPU, so that convolution and backpropagation operations do not use non-deterministic algorithms that can produce variations in output between executions. With these settings in place, the entire model training pipeline, from data shuffling, random augmentation, parameter initialization, to model optimization, can be executed consistently, resulting in more reliable, replicable, and scientifically valid experimental results.

2.4. Convolutional Neural Network (CNN)

2.4.1. CNN Baseline

The CNN model uses the MobileNetV2 architecture with a transfer learning approach as a feature extractor. All weights in the backbone are frozen, and the classification section is rebuilt using Global Average Pooling, one hidden layer with 128 units, dropout as regularization, and sigmoid output for binary classification. The dataset consists of images labeled NORMAL and PNEUMONIA that are processed into 224×224 pixels and normalized. The model is trained using the Adam optimizer and binary cross-entropy, with ModelCheckpoint and EarlyStopping mechanisms to maintain validation performance. After training, the best model is evaluated using test data through accuracy, loss, confusion matrix, and classification report metrics.

2.4.2. CNN Fine-Tuned

During the fine-tuning stage, most of the MobileNetV2 layers are reactivated, particularly the 100 deepest layers, to enable weight adjustment to the characteristics of X-ray images. The model is recompiled using a smaller learning rate and trained as a continuation of the baseline model. The training process is monitored using early stopping and model checkpoints. Evaluation is performed using the same metrics as the baseline, and training curve analysis is used to observe the effects of fine-tuning on model stability and generalization capabilities.

2.5. Vision Transformer (ViT)

2.5.1. Vit Baseline

The Vision Transformer (ViT-Base-Patch16-224) model is used as a feature extractor with the entire backbone frozen. Only the classification layer is trained for a two-class task using Cross-Entropy Loss and the Adam optimizer. Data is processed through a PyTorch-based pipeline, including image resizing to 224×224 pixels and feature extraction using ViTFeatureExtractor. Training is run with early stopping, and model performance is evaluated based on accuracy, loss, confusion matrix, and classification report.

2.5.2. Vit Fine-Tuned

The fine-tuning scenario was conducted using a partial unfreezing approach on the last three ViT encoder blocks. The classification layer was equipped with dropout as regularization, and training was performed using AdamW with weight decay and a cosine annealing scheduler. An early stopping procedure was applied to prevent overfitting. After training, the best model was evaluated using standard

metrics (accuracy, loss, precision, recall, F1-score). Accuracy and loss curve analysis was used to assess the convergence improvement due to fine-tuning.

3. Result and Discussion

The training process across all models shows fairly consistent dynamics, where the accuracy and loss curves on both training and validation data move towards a convergent pattern as the number of epochs increases. Although the characteristics of each architecture differ, ranging from CNNs that tend to be stable from the outset to Vision Transformers that are more sensitive to the number of epochs, each model can reach a performance equilibrium after several training iterations. The early stopping mechanism helps ensure that each model stops at its best epoch, which is reflected in the varying number of iterations: baseline models tend to converge faster, while the fine-tuning process requires additional epochs to adjust internal weights. In general, the entire training process is stable without extreme fluctuations, so that the more detailed results in the next subsection can accurately describe the performance of each model.

3.1. CNN Model Evaluation (MobileNetV2)

This section discusses the performance of two training scenarios on the MobileNetV2 architecture, namely the baseline model that uses a feature extraction approach and the fine-tuned model with partial layer opening. The analysis is based on the training curve, classification report, and confusion matrix to assess the generalization capabilities of each model.

3.1.1. CNN Baseline

The baseline model uses the MobileNetV2 architecture with the entire backbone frozen so that it functions as a feature extractor. Training is only performed on the classification layer to obtain baseline performance before the fine-tuning process. The training results show a stable convergence pattern, where the validation accuracy increases to 97.33% with a consistent decrease in validation loss as the number of epochs increases. The accuracy and loss curves are shown in Figure 3.

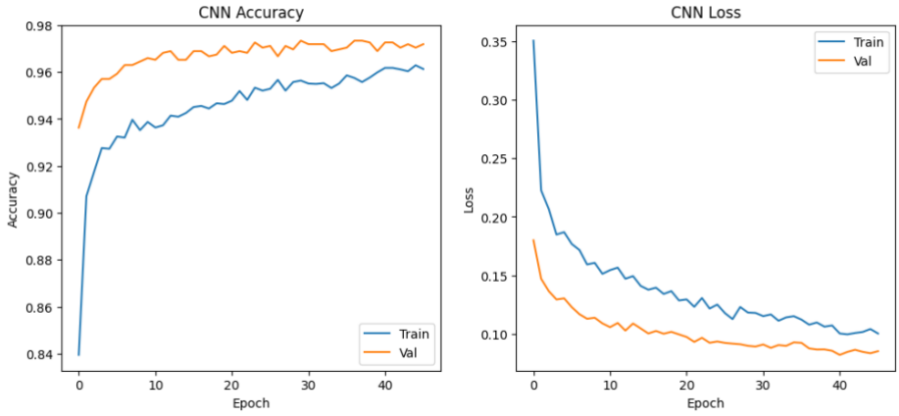

Fig. 3: Accuracy and loss curves on the baseline CNN model

Evaluation using 1,350 test images resulted in an accuracy of 97.63%, indicating that the feature representation obtained from the MobileNetV2 backbone is effective enough to distinguish between normal and pneumonia images without additional weight adjustments. The precision, recall, and f1-score values for both classes were in the range of 0.97 - 0.98, as shown in Table 1. The high recall value for the pneumonia class indicates good detection of pathological patterns, while the high precision indicates a low rate of positive prediction errors.

Table 1: Classification Report CNN Baseline

Kelas	Precision	Recall	F1-score	Support
Normal	0.98	0.97	0.98	675
Pneumonia	0.97	0.98	0.98	675
Accuracy			0.98	1350

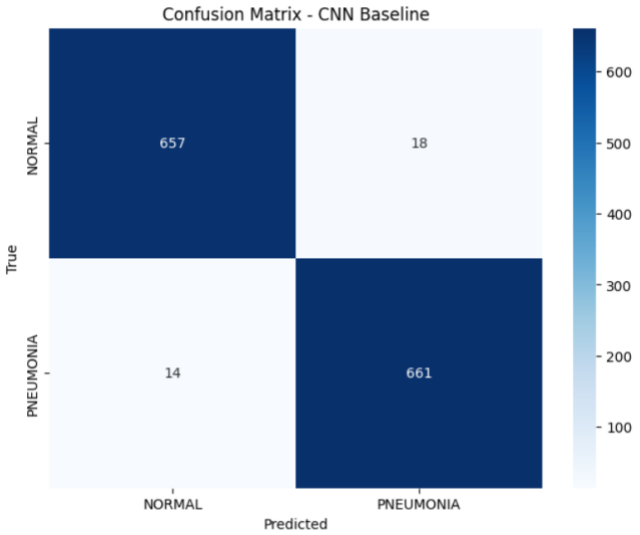

Fig. 4: Confusion matrix model CNN baseline

Classification errors generally occur in images with low contrast or radiographic artifacts, which are often found in public X-ray datasets. Overall, these results show that MobileNetV2 is capable of providing a strong baseline with stable generalization. However, because its backbone is not frozen, its ability to adapt to subtle variations in radiographic patterns is still limited, requiring fine-tuning to improve medical feature representation.

3.1.2. CNN Fine-Tuned

Fine-tuning was performed to improve MobileNetV2's ability to recognize chest radiography characteristics that were not fully captured in the baseline model. At this stage, most of the layers in the backbone were reactivated so that the model could learn higher-level feature representations that were more relevant to the pathological patterns of pneumonia

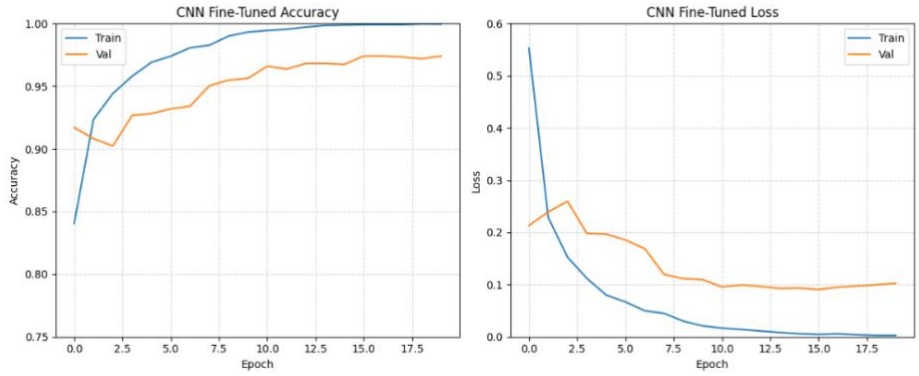

Fig. 5: Accuracy and loss curves on the fine-tuned CNN model

After the advanced training process, the model showed moderate performance improvement compared to the baseline. Validation accuracy increased to 97.41% with a stable downward trend in validation loss to 0.09. In the early stages of fine-tuning, the model underwent readjustment to new weights, but performance stabilized after several epochs and reached convergence at around the 62nd epoch.

Evaluation of 1,350 test images resulted in an accuracy of 97.41%, which is consistent with the validation value. The classification report shows balanced performance in both classes with an f1-score of 0.97, as shown in Table 2. High precision in the normal class (0.99) indicates a low false positive rate, while high recall in the pneumonia class (0.99) indicates good ability to detect abnormal cases.

Table 2: Classification Report CNN Fine-tuned

Kelas	Precision	Recall	F1-score	Support
Normal	0.99	0.96	0.97	675
Pneumonia	0.96	0.99	0.97	675
Accuracy			0.97	1350

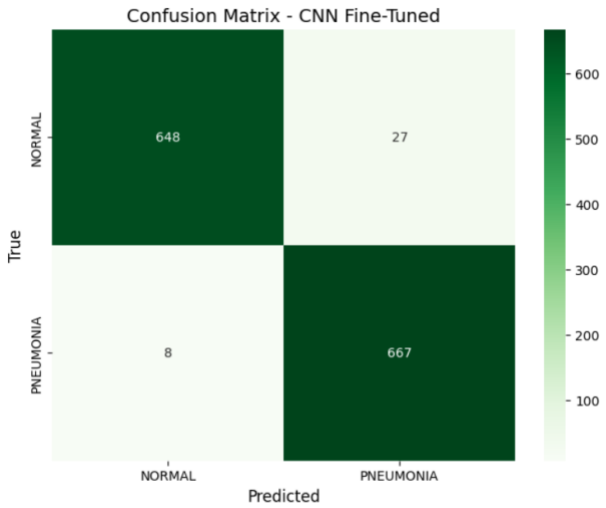

Fig. 6: Confusion matrix model CNN fine-tuned

Some classification errors occurred in images of pneumonia with low intensity or artifacts resembling normal lung structures. However, the number of errors was relatively small and did not significantly impact overall performance. These results confirm that opening the deepest layer successfully enriched the representation of lung texture features, making the model more adaptive to complex radiographic image variations.

The fine-tuning strategy provides increased sensitivity to pathological patterns and better performance stability compared to the baseline. Although the margin of improvement is not very large, the fine-tuned model still shows advantages in terms of representation depth and adaptability, making it worth considering for implementation in deep learning-based pneumonia diagnosis support systems.

3.2. Vision Transformer Model Evaluation (ViT-Base-Patch16-224)

This section presents an evaluation of the performance of the Vision Transformer (ViT) in two training scenarios, namely the baseline ViT and the fine-tuned ViT. In the baseline configuration, the entire transformer backbone is kept frozen to assess the initial representation capabilities without adjusting the weights to the radiography domain. Meanwhile, the fine-tuning scenario is performed by partially unfreezing the encoder blocks so that the model can learn the texture patterns and global structures characteristic of X-ray images. Performance analysis includes accuracy and loss curves during training, classification reports, confusion matrices, and comparisons between the two configurations to assess the effect of partial unfreezing on the model's generalization ability.

3.2.1. ViT Baseline

The Vision Transformer (ViT) model was used as a comparison to the CNN architecture to assess the effectiveness of the self-attention mechanism in detecting pneumonia in chest X-ray images. In this baseline scenario, the entire ViT backbone was frozen, so training was only performed on the classification layer. This approach aimed to evaluate the initial representation capability of ViT without weight adjustment to the radiography domain.

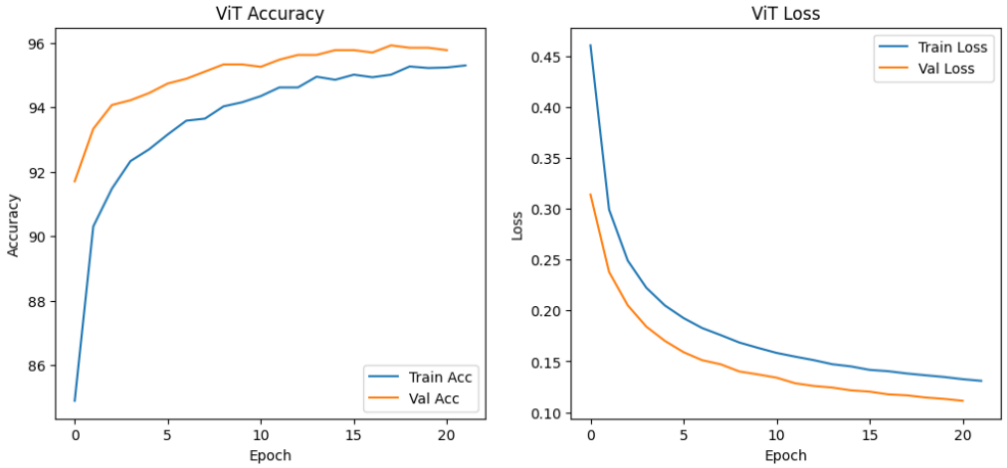

Fig. 7 : Accuracy and loss curves on the baseline ViT model

During the training process, the model showed a steady increase in accuracy, reaching around 95.93%, with a consistent downward trend in validation loss. This shows that ViT is able to recognize general patterns that distinguish between normal images and pneumonia even though the backbone was not retrained. Evaluation on 1,350 test images resulted in an accuracy of 95.85% and a test loss of 0.1102,

indicating fairly good generalization for the baseline configuration. Identical precision, recall, and f1-score values for both classes (0.96 each) indicate that the model is able to balance sensitivity and specificity in classification, as shown in Table 3.

Table 3 : Classification Report ViT Baseline

Kelas	Precision	Recall	F1-score	Support
Normal	0.96	0.96	0.96	675
Pneumonia	0.96	0.96	0.96	675
Accuracy			0.96	1350

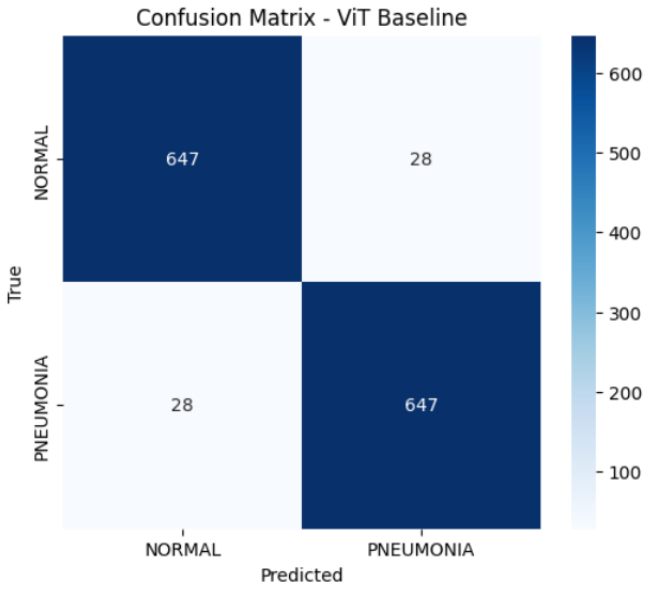

Fig. 8 : Confusion matrix model ViT baseline

Classification errors mostly occur in images of mild pneumonia that visually resemble normal lung tissue. However, the proportion of errors is relatively small, so the model remains reliable in identifying common radiographic abnormalities.

These results indicate that ViT has strong feature representation capabilities, but its adaptability to radiographic textures is still limited when the backbone is not retrained. Therefore, a fine-tuning stage is necessary to adjust the internal weights to be more responsive to morphological variations in X-ray images, which is ultimately expected to significantly improve performance.

3.2.2. ViT Fine-Tuned

The fine-tuning approach on Vision Transformer (ViT) was performed to improve the model's ability to recognize pathological patterns in chest X-ray images in greater depth. Unlike the baseline configuration, which only trains the classification layer, this stage opens part of the encoder so that the model can adjust weights to medical domain characteristics.

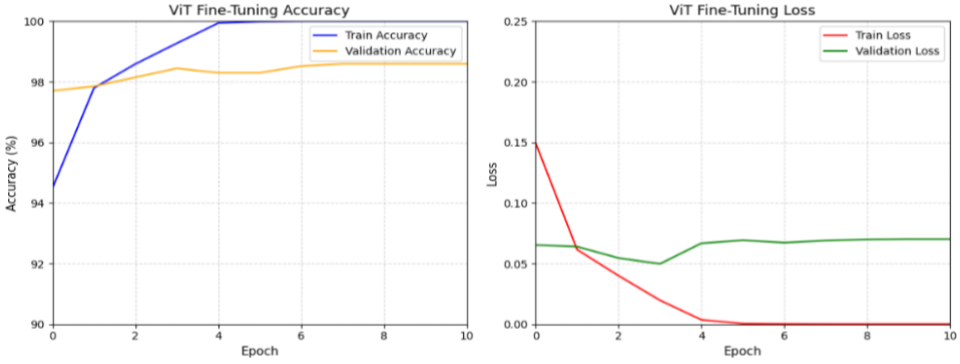

Fig. 9 : Accuracy and loss curves on fine-tuned ViT models

During training, the model showed a more significant improvement in performance compared to the baseline. Validation accuracy increased from around 97.70% in the early stages to 98.59% before the early stopping mechanism halted training at epoch 11. The training curve showed a stable downward trend in loss on both the training and validation data, as well as a convergent increase in accuracy, indicating that the weight adjustment on the encoder block was successful without causing overfitting.

Evaluation on 1,350 test images resulted in an accuracy of 98.59%, a significant improvement over the baseline model. The test loss value of 0.0702 indicates solid generalization capabilities in recognizing radiographic patterns of pneumonia, even though the fine-tuning process was only performed on a small portion of the transformer layers.

Table 4 : Classification Report ViT Fine-Tuned

Kelas	Precision	Recall	F1-score	Support
Normal	0.99	0.99	0.99	675
Pneumonia	0.99	0.99	0.99	675
Accuracy			0.99	1350

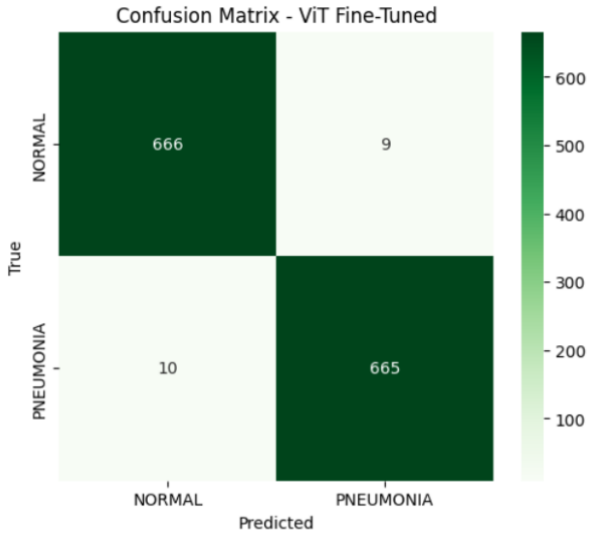

Fig. 10 : Confusion matrix model ViT fine-tuned

The confusion matrix shows that almost all images in both classes were classified correctly. Prediction errors only occurred in a small number of images with subtle opacity patterns or radiographic textures resembling normal lung tissue. However, the number of errors was so minimal that it did not significantly affect overall performance.

The 2.74% increase in accuracy compared to the baseline shows that opening the last three encoder blocks is very effective in improving the quality of ViT feature representation. Fine-tuning allows the model to adjust spatial attention (self-attention) to typical radiographic patterns, such as lung intensity distribution, infiltrate shadows, and opacity areas that are indicators of pneumonia.

Overall, ViT Fine-Tuned is the model with the highest performance in this study. Its ability to learn the global structure of images through the attention mechanism results in very high sensitivity and specificity, making it a strong candidate for implementation in X-ray image-based computer-aided diagnosis systems. Its performance stability and near-perfect accuracy show that this approach has great potential to support fast and accurate clinical screening processes.

3.3. Performance Comparison of All Models

An evaluation was conducted on four main models: baseline CNN, fine-tuned CNN, baseline ViT, and fine-tuned ViT, each of which was trained using chest X-ray images with two target classes (Normal and Pneumonia). Model performance was compared based on the accuracy, precision, recall, and f1-score metrics using 1,350 test images. An overview of the overall performance is shown in Table 5.

Table 5: Comparison of CNN and ViT Model Performance

Model	Architecture	Training Strategy	Accuracy (%)	Precision	Recall	F1-Score
CNN Baseline	MobileNetV2	Transfer learning (frozen)	97.63	0.98	0.98	0.98
CNN Fine-Tuned	MobileNetV2	Full Fine-tuning	97.41	0.97	0.97	0.97

ViT Baseline	ViT-Base Patch16	Transfer learning (frozen)	95.85	0.96	0.96	0.96
ViT Fine-Tuned	ViT-Base Patch16	Partial unfreeze + regularization	98.59	0.99	0.99	0.99

3.4. Model Performance Analysis

In general, all models showed high performance with accuracy above 95%, confirming the effectiveness of the transfer learning approach for medical image classification. However, each model showed different performance characteristics according to its architecture and training strategy.

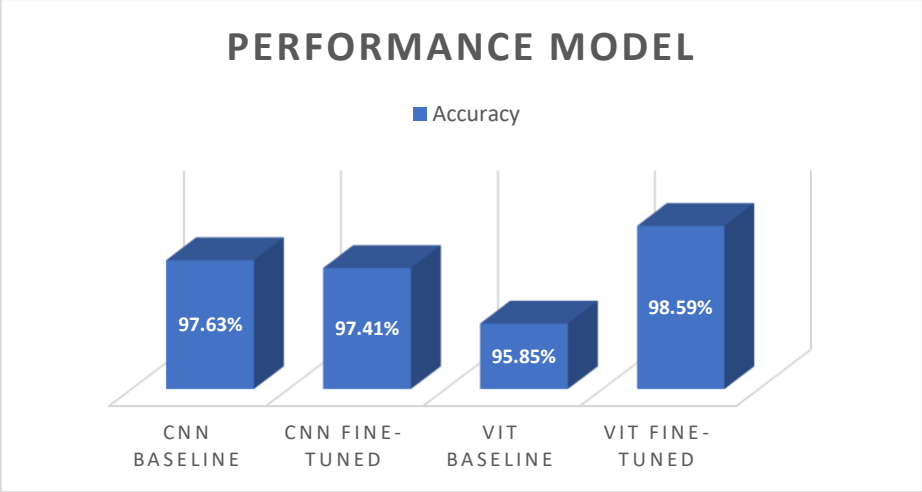

Fig. 11: Comparison of all models

The MobileNetV2-based baseline CNN model achieved an accuracy of 97.63% and an F1-score of 0.98. These results show that feature representations from ImageNet pretrained weights are quite effective in extracting radiographic patterns, even without retraining on the backbone. Stable performance and good generalization indicate that this model has a good balance between bias and variance.

The fine-tuned CNN model experienced a slight decline in performance with an accuracy of 97.41%, reflecting mild overfitting due to retraining many layers on a relatively limited medical dataset. Although in theory fine-tuning can improve performance, in this case the baseline has already achieved very high accuracy, leaving little room for improvement. Opening up to 100 layers also has the potential to trigger catastrophic forgetting, which is the disruption of previously optimal feature representations from ImageNet pretraining, mainly because the fine-tuning learning rate is still too large for sensitive weight updates. Furthermore, the domain difference between ImageNet and X-ray images means that aggressive fine-tuning is not always effective, as the pretrained features become less stable after being modified. Overall, these conditions show that fine-tuning does not always guarantee improved performance, especially when the baseline is already very good and the data variation is not large enough to support weight updates across many layers.

The ViT baseline shows an initial accuracy of 95.85%, which is lower than the CNN model. This is understandable because the Vision Transformer architecture is essentially designed to work optimally on large datasets. Without internal weight updates, the feature representations resulting from pretraining on natural images are not yet fully compatible with the characteristics of medical images.
In contrast, the fine-tuned ViT experienced a very significant improvement in performance and became the best model with an accuracy of 98.59% and an f1-score of 0.99. The partial unfreezing strategy on the last encoder block allowed the model to selectively adapt to radiographic patterns without sacrificing the stability of the initial weights. The addition of dropout and weight decay also contributed to preventing overfitting, thereby maintaining high generalization capabilities.

Overall, this comparison shows that ViT has an advantage in capturing complex global patterns in X-ray images through the self-attention mechanism, while CNN is stronger at extracting local patterns but less effective at broad spatial representation.

3.5. Interpretation and Implications

Comparatively, fine-tuned ViT showed the best performance compared to the other three models. This advantage indicates that the global attention mechanism in Transformer is more adaptive to variations in texture and radiographic intensity than the convolutional approach. The partial unfreezing strategy proved to be more effective than full fine-tuning, as it was able to maintain the stability of pre-trained weights while providing flexibility to adapt to new domains.

Although MobileNetV2 continues to demonstrate high performance and good computational efficiency, its limitations in integrating global information are a factor that reduces its sensitivity to diffuse pneumonia patterns. In contrast, ViT, which has been adjusted through fine-tuning, is able to capture the spatial relationships between lung areas more accurately, thus providing more consistent classification results. These findings confirm that the Vision Transformer approach with a targeted fine-tuning strategy has strong potential for use in X-ray image-based computer-aided diagnosis (CAD) systems. The high accuracy and stability of the model indicate that this architecture is worth considering for clinical applications that require rapid and accurate detection of pneumonia.

4. Conclusion

This study compares the performance of four deep learning models: MobileNetV2 baseline, MobileNetV2 fine-tuned, ViT baseline, and ViT fine-tuned in detecting pneumonia in X-ray images. All models show high accuracy above 95%, confirming that transfer learning is effective for medical image classification. MobileNetV2 baseline recorded strong performance with an accuracy of 97.63% and remains an efficient choice for systems with computational limitations. Full fine-tuning on MobileNetV2 did not provide a significant performance improvement. The Vision Transformer baseline achieved an accuracy of 95.85%, but its performance improved sharply after partial fine-tuning. The ViT Fine-Tuned model achieved the best results with an accuracy of 98.59% and an f1-score of 0.99, showing that gradual adjustments to the encoder layer can strengthen the model's ability to recognize complex radiographic patterns. Based on these findings, ViT Fine-Tuned is the most superior model for pneumonia detection, while MobileNetV2 remains relevant for applications that require computational efficiency. Since this study used a single dataset source, additional validation on real clinical data and more varied datasets is still needed to ensure the reliability of the model in medical practice.

In this section you should present the conclusion of the paper. Conclusions must focus on the novelty and exceptional results you acquired. Allow a sufficient space in the article for conclusions. Do not repeat the contents of Introduction or the Abstract. Focus on the essential things of your article.

Acknowledgement

The author would like to express his gratitude and appreciation to all parties who have provided support during this research process. Special thanks go to Mr. Hendra Marcos, S.T., M.Eng. and Mr. Dr. Purwadi, M.Kom. for their guidance, direction, and valuable scientific input, which enabled this research to be completed successfully. The author also thanks the academic institutions and laboratories that provided computing facilities and a supportive research environment.

Appreciation is also given to the provider of the Chest X-ray Images (Pneumonia) dataset, which enabled this research to be conducted openly. The author is also grateful for the moral support from family, fellow students, and all those who cannot be mentioned one by one but who have contributed to the completion of this research. The author hopes that the results of this research can provide benefits and contribute to the development of artificial intelligence-based disease detection technology.

References

[1] R. G. Bender *et al.*, "Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990–2021: a systematic analysis from the Global Burden of Disease Study 2021," 2024. doi: 10.1016/S1473-3099(24)00176-2.
[2] K. K. (Kemenkes), "Pneumonia Terus Ancam Anak-anak," 2024. [Online]. Available: https://kemkes.go.id/id/pneumonia-terus-ancam-anak-anak
[3] U. N. C. F. (UNICEF), "A child dies of pneumonia every 43 seconds." [Online]. Available: https://data.unicef.org/topic/child-health/pneumonia/
[4] P. Rosyani, S. Saprudin, and R. Amalia, "Klasifikasi Citra Menggunakan Metode Random Forest dan Sequential Minimal Optimization (SMO)," *J. Sist. dan Teknol. Inf.*, vol. 9, no. 2, p. 132, 2021, doi: 10.26418/justin.v9i2.44120.
[5] R. A. Saputra, D. R. R. Putra, and M. A. Asyrofi, "Implementasi Convolutional Neural Network (CNN) Untuk Mendeteksi Penggunaan Masker Pada Gambar," *J. Inform. dan Tek. Elektro Terap.*, vol. 11, no. 3, pp. 710–714, 2023, doi: 10.23960/jitet.v11i3.3286.
[6] H. F. Fadhilah and R. Kurniawan, "Keunggulan dan Tantangan dalam Penggunaan Computer Vision untuk Diagnosis Pneumonia Pediatri: A Systematic Review," *J. Biostat. Kependudukan, dan Inform. Kesehat.*, vol. 5, no. 1, 2024, doi: 10.7454/bikfokes.v5i1.1077.
[7] S. Arnandito and T. B. Sasongko, "Comparison of EfficientNetB7 and MobileNetV2 in Herbal Plant Species Classification Using Convolutional Neural Networks," *J. Appl. Informatics Comput.*, vol. 8, no. 1, pp. 176–185, 2024, doi: 10.30871/jaic.v8i1.7927.
[8] O. F. Altal *et al.*, "Hybrid attention-enhanced MobileNetV2 with particle swarm optimization for endometrial cancer classification in CT images," *Informatics Med. Unlocked*, vol. 57, no. February, p. 101662, 2025, doi: 10.1016/j.imu.2025.101662.
[9] S. Agustiani, R. Aryanti, S. Khotimatul Wildah, Y. T. Arifin, S. Marlina, and T. Misriati, "Optimisasi Model Deep Learning untuk Deteksi Penyakit Daun Tebu dengan Fine-Tuning MobileNetV2," *J. Informatics Manag. Inf. Technol.*, vol. 4, no. 4, pp. 150–157, 2024, doi: 10.47065/jimat.v4i4.411.
[10] A. Syakuroh, F. Monado, M. Ariani, Hadi, E. Koriyanti, and Erni, "Analisis Akurasi Model Mobilenetv2 Dalam Klasifikasi Citra X-Ray Untuk Deteksi Kondisi Paru-paru," *J. Online Phys.*, vol. 10, no. 3, pp. 67–74, 2025, doi: 10.22437/jop.v10i3.44453.
[11] D. F. Rajendra and A. K. Wardhana, "Development of MobileNetV2 for CT-Scan Lung Classification Using Transfer Learning," vol. 9, no. 5, pp. 2703–2710, 2025, doi: 10.30871/jaic.v9i5.10282.
[12] A. Musha, A. Al Mamun, A. Tahabilder, J. Hossen, B. Jahan, and S. Ranjbari, "A deep learning approach for COVID-19 and pneumonia detection from chest X-ray images," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 4, pp. 3655–3664, 2022, doi: 10.11591/ijece.v12i4.pp3655-3664.
[13] M. F. Fadhlurrahman and Y. Wihardi, "Pengenalan Ekspresi Wajah Peserta Didik di Ruang Kelas Menggunakan Vision Transformer (ViT)," vol. 4, no. 2, pp. 1047–1058, 2025.
[14] N. S. Mahajaya, P. D. W. Ayu, and R. R. Huizen, "Pengaruh optimizer adam, adamw, sgd, dan lamb terhadap model vision transformer pada klasifikasi penyakit paru-paru," *Spinter 2024*, vol. 1, no. 2, pp. 818–823, 2024, [Online]. Available: https://spinter.stikom-bali.ac.id/index.php/spinter/article/view/222/188
[15] R. R. Ar, Agusriyati, and S. Moka, "Pendeteksian Dini Stunting Pada Balita Menggunakan Vision Transformer (VIT) Berbasis Citra Tubuh," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3S1, pp. 896–902, 2025, doi: 10.23960/jitet.v13i3S1.7888.
[16] S. Angara, N. R. Mannuru, A. Mannuru, and S. Thirunagaru, "A novel method to enhance pneumonia detec1on via a model-level ensembling of CNN and vision transformer," vol. 6, no. 1, pp. 21–28, 2024, doi: 10.48550/arXiv.2401.02358.
[17] T. B. Sasongko, H. Haryoko, and A. Amrullah, "Analisis Efek Augmentasi Dataset dan Fine Tune pada Algoritma Pre-Trained Convolutional Neural Network (CNN)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 4, pp. 763–768, 2023, doi: 10.25126/jtiik.20241046583.
[18] M. A. Mutasodirin and F. M. Falakh, "Efficient Weather Classification Using DenseNet and EfficientNet," *J. Inform. J. Pengemb. IT*, vol. 9, no. 2, pp. 173–179, 2024, doi: 10.30591/jpit.v9i2.7539.
[19] N. S. Mahajaya, Putu Desiana Wulaning Ayu, and Roy Rudolf Huizen, "Classification of Lung Diseases in X-Ray Images Using Transformer-Based Deep Learning Models," *J. Nas. Pendidik. Tek. Inform.*, vol. 13, no. 3, pp. 494–505, 2024, doi: 10.23887/janapati.v13i3.81425.
[20] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *CellPress*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.