A Fine-Tuned Transfer Learning Vision Transformer Framework for Lungs X-Ray Image Classification

I Gusti Ngurah Lanang Wijayakusuma [1]*, Made Sudarma [2], Ni Putu Dian Astutik [3]***
* Mathematics, Universitas Udayana
** Electrical Engineering, Universitas Udayana
lanang_wijaya@unud.ac.id [1], msudarma@unud.ac.id [2], dianastutik03@gmail.com [3]

Article Info

Article history:

Received 2025-11-27
Revised 2026-01-01
Accepted 2026-01-07

Keyword:

Artificial Intelligence,
Fine-tuning,
Transfer Learning,
Vision Transformer,
X-ray.

ABSTRACT

Lung diseases constitute a significant source of morbidity and therefore require diagnostic frameworks that provide both high accuracy and operational efficiency. This study proposes the development of a Vision Transformer (ViT)-based classification model for lung X-ray images, employing transfer learning and fine-tuning techniques to improve detection performance across five disease categories. Experimental results demonstrate stable and effective model convergence, as reflected by the consistent decrease in loss metrics throughout the learning process. Evaluation on an independent test dataset shows that the proposed approach achieves an accuracy of 0.958, indicating strong and balanced generalization performance. Further analysis using a confusion matrix reveals that the ViT model is capable of recognizing subtle and complex radiographic patterns with low misclassification rates, particularly achieving high recall for major pathological classes, which is critical for minimizing false negatives in clinical screening scenarios. Overall, this study demonstrates that the application of transfer learning with fine-tuning on a Vision Transformer architecture yields competitive performance for multi-class lung X-ray classification when trained on a balanced dataset. These findings are consistent with prior evidence highlighting the effectiveness of ViT in capturing global contextual information in medical imaging tasks.

I. Introduction

Lung diseases represent one of the most urgent public health challenges in Indonesia, exerting a substantial impact on the population. Conditions such as tuberculosis, pneumonia, and lung cancer are major causes of morbidity and mortality, particularly in regions with limited healthcare infrastructure. The shortage of radiology specialists and the unequal distribution of medical technology further exacerbate this issue. Therefore, innovative solutions capable of supporting accurate early diagnosis are essential to reducing mortality rates [1], [2], [3], [4], [5], [6].

During the COVID-19 pandemic, artificial intelligence (AI) gained prominence as a potential solution to the limitations of medical resources. AI-based methods, especially those applied to medical imaging modalities such as chest X-rays, demonstrate the ability to recognize complex patterns that are difficult for untrained observers to identify and typically require radiological expertise. With properly designed algorithms, AI can assist clinicians in detecting abnormalities more rapidly, accurately, and efficiently. This capability not only alleviates the workload of healthcare professionals but also accelerates clinical decision-making, ultimately improving patient outcomes [7], [8].

Modern approaches employing attention mechanisms, such as the Vision Transformer (ViT), have been extensively applied in medical image analysis and proven effective in extracting visual features. The ViT architecture employs self-attention to capture global relationships among various regions within an image. For lung X-rays, where subtle differences often distinguish normal from abnormal images, global context is crucial because abnormalities frequently arise from structural relationships across regions rather than isolated local features [9], [10].

Previous studies have demonstrated that Vision Transformer (ViT) models deliver competitive and reliable performance in a broad range of image classification tasks, with growing adoption in medical imaging analysis applications that involve subtle visual variations. Compared with conventional convolutional neural networks (CNNs), ViT offers an advantage in modeling long-range dependencies, which is beneficial for analyzing complex lung X-ray images, where disease indicators may not be localized to specific regions [9], [10], [11], [12], [13], [14].

Despite these advantages, ViT models are inherently data-intensive and, in most cases, demand large-scale training datasets to achieve optimal performance, which poses challenges when data availability is limited. Unlike CNNs, ViT architectures lack built-in inductive biases such as locality and translation invariance, which allow CNNs to learn efficiently from smaller datasets [15], [16], [17], [18]. As a result, ViT models trained from scratch on limited or medium-sized datasets are more susceptible to overfitting and may underperform compared to CNN-based approaches.

Another major challenge in lung X-ray analysis lies in the subtle and difficult-to-recognize nature of disease features. Early indicators of pneumonia or tuberculosis may appear as minor structural changes that are difficult to observe in grayscale images. Additionally, class imbalance within many datasets increases the complexity of the classification task, as certain disease categories occur far less frequently than others, affecting model performance in detecting minority classes [19], [20], [21], [22], [23].

Transfer learning has emerged as a practical approach for mitigating these constraints. By leveraging models that were previously trained on extensive, generic datasets, transfer learning enables the transfer of prior knowledge into specific domains, such as medical X-ray analysis. Fine-tuning, in which selected model layers are re-optimized using the target dataset, has shown substantial benefits, particularly when training data are limited or imbalanced. This approach minimizes the requirement to train a model from the ground up and accelerates model development [19], [24], [25], [26], [27], [28].

Recent research efforts have further demonstrated that the combination of ViT architectures with transfer learning can yield promising results in chest X-ray classification tasks. In parallel, data augmentation techniques—including image rotation, horizontal and vertical flipping, and contrast adjustment—are commonly employed to increase data diversity and enhance model robustness. When integrated with transfer learning, these strategies can help mitigate data limitations and improve classification performance [24], [29], [30], [31].

Although Vision Transformer models with transfer learning have been widely investigated for chest X-ray classification, most existing studies primarily focus on large-scale and well-curated datasets. Comparatively less attention has been given to systematic fine-tuning strategies and data preprocessing techniques, specifically aimed at addressing the challenges arising from limited and small-to-medium-sized datasets, which are commonly encountered in medical imaging research.

To address this gap, this study presents an empirical investigation of a fine-tuned transfer learning Vision Transformer framework for lung X-ray classification, with an emphasis on methodological robustness rather than direct clinical deployment. The proposed approach systematically examines fine-tuning strategies for ViT when applied to small- to medium-sized lung X-ray datasets, aiming to improve feature representation and classification reliability under realistic data constraints.

II. METHODS

This study follows a systematic workflow to ensure that each stage of data processing and model development is conducted in a structured and measurable manner. The workflow begins with data collection, followed by the initial pre-processing step to standardize the images and prepare inputs suitable for the Vision Transformer (ViT) architecture, and concludes with the analysis of the model using the optimal hyperparameter configuration. The workflow is illustrated in Figure 1.

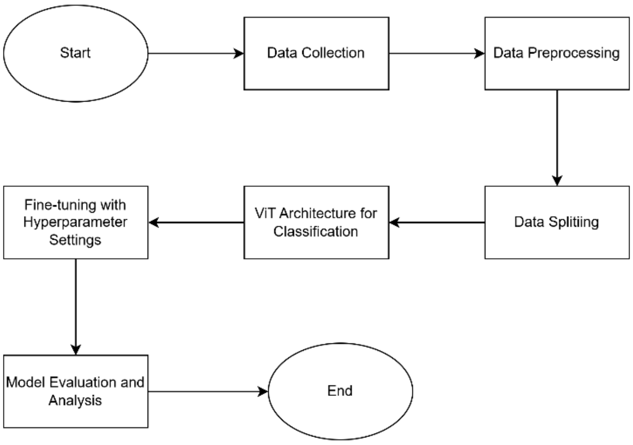

Start
Data Collection
Data Preprocessing
Data Splititing
ViT Architecture for Classification
Fine-tuning with Hyperparameter Settings
Model Evaluation and Analysis
End

Figure 1. Research Flow

A. Data Collection

In this study, the dataset employed consists of lung X-ray images categorized into four classes: COVID-19, Pneumonia, Pneumothorax, and Normal, comprising a total of 20.000 images. The data were collected by downloading and curating images from several publicly available datasets hosted on Kaggle, namely Chest X-ray (Covid-19 & Pneumonia), Chest X-Ray (Pneumonia, Covid-19, Tuberculosis), and the NIH Chest X-rays dataset. The dataset exhibits a balanced distribution among four classes, each represented by 5,000 images.

B. Data Pre-processing

For the purpose of assuring the accuracy and consistency of the inputs used during model training, a series of preprocessing procedures was applied. These procedures were designed to standardize image characteristics and reinforce the model's capacity to handle diverse input variations in lung X-ray imagery through several simple augmentation techniques. The steps performed in this study are as follows:

1) *Resize (256 pixels):*

All images were resized to 256 pixels on their shortest side. This ensures consistent image dimensions, facilitating batch processing and reducing computational complexity without removing diagnostically important pulmonary information.

2) *Center Crop (224 x 224):*

After resizing, images were centrally cropped to 224 × 224 pixels, which corresponds to the standard input dimensions required by Vision Transformer (ViT) models pretrained on ImageNet. This step preserves the most diagnostically relevant central lung region while reducing noise at the periphery, such as radiographic labels or background artifacts.

3) *Random Horizontal Flip:*

During training, the images were subjected to horizontal flipping with a predetermined probability. This augmentation enhances spatial invariance, enabling the model to learn abnormal patterns regardless of orientation.

4) *Random Rotation (±5 degrees):*

Images were subjected to small random rotations within the range of -5° to +5°, mimicking real-world variations in patient positioning. This improves the model's robustness to slight misalignments that do not affect clinical interpretation.

5) *Color Jitter (Brightness and Contrast):*

Image brightness and contrast were randomly adjusted within ±20%. This augmentation helps the model adapt to differences in lighting conditions and contrast variations across X-ray machines or acquisition environments, improving cross-domain generalization.

6) Convert to Tensor:

Since the study utilizes the PyTorch framework, all images were converted into PyTorch tensors with pixel intensities rescaled to a 0–1 interval. This step is required for processing data within deep learning pipelines.

7) Normalize:

The images were normalized by applying a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225], ensuring that pixel intensities were standardized for consistent input to the model, by following the ImageNet statistical distribution. Normalization aligns the statistical distribution of the X-ray inputs with that learned during the pretraining phase of the ViT model, which promotes faster convergence and more stable training.

C. ViT Architecture for Lung X-Ray

The Vision Transformer (ViT) architecture used in this study is a transformer-based model fine-tuned to perform diagnostic classification of lung X-ray scans into four categories. In general, this architecture features two fundamental components: the X-ray ViT backbone and the multilayer perceptron (MLP) classifier, as shown in Figure 2.

In the backbone, the process begins with patch embedding, where an input image with three color channels is divided into small patches of size 16 × 16 pixels. Each image patch is subsequently transformed into a 768-dimensional embedding vector through a 2D convolution operation employing a 16 × 16 kernel and corresponding stride. Mathematically, the number of patches produced is (H/16 × W/16) for an image of size H × W × 3. This linear projection serves as a substitute for the spatial encoding produced by CNNs, embedding each patch into a latent representation.

The integrated patches are subsequently processed by a transformer encoder with 12 stacked layers. Each transformer encoder layer is architecturally composed of two primary submodules: a feed-forward network (FFN) and a multi-head self-attention mechanism.

. Within the self-attention mechanism, the input embedding $X \in R^{N \times d}$, where NDenotes the total number of patches plus a class token and d corresponds to the embedding dimension (768), is linearly transformed into three matrices, referred to as Query (Q), Key (K), and Value (V). Based on these representations, the attention scores are then derived as expressed in Equation (1):

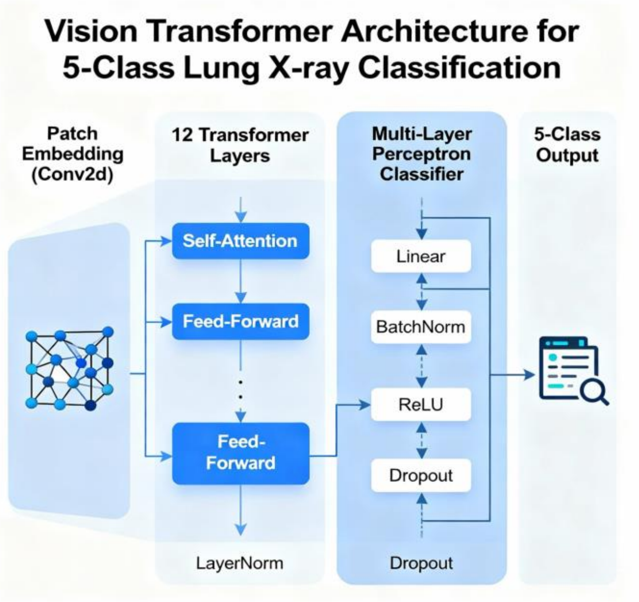


Figure 2. ViT Architecture for Lung X-ray

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

The output is then linearly projected. By employing self-attention, the model effectively encodes global relationships spanning across the image patches, offering an advantage over CNNs, which primarily capture local features. Meanwhile, the FFN applies non-linear transformations with an intermediate dimension of 3072, increasing the model's representational capacity.

In the classifier section, the ViT is connected to a fully connected multilayer perceptron (MLP) consisting of several linear layers. The first layer reduces the embedding dimension from 768 to 512 units, followed by batch normalization for stabilization. A ReLU activation function introduces non-linearity into the feature mapping.

D. Fine-Tuning with Hyperparameter Settings

As shown in Table I, the hyperparameter values were selected through a series of trial-and-error experiments aimed at achieving the most stable and consistent validation performance. In addition to tuning optimization parameters, the fine-tuning strategy also involved determining which components of the Vision Transformer architecture should be updated during training. The model employs a ViT-B/16 backbone consisting of a patch embedding layer followed by 12 transformer encoder layers and a task-specific classification head.

TABLE I
HYPERPARAMETER SETTINGS

Hyperparameter	**Value**
Learning rate	2e-4
Train/eval batch size	16:16
Epochs	12
Ratio train/val/test	85:10:5

During fine-tuning, all transformer encoder layers and the classification head were unfrozen, allowing the model to fully adapt pretrained representations to the lung X-ray domain. This full fine-tuning approach was empirically found to outperform partial layer freezing in preliminary experiments, particularly in capturing high-level contextual features relevant to medical image classification. The classification head was implemented as a multi-layer fully connected network with batch normalization, ReLU activation, and dropout regularization to enhance discriminative capability while mitigating overfitting.

With respect to optimization, a learning rate of 2×10^{-4} achieved an optimal balance between convergence speed and gradient stability. Throughout the training phase, a batch size of 16 was used to ensure stable gradient updates, while a batch size of 16 was applied during evaluation to improve computational efficiency. Training was conducted for 12 epochs, which was sufficient to reach convergence without observable overfitting on the validation set. Data splitting was performed using an 85:10:5 ratio for training, validation, and testing subsets, respectively, ensuring adequate data for model learning while maintaining objective performance evaluation. This combination of fine-tuning strategy and hyperparameter configuration yielded the highest validation performance compared to alternative settings explored during preliminary experiments.

E. Model Evaluation

The model was evaluated using a framework that examines multiple performance aspects. A confusion matrix served as the primary diagnostic tool, providing insights into model behavior, error patterns, and performance on real-world data. The evaluation process categorized model outputs into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), each representing outcomes with distinct clinical implications.

Accuracy represents a general indication of model performance. Nevertheless, in medical imaging applications such as lung X-ray classification, reliance on alone can be misleading when class frequencies are imbalanced. A model may appear highly accurate simply by predicting the majority class—such as “normal”—while failing to detect less frequent but clinically critical abnormalities. Precision provides a more clinically meaningful perspective by indicating how many of the model’s positive predictions truly correspond to pathological findings.

Next recall, this metric is particularly crucial in healthcare and clinical decision-support systems, as a low recall indicates a higher likelihood of false negative predictions, where existing abnormalities are overlooked by the model. In the context of lung X-ray analysis, such missed detections may involve clinically significant conditions, including pneumonia, pneumothorax, or other pulmonary abnormalities, which require timely medical intervention.

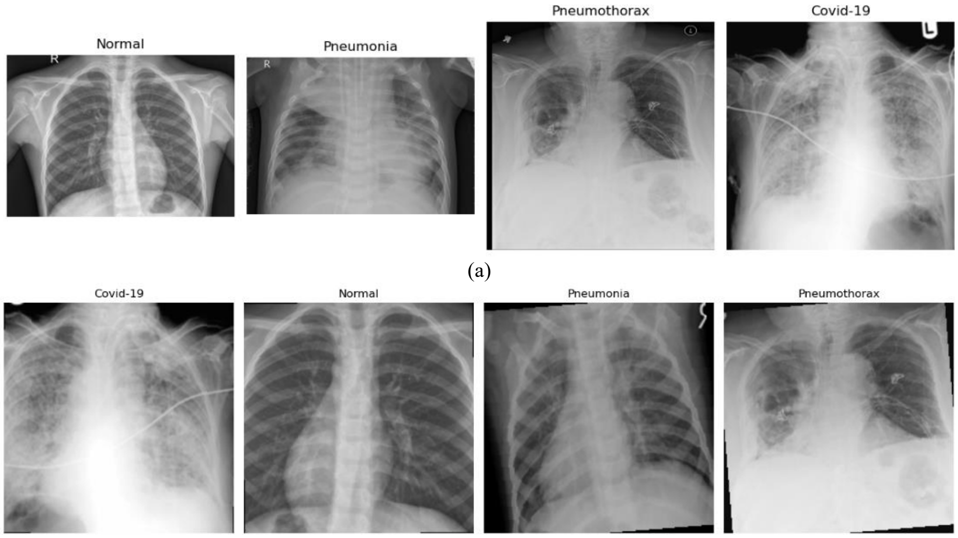
The F1-score integrates both precision and recall through their harmonic mean, producing a balanced metric that captures the trade-off between avoiding false alarms and minimizing missed diagnoses. In the context of lung X-ray analysis, the F1-score thus offers a more realistic and clinically aligned assessment of the model’s diagnostic capability, especially when both over-detection and under-detection carry significant consequences for patient care.

III. RESULTS AND DISCUSSION

This section presents the evaluation results of the fine-tuned Vision Transformer (ViT) model developed for multi-class lung X-ray classification, using Accuracy, Precision, Recall, and F1-Score as the primary performance indicators.

Before the fine-tuning process, all X-ray images underwent a preprocessing stage aimed at ensuring data consistency and accelerating model convergence. Examples of the images before preprocessing are illustrated in Figure 3(a), while the corresponding results after preprocessing are shown in Figure 3(b).

The overall training trajectory demonstrated stable and progressive learning, characterized by a consistent decrease in both training and validation loss, indicating that the model converged effectively without exhibiting signs of overfitting. As illustrated in Figure 4, this downward trend reflects the model’s increasing ability to generalize across samples, while Figure 5 highlights notable improvements across all evaluated metrics as training progressed.



(a)

(b)

Figure 3. Image (a) Before Preprocessing, (b) After Preprocessing

The reduction in training loss provides additional evidence of successful optimization. Starting at 0.1581 during the first epoch, the loss steadily declined to 0.0202 by the seventh epoch, slightly increased at the eighth epoch, but continued to decrease thereafter until reaching a minimum value at the twelfth epoch, demonstrating stable convergence behavior.

The validation loss also exhibited a general downward trend, decreasing from 0.0462 to 0.0092, although minor fluctuations were observed at epoch 3 (0.0519). Such early fluctuations are commonly encountered during the optimization process as the model adjusts its parameters. After epoch 8, the validation loss stabilized below 0.1, indicating strong generalization capability and the absence of significant overfitting.

The training performance across all evaluation metrics is presented in Figure 5, which shows that metric values began at 0.986 in the first epoch and gradually converged to 0.998, with a total training time of 37.13 minutes. These findings demonstrate that the fine-tuned model achieved near-optimal performance on the training dataset, as evidenced by its efficient convergence behavior and stable learning process. The strong results observed during training indicate that the selected fine-tuning strategy and hyperparameter configuration were effective in enabling the model to learn representative features from the data.

Nevertheless, it should be emphasized that metrics derived exclusively from the training data provide only a partial representation of the model's actual predictive capacity and cannot adequately characterize its behavior when exposed to unfamiliar inputs. Performance observed during training may be influenced by data memorization rather than genuine learning. To establish a more trustworthy and unbiased estimation of the model's generalization capability, a separate evaluation was therefore performed using an independent dataset that remained entirely excluded from both the training and validation processes. The quantitative performance metrics obtained from this evaluation are presented in Table 2.

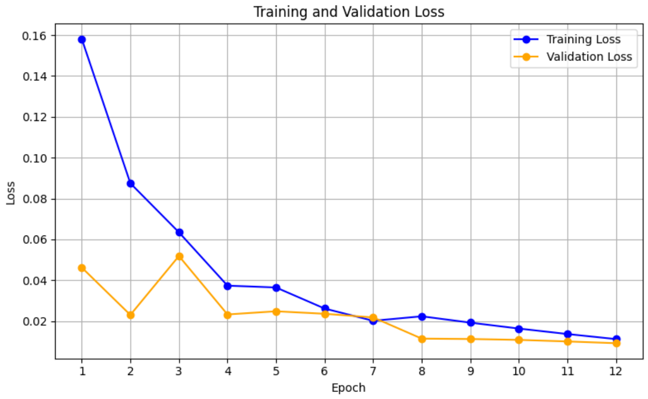


Figure 4. Training vs Validation Loss for ViT X-ray Classification

Evaluation on the test dataset reveals a pattern of stable and reliable predictive behavior, with an overall accuracy of 0.958, corresponding to 95.8% of the samples being assigned to their correct categories. The recorded precision of 0.959 indicates a high level of confidence in the model's positive predictions, suggesting that identified abnormal cases are largely trustworthy and unlikely to result in unwarranted clinical actions.

TABLE II
MODEL EVALUATION ON TEST DATA

Test Metrics			
Accuracy	Precision	Recall	F1
0,958	0,959	0,958	0,9582

Meanwhile, the recall value of 0.958 demonstrates strong sensitivity in capturing relevant cases across all classes, reducing the likelihood of overlooked pathological findings. The resulting F1-score of 0.9582 provides further evidence of consistent performance across complementary evaluation perspectives, reinforcing the robustness of the model when applied to previously unseen data that were not included during either training or validation.

TABLE III
METRICS PERFORMANCE ON EACH CLASS

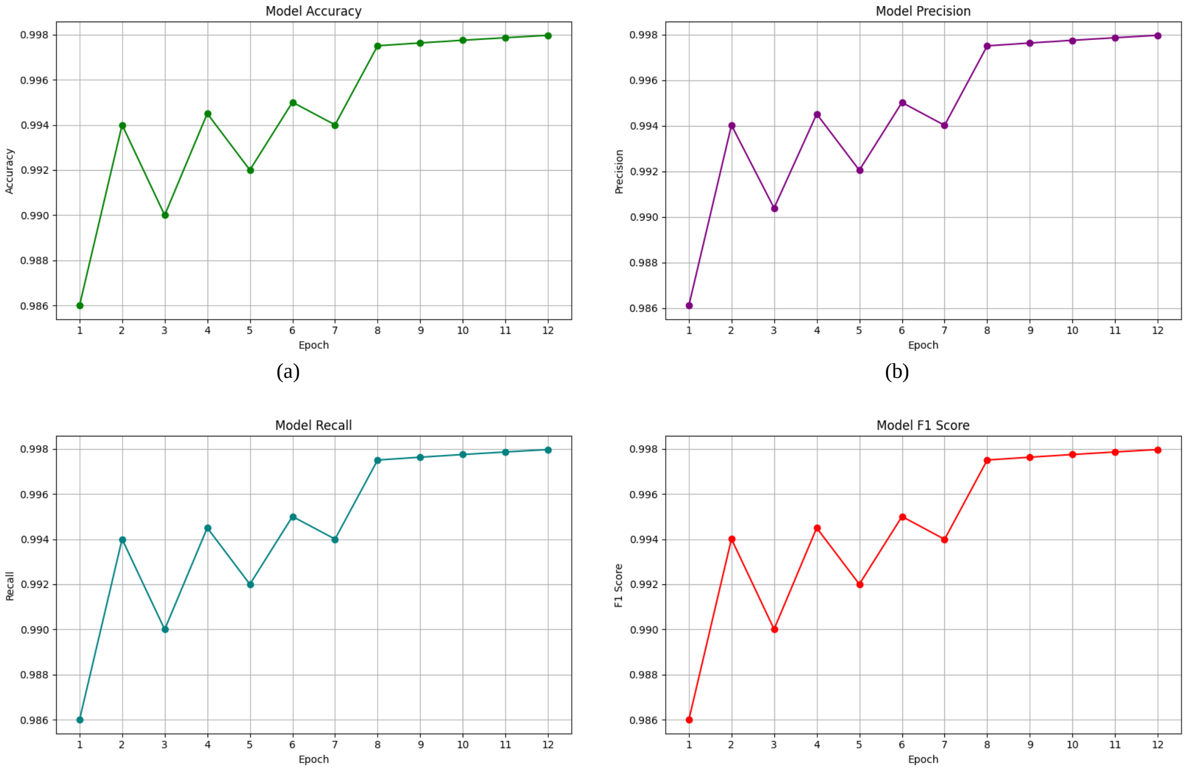
	Precision	Recall	F1-Score	Support
Covid-19	0,98	0,99	0,99	260
Pneumonia	0,97	0,99	0,98	234
Pneumothorax	0,93	0,96	0,94	249
Normal	0,95	0,90	0,92	257

To enable a more thorough interpretation of the evaluation results, the class-wise performance of the model on the test dataset was examined using the confusion matrix of the Vision Transformer (ViT), as shown in Figure 6, for the four-class X-ray classification task comprising COVID-19, Pneumonia, Pneumothorax, and Normal. The corresponding class-specific performance metrics are summarized in Table 3.

In the first class category, namely Covid-19, out of 260 images, only 257 were correctly classified, while only 2 samples were incorrectly predicted as Normal, and 1 sample was misclassified as Pneumothorax. The resulting recall score of 0.99 reflects the model's strong ability to detect Covid-19 cases, indicating a minimal occurrence of false negative predictions. This minimizes the likelihood of infected patients being undetected, thereby supporting effective infection control and timely clinical decision-making.

In the Pneumonia class, out of 234 images, 231 were correctly classified, while 2 images were misclassified as Normal and 1 image as COVID-19. The high recall value demonstrates the model's reliable capability in detecting pneumonia, which is essential for preventing delayed diagnosis and subsequent clinical complications.

For the Pneumothorax class, 239 out of 249 images were correctly classified. Most misclassifications occurred when pneumothorax cases were predicted as Normal. Although a small number of false negatives remain, this performance indicates the model's potential as an effective early screening tool to assist radiologists in identifying critical cases.



(a)

(b)

(c)

(d)

Figure 5. Performance of (a) Accuracy, (b) Precision, (c) Recall, (d) F1-Score

. In the Normal class, only 231 out of 260 images were correctly classified, indicating relatively lower performance compared to the pathological classes. Misclassifications predominantly occurred as Pneumothorax (18 cases) and, to a lesser extent, as Pneumonia. The lower recall value (0.90) suggests that some normal lung images were incorrectly identified as pathological, resulting in false positives. In clinical practice, this may lead to unnecessary follow-up examinations; however, from a patient safety perspective, such errors are generally more tolerable than false negatives.

The macro-averaged metrics (Precision, Recall, and F1-score = 0.96) indicate that the model delivers balanced performance across all classes without significant bias toward any specific category. Additionally, the consistent weighted average metrics (0.96) confirm that the model maintains stable performance despite variations in sample size among classes.

Overall, the confusion matrix and classification report results demonstrate that the Vision Transformer model for lung X-ray classification exhibits strong and balanced diagnostic capability. The high recall values achieved for major disease classes—COVID-19, pneumonia, and pneumothorax—are particularly relevant in clinical settings, as they minimize the risk of missing pathological cases.

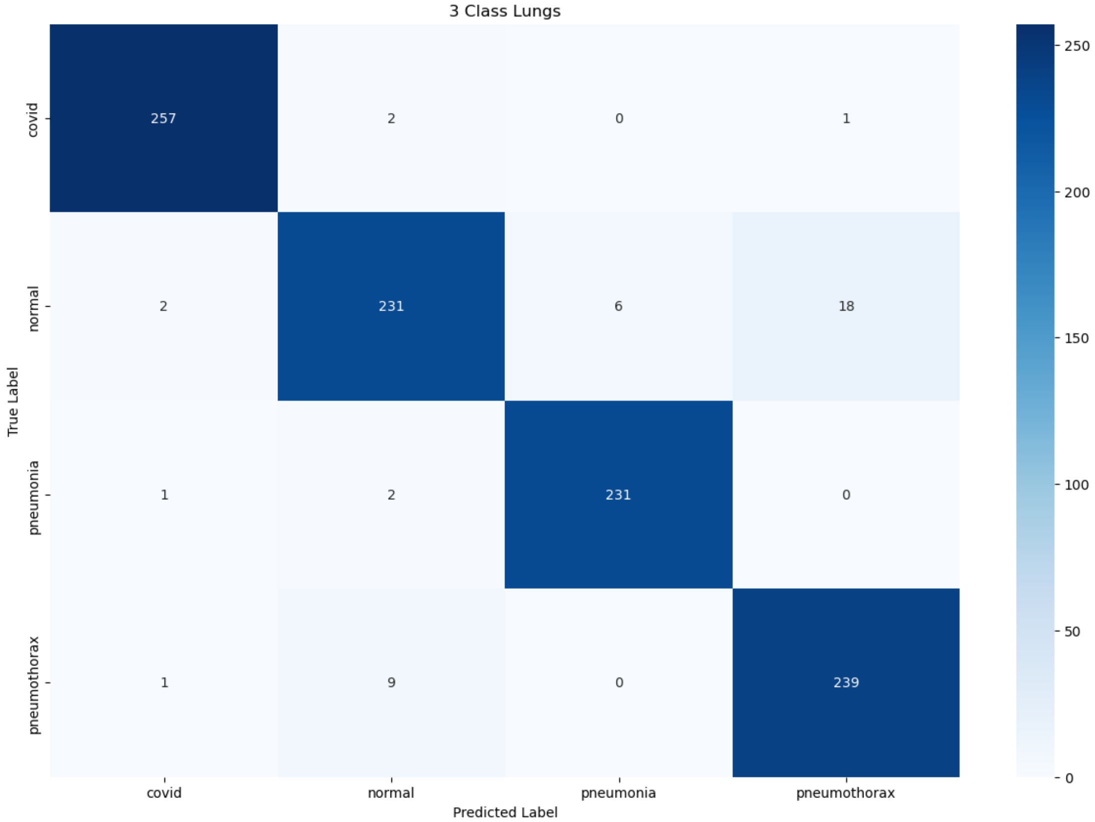


Figure 6. Confusion Matrix of ViT Lung X-Ray Classification

Subsequently, a quantitative comparison between the proposed approach and several established baseline models is summarized in Table IV, enabling a clear assessment of relative performance.

TABLE IV
STUDY COMPARISON

Study	Methods	Dataset	Accuracy
[32]	VGG16	Kaggle's pneumonia detection dataset (5.856 images)	95,4%
[33]	EfficientNet + Noisy Student	Combined 3 datasets (25.966 images)	86,6%
[9]	VIT	NIH ChestXray 14 dataset (112.120 images)	83,4%
Our	VIT	20.000 images	95,8%

Table IV summarizes the performance comparison between the proposed model and prior studies that employ different deep learning architectures, varying dataset sizes, and heterogeneous data sources. Study [32] utilized the VGG16 architecture on the Kaggle pneumonia dataset, consisting of 5,856 images, and achieved an accuracy of 95.4%, demonstrating that conventional CNN architectures can still deliver competitive performance on relatively small and focused datasets. In contrast, the study [33], which applied EfficientNet with a Noisy Student approach on a larger combined dataset of 25,966 images, achieved an accuracy of only 86.6%, suggesting that increased data complexity and source heterogeneity may negatively affect model stability.

In Study [9], a Vision Transformer (ViT) model was evaluated using the NIH ChestXray14 dataset comprising more than 112,120 images, yet it achieved an accuracy of 83.4%. Despite the large dataset size, the complex class distribution and diverse pathological labels are suspected to have contributed to the reduced overall classification performance. Conversely, the proposed Vision Transformer–based model, trained and fine-tuned on 20,000 lung X-ray images, achieved the highest accuracy of 95.8%. These findings demonstrate that an appropriate transfer learning and fine-tuning strategy, combined with effective data management, can yield superior performance even without extremely large-scale datasets.

Overall, the results presented in Table IV indicate that the proposed Vision Transformer–based model achieves competitive, and in several cases superior, performance compared to existing baseline studies, despite being trained on a comparatively moderate-sized dataset. This suggests that the adopted fine-tuning strategy and data preprocessing pipeline play a crucial role in maximizing model effectiveness, particularly in medical imaging scenarios where data availability and class distribution are often constrained.

Nonetheless, there remains room for further improvement. Future work may explore more efficient Vision Transformer variants, employ generative model–based data augmentation, evaluate cross-hospital generalization, and integrate the system into real-world clinical workflows to assess operational performance. These directions are expected to enhance model reliability and expand its applicability in modern radiological practice.

IV. Conclusion

This study demonstrates that the application of a transfer learning framework combined with systematic fine-tuning on the Vision Transformer (ViT) architecture results in strong and stable performance for multi-class lung X-ray classification. The presented configuration model attained an overall test accuracy of 95.8%, accompanied by consistently strong precision, recall, and F1-score values, indicating robust generalization to previously unseen data. Notably, the model exhibited particularly high recall for clinically critical classes, including COVID-19, pneumonia, and pneumothorax, highlighting its effectiveness in identifying pathological conditions that require timely clinical intervention.

The class-wise analysis based on the confusion matrix further confirms the model’s capability to capture complex radiographic patterns across different lung conditions. The low rate of false negatives observed in disease classes is especially significant in a clinical context, as it reduces the risk of missed diagnoses. Although relatively higher misclassification rates were observed in the Normal class—primarily due to confusion with pneumothorax and pneumonia—such errors predominantly result in false positives, which are generally more tolerable in medical screening scenarios than false negatives. Collectively, these findings imply that the proposed ViT model is well-suited as an early screening or decision-support tool to assist radiologists in clinical practice.

When compared with existing baseline studies, the proposed model achieved competitive and, in several cases, superior performance despite being trained on a moderately sized dataset. This outcome underscores the importance of effective data preprocessing and fine-tuning strategies in maximizing the performance of Vision Transformer models, particularly in medical imaging tasks where data availability and class distribution are often constrained.

Nevertheless, the findings of this study should be interpreted with caution. Although the model was evaluated using combined public datasets with a balanced class distribution, these datasets may not fully reflect real-world clinical imaging conditions. Variations in image acquisition protocols, equipment, and patient demographics across clinical settings may limit the generalizability of the proposed model when deployed in heterogeneous healthcare environments. Therefore, future investigations are encouraged to focus on validating the proposed approach using multi-center and cross-institutional datasets that better capture real clinical variability, as well as assessing model performance within operational clinical workflows. Such efforts are necessary to ensure the reliability, robustness, and practical applicability of Vision Transformer–based systems in modern radiological practice.

Acknowledgement

This research was supported by the Faculty of Mathematics and Natural Sciences, Udayana University, through the Program Unggulan Program Studi (PUPS) research scheme, under grant number B/229.127/UN14.4.A/PT.01.03/2025. The authors express their sincere gratitude for the funding and institutional support that made this work possible.

References

[1] E. Bhattacharya and D. Bhattacharya, “A Review of Recent Deep Learning Models in COVID-19 Diagnosis,” *European Journal of Engineering and Technology Research*, vol. 6, no. 5, 2021.

[2] E. Prompetchara, C. Ketloy, and T. Palaga, “Allergy and Immunology Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic,” *Asian Pasific Journal of Allergy and Immunology*, 2020, doi: 10.12932/AP-200220-0772.

[3] A. S. Simbirtsev, “Immunopathogenesis and perspectives for immunotherapy of coronavirus infection,” *HIV Infection and Immunosuppressive Disorders*, vol. 12, no. 4, pp. 7–22, 2020, doi: 10.22328/2077-9828-2020-12-4-7-22.

[4] G. Li *et al.*, “Coronavirus infections and immune responses,” *J Med Virol*, vol. 92, no. 4, pp. 424–432, Apr. 2020, doi: 10.1002/JMV.25685.

[5] H. Mary Shyni and E. Chitra, “A comparative study of X-ray and CT images in COVID-19 detection using image processing and deep learning techniques,” *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100054, Jan. 2022, doi: 10.1016/J.CMPBUP.2022.100054.

[6] H. Mohammad-Rahimi, M. Nadimi, A. Ghalyanchi-Langeroudi, M. Taheri, and S. Ghafouri-Fard, “Application of Machine Learning in Diagnosis of COVID-19 Through X-Ray and CT Images: A Scoping Review,” *Front Cardiovasc Med*, vol. 8, Mar. 2021, doi: 10.3389/FCVM.2021.638011/FULL.

[7] A. U. Haq, J. P. Li, S. Ahmad, S. Khan, M. A. Alshara, and R. M. Alotaibi, “Diagnostic Approach for Accurate Diagnosis of COVID-19 Employing Deep Learning and Transfer Learning Techniques through Chest X-ray Images Clinical Data in E-Healthcare,” *Sensors 2021, Vol. 21, Page 8219*, vol. 21, no. 24, p. 8219, Dec. 2021, doi: 10.3390/S21248219.

[8] S. Hassantabar, M. Ahmadi, and A. Sharifi, “Diagnosis and detection of infected tissue of COVID-19 patients based on lung x-

ray image using convolutional neural network approaches," *Chaos Solitons Fractals*, vol. 140, p. 110170, Nov. 2020, doi: 10.1016/J.CHAOS.2020.110170.

[9] L. Huang, J. Ma, H. Yang, and Y. Wang, "Research and implementation of multi-disease diagnosis on chest X-ray based on vision transformer," *Quant Imaging Med Surg*, vol. 14, no. 3, pp. 2539–2555, Mar. 2024, doi: 10.21037/QIMS-23-1280/COIF.

[10] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," *Sci Rep*, vol. 14, no. 1, p. 2487, Dec. 2024, doi: 10.1038/S41598-024-52703-2.

[11] Ş. Öztürk, M. Y. Turalı, and T. Çukur, "HydraViT: Adaptive Multi-Branch Transformer for Multi-Label Disease Classification from Chest X-ray Images," *Biomed Signal Process Control*, vol. 100, Oct. 2023, doi: 10.1016/j.bspc.2024.106959.

[12] S. Ghosh, A. Bandyopadhyay, M. Bose, and K. C. Santosh, "Vision Transformers Excel in Chest X-Ray Analysis," *Proceedings - 2025 IEEE Conference on Artificial Intelligence, CAI 2025*, pp. 495–500, 2025, doi: 10.1109/CAI64502.2025.00090.

[13] S. Regmi, A. Subedi, U. Bagci, D. Jha, and P. Campus, "Vision Transformer for Efficient Chest X-ray and Gastrointestinal Image Classification," p. 111, Apr. 2023, doi: 10.1117/12.3045810.

[14] O. Uparkar, J. Bharti, R. K. Pateriya, R. K. Gupta, and A. Sharma, "Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images," *Procedia Comput Sci*, vol. 218, pp. 2338–2349, Jan. 2023, doi: 10.1016/J.PROCS.2023.01.209.

[15] Y. Shen *et al.*, "MoViT: Memorizing Vision Transformers for Medical Image Analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14349 LNCS, pp. 205–213, Mar. 2023, doi: 10.1007/978-3-031-45676-3_21.

[16] B. Zhang and Y. Zhang, "MSCViT: A Small-size ViT architecture with Multi-Scale Self-Attention Mechanism for Tiny Datasets," Jan. 2025, Accessed: Nov. 26, 2025. [Online]. Available: https://arxiv.org/pdf/2501.06040

[17] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020, Accessed: Nov. 26, 2025. [Online]. Available: https://arxiv.org/pdf/2010.11929

[18] M. J. Horry, S. Chakraborty, B. Pradhan, N. Shulka, and M. Almazroui, "Two-Speed Deep-Learning Ensemble for Classification of Incremental Land-Cover Satellite Image Patches," *Earth Systems and Environment 2023 7:2*, vol. 7, no. 2, pp. 525–540, Mar. 2023, doi: 10.1007/S41748-023-00343-3.

[19] K. S. Charan, O. V. Krishna, P. V. Sai, and A. K. Ilavarasi, "Transfer Learning Based Multi-Class Lung Disease Prediction Using Textural Features Derived From Fusion Data," *IEEE Access*, vol. 12, pp. 108248–108262, 2024, doi: 10.1109/ACCESS.2024.3435680.

[20] P. Misra, N. Panigrahi, S. Gopal Krishna Patro, A. O. Salau, and S. S. Aravinth, "PETLFC: Parallel ensemble transfer learning based framework for COVID-19 differentiation and prediction using deep convolutional neural network models," *Multimedia Tools and Applications 2023 83:5*, vol. 83, no. 5, pp. 14211–14233, Jul. 2023, doi: 10.1007/S11042-023-16084-4.

[21] K. Rajagopalan and S. Babu, "The detection of lung cancer using massive artificial neural network based on soft tissue technique," *BMC Medical Informatics and Decision Making 2020 20:1*, vol. 20, no. 1, pp. 282-, Oct. 2020, doi: 10.1186/S12911-020-01220-Z.

[22] A. Victor Ikechukwu and S. Murali, "CX-Net: an efficient ensemble semantic deep neural network for ROI identification from chest-x-ray images for COPD diagnosis," *Mach Learn Sci Technol*, vol. 4, no. 2, p. 025021, May 2023, doi: 10.1088/2632-2153/ACD2A5.

[23] K. Wang, X. Zhang, S. Huang, F. Chen, X. Zhang, and L. Huangfu, "Learning to Recognize Thoracic Disease in Chest X-Rays with Knowledge-Guided Deep Zoom Neural Networks," *IEEE Access*, vol. 8, pp. 159790–159805, 2020, doi: 10.1109/ACCESS.2020.3020579.

[24] M. Liu, L. Dong, Q. Jiao, C. Gu, and M. Lee, "Deep Transfer Learning Using Real-World Image Features for Medical Image Classification, with a Case Study on Pneumonia X-ray Images," *Bioengineering 2024, Vol. 11, Page 406*, vol. 11, no. 4, p. 406, Apr. 2024, doi: 10.3390/BIOENGINEERING11040406.

[25] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval 2021 11:1*, vol. 11, no. 1, pp. 19–38, Sep. 2021, doi: 10.1007/S13735-021-00218-1.

[26] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "DeTrac: Transfer Learning of Class Decomposed Medical Images in Convolutional Neural Networks," *IEEE Access*, vol. 8, pp. 74901–74913, 2020, doi: 10.1109/ACCESS.2020.2989273.

[27] G. H. Huang, Q. J. Fu, M. Z. Gu, N. H. Lu, K. Y. Liu, and T. B. Chen, "Deep Transfer Learning for the Multilabel Classification of Chest X-ray Images," *Diagnostics*, vol. 12, no. 6, p. 1457, Jun. 2022, doi: 10.3390/DIAGNOSTICS12061457/S1.

[28] Z. Alammar, L. Alzubaidi, J. Zhang, Y. Li, W. Lafta, and Y. Gu, "Deep Transfer Learning with Enhanced Feature Fusion for Detection of Abnormalities in X-ray Images," *Cancers 2023, Vol. 15, Page 4007*, vol. 15, no. 15, p. 4007, Aug. 2023, doi: 10.3390/CANCERS15154007.

[29] M. A. Sufian *et al.*, "AI-Driven Thoracic X-ray Diagnostics: Transformative Transfer Learning for Clinical Validation in Pulmonary Radiography," *Journal of Personalized Medicine 2024, Vol. 14, Page 856*, vol. 14, no. 8, p. 856, Aug. 2024, doi: 10.3390/JPM14080856.

[30] R. Fan and S. Bu, "Transfer-Learning-Based Approach for the Diagnosis of Lung Diseases from Chest X-ray Images," *Entropy 2022, Vol. 24, Page 313*, vol. 24, no. 3, p. 313, Feb. 2022, doi: 10.3390/E24030313.

[31] P. K. Pagadala, S. L. Pinapatruni, C. R. Kumar, S. Katakam, L. S. K. Peri, and D. A. Reddy, "Enhancing Lung Cancer Detection from Lung CT Scan Using Image Processing and Deep Neural Networks," *Revue d'Intelligence Artificielle*, vol. 37, no. 6, pp. 1597–1605, Dec. 2023, doi: 10.18280/RIA.370624.

[32] S. Sharma and K. Guleria, "A Deep Learning based model for the Detection of Pneumonia from Chest X-Ray Images using VGG-16 and Neural Networks," *Procedia Comput Sci*, vol. 218, pp. 357–366, Jan. 2023, doi: 10.1016/J.PROCS.2023.01.018.

[33] M. Nishio *et al.*, "Deep learning model for the automatic classification of COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy: a multi-center retrospective study," *Sci Rep*, vol. 12, no. 1, p. 8214, Dec. 2022, doi: 10.1038/S41598-022-11990-3.