Journal of Deep Learning, Computer Vision and Digital Image Processing
E-ISSN: 2986-8939 P-ISSN: 2986-8920
Vol. 4, No. 1, March 2026
DOI. https://doi.org/10.61255/decoding.v4i1.1277

Vision Transformer Transfer Learning for Automated COVID-19 Pneumonia Detection in Chest Radiography Using Binary Classification

Vina Annisa Sofyan✉
Universitas Negeri Makassar, Makassar, Indonesia

Nur Amalia Ramadhan
Universitas Negeri Makassar, Makassar, Indonesia

Marwan Ramdhany Edy
Universitas Negeri Makassar, Makassar, Indonesia

Ardiansah
Universitas Negeri Makassar, Makassar, Indonesia

Andini Noviyanti Fitriani
Universitas Negeri Makassar, Makassar, Indonesia

Syahril Ramadhan
Universitas Negeri Makassar, Makassar, Indonesia

OPEN ACCESS

ARTICLE HISTORY
Received: January 22, 2026
Revised: February 15, 2026
Accepted: March 18, 2026

KEYWORDS
Binary Classification; COVID-19 Pneumonia; Chest Radiography; Transfer Learning; Vision Transformer.

ABSTRACT

Purpose – Automated detection of COVID-19-associated pneumonia from chest radiography can support clinical decision-making when radiographic interpretation is challenged by limited data and image variability. This study evaluates a Vision Transformer (ViT) transfer learning model for binary classification of chest X-ray images into COVID with pneumonia and COVID without pneumonia classes.

Methods – A publicly accessible COVID-19 radiography dataset containing 513 chest X-ray images was used. Images were resized to 224 × 224 pixels, converted into tensors, and normalized using ImageNet parameters. A pre-trained ViT model from the timm library was fine-tuned for binary classification. The data were divided into stratified training, validation, and testing subsets, and performance was assessed using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

Findings – The confusion matrix consisted of 43 true positives, 3 false positives, 35 true negatives, and 0 false negatives, representing 81 testing images. Based on these values, the model achieved an accuracy of 96.30%, precision of 0.93, recall of 1.00, F1-score of 0.97, and ROC-AUC of 1.00. These findings indicate that the ViT transfer learning model demonstrated high sensitivity in detecting COVID with pneumonia cases, although three false positive predictions were still observed in the COVID without pneumonia class.

Research implications – The study offers a cautious evaluation of ViT transfer learning for COVID-19 pneumonia classification using a limited dataset. Larger multi-site datasets, cross-validation, and CNN baseline comparisons are required to strengthen clinical generalizability.

Originality – This study is positioned as an evaluation of ViT transfer learning for COVID-19 pneumonia classification rather than as a ViT-based data augmentation method.

Correspondence Author: ✉vinaannisasofyan7@gmail.com
To cite this article : Sofyan, V. A., Ramadhan, N. A., Edy, M. R., Ardiansah, Fitriani, A. N., & Ramadhan, S. (2026). Vision transformer transfer learning for automated COVID-19 pneumonia detection in chest radiography using binary classification. *Journal of Deep Learning, Computer Vision and Digital Image Processing*, *4*(1), 117-128. Doi. https://doi.org/10.61255/decoding.v4i1.1277

This is an open access article under the CC BY-SA license

INTRODUCTION

The accelerating pace of technological evolution across multiple sectors, healthcare in particular, has established digital imaging as an indispensable pillar of contemporary clinical workflows. Chest radiography, as one of the most widely employed medical imaging modalities, serves as a primary investigative tool for identifying a range of pulmonary disorders, including pneumonia-a condition routinely screened through radiographic examination. Despite its widespread clinical adoption, however, the efficacy of radiographic analysis is compromised by persistent constraints in image quality. Degradation stemming from insufficient spatial resolution, compression-related data loss, hardware inadequacies, signal interference, and acquisition artifacts collectively impair diagnostic precision [1][2]. At the image level, reduced resolution (LR) tends to obscure critical pathological markers, impeding disease identification and complicating clinical interpretation [3][4]. Compounding this challenge is the structural imbalance in sample distribution inherent to most chest X-ray repositories, where pneumonia cases occupy a disproportionately smaller fraction of the dataset relative to normal scans-a condition that may systematically bias learning algorithms and curtail their effectiveness in correctly identifying minority-class instances [5].

In response to these quality-related constraints, super-resolution imaging techniques have gained significant traction as a computationally viable solution, reconstructing high-fidelity, high-resolution (HR) representations from degraded low-resolution (LR) inputs. The landscape of image quality enhancement in medical contexts spans a broad methodological spectrum, from classical interpolation algorithms to contemporary deep learning frameworks. Among these, Convolutional Neural Networks (CNNs) have emerged as the predominant computational paradigm for thoracic pathology recognition including pneumonia identification from chest radiographs, owing to their exceptional proficiency in hierarchical automated feature extraction [6]. Nonetheless, the localized receptive field inherent to convolutional operations imposes a fundamental constraint: CNN architectures primarily rely on local receptive fields, which may limit their ability to directly capture long-range spatial relationships across distant image regions, a limitation that proves consequential when analyzing the complex, spatially distributed patterns present in diverse medical images. Under conditions of imaging noise or acquisition-induced degradation routinely encountered in clinical environments this architectural shortcoming manifests as measurable accuracy deterioration [7][8].

Bridging the representational gap inherent in convolution-based systems, Vision Transformer (ViT) introduces a fundamentally different computational strategy: rather than scanning localized regions sequentially, ViT partitions the input image into a structured sequence of fixed-size patch tokens, each projected into an embedding space before being jointly processed by multi-head self-attention modules that can relate any two spatial positions simultaneously [9]. This token-level attention formulation grants the model access to image-wide contextual cues that are structurally inaccessible to convolutional architectures, and comparative evaluations have confirmed that this property translates into measurable classification advantages across diverse image recognition benchmarks, including medical image datasets [10]. In the specific context of chest radiographic pneumonia detection, global attention proves especially valuable: pneumonia-related opacity and consolidation often manifest across spatially scattered lung regions, and ViT's capacity to jointly process these distributed cues enables it to construct more diagnostically informative feature representations than architectures confined to local receptive windows [11]. Equally important in clinical imaging contexts is ViT's demonstrated resilience under limited training data conditions and its comparative stability when class distributions are skewed both of which are persistent realities in medical dataset construction [12][13].

The empirical literature on ViT applications in medical imaging has yielded converging evidence of the architecture's diagnostic potential. In the context of pneumonia screening specifically, ViT-based systems have demonstrated measurably superior classification metrics compared to convolutional baselines [14], with reported improvements in both sensitivity and specificity observed even when the experimental conditions involved limited data availability and inconsistent image acquisition outcomes that researchers trace to ViT's capacity for holistic image comprehension, a representational property that CNN operations cannot replicate by design. A separate research

thread [15] has established that integrating data augmentation protocols within ViT-based training pipelines provides an effective mechanism for addressing label imbalance and extending classification reliability across radiographic presentations that differ substantially from the training distribution. The contribution of transfer learning as an enabler of strong ViT performance in constrained data environments has also been independently documented [16], with findings indicating that initializing model weights from large-scale pre-trained checkpoints produces substantial gains in downstream task accuracy and prediction stability effects of particular practical relevance given the inherent data scarcity challenges of clinical imaging research.

Drawing from accumulated evidence that positions ViT as a high-performing architecture for medical image analysis particularly for pneumonia detection the present study undertakes a systematic exploration of transformer-based methods in elevating chest radiographic classification accuracy. Central to this research is the evaluation of a pre-trained ViT model through transfer learning to classify COVID-19 radiographic images into COVID with pneumonia and COVID without pneumonia classes. Rather than proposing a new augmentation mechanism, this study focuses on assessing the feasibility of ViT transfer learning under limited dataset conditions using standardized preprocessing, stratified data partitioning, and multi-metric evaluation. The anticipated output of this combined strategy is a classification system that demonstrates both superior reliability and heightened accuracy in pneumonia identification, with direct implications for clinical decision support systems prioritizing diagnostic speed and automation.

METHOD

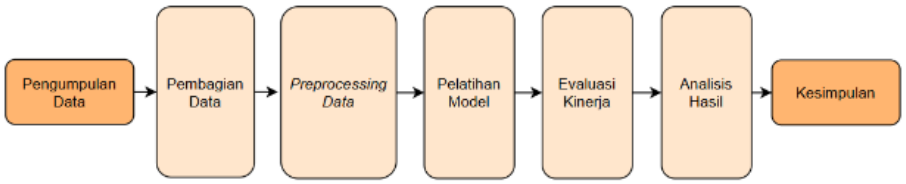


Figure 1. Research Flow

The methodological framework of this study centers on the construction and empirical validation of a Vision Transformer (ViT) transfer learning model for binary classification of COVID-19 radiographic images into two diagnostic categories: COVID with pneumonia and COVID without pneumonia. The research workflow was structured to include data collection, stratified data splitting, image preprocessing, ViT model implementation, model training, and performance evaluation. These stages were organized to support transparent reporting of artificial intelligence-based medical imaging research [8].

Data Collection

The dataset used in this study was obtained from a publicly accessible Kaggle repository containing 513 chest X-ray images of COVID-19 patients with and without pneumonia [9]. Each image belongs to one of two diagnostic categories: COVID with pneumonia and COVID without pneumonia. Therefore, the classification task in this study is COVID-specific and should not be generalized to all forms of pneumonia without further validation using broader pneumonia datasets.

Data Splitting

A stratified allocation scheme was applied to distribute the dataset into training, validation, and testing subsets. The final data partition consisted of 359 training images, 73 validation images, and 81 testing images. The training subset was used for model weight updating, the validation subset was used to monitor model performance during fine-tuning, and the testing subset was reserved for final performance evaluation. Stratification was applied to preserve the class distribution across the three subsets and to reduce the risk of biased performance estimation. Given the limited dataset size, future studies should consider k-fold cross-validation to obtain a more stable estimate of model generalization in medical imaging AI studies [10].

Table 1. Data Distribution

Class	Training Data	Validation Data	Testing Data	Total
Dataset	359 Images	73 Images	81 Images	513 Images

Image Preprocessing

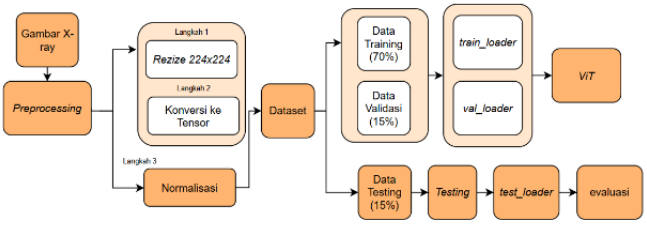


Figure 1. Image Preprocessing Stages

Prior to model ingestion, each radiographic image underwent spatial standardization through resizing to 224×224 pixels, aligning with the dimensional input specifications of the ViT architecture. The spatially normalized images were subsequently converted to tensor representations and subjected to channel-wise intensity normalization using parameters derived from the ImageNet corpus (mean = [0.485, 0.456, 0.406]; standard deviation = [0.229, 0.224, 0.225]). Post-normalization, the processed samples were organized into three dedicated data loader objects train loader, val loader, and test loader corresponding to the 70%, 15%, and 15% allocation proportions respectively, a configuration designed to uphold methodological rigor and impartiality in model performance estimation. No ViT-driven synthetic image generation was applied in the experimental pipeline; therefore, this study is positioned as ViT transfer learning rather than ViT-based data augmentation.

Vision Transformer (ViT) Implementation

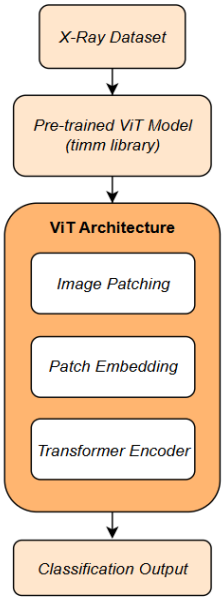


Figure 2. Vision Transformer (ViT) Implementation Stages

Serving as the foundational classification backbone, the ViT model operates by decomposing each input image into a grid of non-overlapping fixed-size patches measuring 16×16 pixels apiece which are individually projected into high-dimensional embedding vectors before being processed sequentially through stacked transformer encoder layers. This patch-based global attention mechanism confers upon ViT a distinctive advantage in perceiving long-range spatial interdependencies spanning the full image extent, fundamentally distinguishing it from CNN

architectures constrained to local convolutional receptive fields. The ViT model was implemented using a pre-trained checkpoint from the PyTorch Image Models (timm) library [11]. The final classification layer was adjusted for binary classification to produce predictions for the COVID with pneumonia and COVID without pneumonia classes. Throughout this fine-tuning phase, the model's learned weight parameters were progressively recalibrated to align with the distinctive visual signatures characteristic of chest radiographic imagery.

Disease Detection Model Training (ViT)

Model training was conducted using AdamW as the optimizer, a learning rate of 0.001, binary cross-entropy as the loss function, and a batch size of 32. The binary cross-entropy loss was selected because the task involved two diagnostic classes. The optimization pipeline ran for 5 epochs; therefore, the limited number of training epochs should be acknowledged as a methodological limitation unless retraining with additional epochs and learning curve analysis is conducted.

Binary Cross-Entropy Loss Function Formula:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \qquad (1)$$

Performance Evaluation

Model performance was evaluated using the held-out testing subset. The confusion matrix was used to calculate true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In this study, TP refers to COVID with pneumonia images correctly classified as positive; FP refers to COVID without pneumonia images incorrectly classified as positive; TN refers to COVID without pneumonia images correctly classified as negative; and FN refers to COVID with pneumonia images incorrectly classified as negative. Accuracy, precision, recall, F1-score, TPR, and FPR were calculated using the following formulas [12].

Accuracy Formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{TP+TN+FP+FN} \qquad (2)$$

Precision is operationalized as the ratio of true positive predictions to the aggregate volume of positive predictions issued, comprising both TP and FP. This metric reflects the discriminative reliability of the model's positive classifications.

Precision Formula:

$$\text{Precision} = \frac{\text{TP}}{TP + FP} \qquad (3)$$

Recall, also referred to as sensitivity or true positive rate, measures the proportion of actual positive cases that are correctly identified by the model. In this study, recall indicates the model's ability to detect COVID with pneumonia cases.

Recall Formula:

$$\text{Recall} = \frac{\text{TP}}{TP + FN} \qquad (4)$$

F1-score constitutes a harmonic synthesis of precision and recall into a unified scalar indicator, rendering it especially informative under conditions of class imbalance. In radiographic disease classification, this composite metric is important because it accounts for both false positive and false negative errors.

F1-Score Formula:

$$f1 - Score = 2\text{x}\ \frac{\text{Presisi x Recall}}{\text{Presisi + Recall}} \qquad (5)$$

For ROC analysis, the true positive rate (TPR) and false positive rate (FPR) were calculated to evaluate classifier discrimination across decision thresholds.

True Positive Rate Formula:

$$\text{TPR} = \frac{\text{TP}}{TP + FP} \quad (6)$$

False Positive Rate Formula:

$$\text{FPR} = \frac{\text{FP}}{FP + TN} \quad (7)$$

The confusion matrix structure used to calculate these metrics is presented in Table 2.

Table 2. Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	*TP*	*FN*
Actual Negative	*FP*	*TN*

RESULTS AND DISCUSSION

Evaluation Results of the ViT Model

The performance of the Vision Transformer (ViT) transfer learning model was evaluated using the held-out testing subset consisting of 81 images. The confusion matrix showed that the model produced 43 true positive predictions, 3 false positive predictions, 35 true negative predictions, and 0 false negative predictions. All performance metrics were derived from this confusion matrix to ensure mathematical consistency between the classification outcomes and the reported evaluation values.

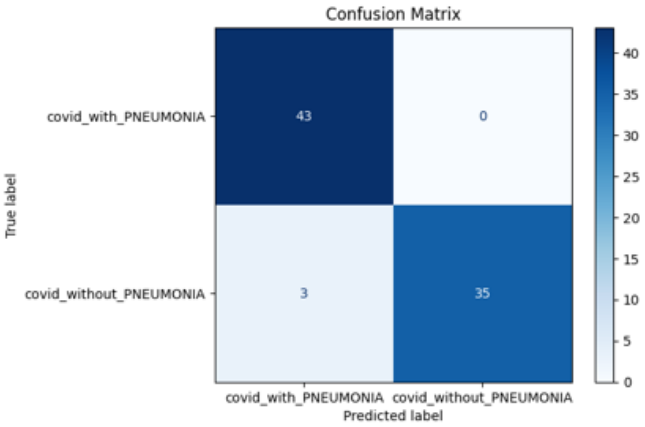


Figure 4. Results of ViT Confusion Matrix Analysis

The confusion matrix demonstrates that all 43 COVID with pneumonia images in the testing subset were correctly classified as positive. This result indicates that the model achieved complete positive-case detection within the evaluated test subset. From a clinical screening perspective, this is an important finding because false negative predictions may cause actual pneumonia cases to be missed, potentially delaying further clinical assessment or treatment. In this study, the absence of false negatives means that no COVID with pneumonia image was incorrectly classified as COVID without pneumonia.

However, the model also produced 3 false positive predictions. These false positives occurred when COVID without pneumonia images were incorrectly classified as COVID with pneumonia. Although false positives are generally less dangerous than false negatives in screening-oriented medical tasks, they remain clinically relevant because they may lead to unnecessary follow-up examinations, increased patient anxiety, or additional clinical workload. Meanwhile, 35 COVID without pneumonia images were correctly classified as negative, indicating that the model retained strong ability to recognize non-pneumonic COVID radiographs. The confusion matrix is summarized in Table 3.

Table 3. Confusion Matrix of the ViT Model

Actual Class	Predicted COVID with Pneumonia	Predicted COVID without Pneumonia	Total
COVID with pneumonia	43	0	43
COVID without pneumonia	3	35	38
Total	46	35	81

Based on the confusion matrix, the model achieved an accuracy of 96.30%, precision of 0.93, recall of 1.00, F1-score of 0.97, specificity of 0.92, false positive rate of 0.08, and ROC-AUC of 1.00. The recall value of 1.00 was obtained because no false negative predictions were observed, while the precision value of 0.93 reflects the presence of three false positive predictions. These values show that the model achieved high sensitivity in detecting COVID with pneumonia cases while still producing several incorrect positive classifications in the COVID without pneumonia class.

Table 4. ViT Performance Results

Metric	Formula	Value
Accuracy	(TP + TN) / (TP + TN + FP + FN)	96.30%
Precision	TP / (TP + FP)	0.93
Recall	TP / (TP + FN)	1.00
F1-score	2 × Precision × Recall / (Precision + Recall)	0.97
Specificity	TN / (TN + FP)	0.92
False Positive Rate	FP / (FP + TN)	0.08
ROC-AUC	Area under ROC curve	1.00

The accuracy value of 96.30% indicates that 78 out of 81 test images were classified correctly. Although accuracy provides a general overview of model performance, it should not be interpreted alone in medical image classification, especially when the clinical implications of false positive and false negative predictions are not equivalent. For this reason, precision, recall, F1-score, specificity, and class-specific performance metrics were also analyzed.

The precision value of 0.93 means that most images predicted as COVID with pneumonia were correct. Specifically, among 46 images predicted as positive, 43 were truly COVID with pneumonia, while 3 were actually COVID without pneumonia. This result shows that the model had a relatively low false positive burden. Nevertheless, the presence of false positives indicates that the model may still classify some non-pneumonic COVID images as pneumonic, possibly because of overlapping radiographic patterns, subtle opacity-like features, image noise, or non-pathological variations in lung appearance.

The recall value of 1.00 indicates that the model successfully detected all actual COVID with pneumonia images in the testing subset. This is particularly important for screening-oriented diagnostic support systems because sensitivity is often prioritized when the main objective is to reduce missed disease cases. In the context of radiographic pneumonia screening, false negative predictions are clinically critical because undetected cases may delay appropriate intervention. Therefore, the absence of false negatives in this study represents a favorable result. However, this finding should still be interpreted cautiously because the testing subset consisted of only 81 images.

The F1-score of 0.97 reflects a strong balance between precision and recall. Since F1-score combines both false positive and false negative considerations, it is especially useful when evaluating diagnostic models under potential class imbalance. The specificity value of 0.92 further indicates that the model correctly identified most COVID without pneumonia cases, although three false positive errors reduced the specificity from a perfect score. In practical terms, the model demonstrated stronger sensitivity than specificity, suggesting that it was more conservative in detecting COVID with pneumonia cases.

To provide a more transparent evaluation across both diagnostic categories, class-specific metrics were calculated. Reporting only aggregate metrics can obscure class-level behavior, particularly when the two diagnostic categories differ in clinical significance. The class-specific results are shown in Table 5.

Table 5. Class-Specific Performance Results

Class	Precision	Recall	F1-score	Support
COVID with pneumonia	0.93	1.00	0.97	43
COVID without pneumonia	1.00	0.92	0.96	38
Macro average	0.97	0.96	0.96	81
Weighted average	0.97	0.96	0.96	81

The class-specific evaluation shows that the COVID with pneumonia class achieved a recall of 1.00, confirming that all positive cases were correctly identified. Meanwhile, the COVID without pneumonia class achieved a recall of 0.92, indicating that 35 out of 38 negative cases were correctly identified. The lower recall for the COVID without pneumonia class reflects the three false positive predictions. This class-level interpretation is important because it clarifies that the model's main classification error was not missed pneumonia, but over-identification of pneumonia in a small number of non-pneumonic COVID images.

Figure 5 presents the receiver operating characteristic (ROC) curve of the ViT model. The ROC-AUC value was 1.00, indicating strong separability between COVID with pneumonia and COVID without pneumonia across decision thresholds. ROC-AUC is widely used to evaluate diagnostic discrimination because it summarizes the relationship between true positive rate and false positive rate across thresholds [13], [14]. In this study, the perfect ROC-AUC suggests that the model assigned highly separable prediction scores to the two diagnostic categories. However, this result should be interpreted with caution because the testing subset contained only 81 images. A perfect ROC-AUC on a small dataset may indicate strong class separability in the test subset, but it does not automatically guarantee robust performance on external datasets, different hospitals, different imaging devices, or broader pneumonia categories.

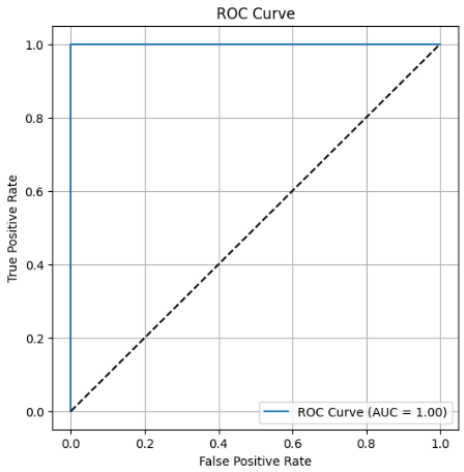


Figure 5. Receiver Operating Characteristic (ROC) of ViT

The results support the potential of ViT transfer learning for COVID-19 pneumonia classification. The model's strong sensitivity is consistent with the theoretical advantage of transformer-based architectures, which can capture long-range spatial relationships through self-attention mechanisms. Previous studies have also reported promising results for transformer-based models in medical imaging tasks, including chest radiography classification and lung disease detection [15], [16]. In particular, ViT-based models have shown potential in learning global image representations that may be useful when pathological patterns are spatially distributed across lung regions [17], [18].

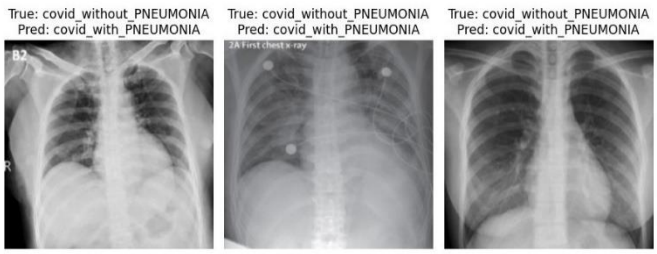
Nevertheless, the findings should be interpreted as evidence of model performance within a constrained experimental setting, not as proof of general clinical readiness. The dataset used in this study is COVID-specific and consists of 513 images. Therefore, the classification task is limited to distinguishing COVID with pneumonia from COVID without pneumonia. It should not be generalized to all forms of pneumonia without additional validation using larger and more diverse chest radiography datasets. Recent reporting guidelines for AI-based clinical prediction and medical imaging studies emphasize the importance of transparent dataset description, adequate validation, and cautious interpretation of model performance [19], [20].

Compared with prior studies on pneumonia and COVID-19 classification using chest X-ray images, the present study contributes a focused evaluation of ViT transfer learning under limited dataset conditions. Several recent studies have explored transformer-based or deep learning-based approaches for chest X-ray classification and have reported high diagnostic performance under specific dataset settings [16], [17], [18]. However, direct comparison across studies is difficult because performance is influenced by dataset size, class composition, preprocessing strategy, training configuration, validation method, and evaluation protocol. Therefore, the high performance observed in this study should be compared cautiously with previous results rather than interpreted as universal superiority.

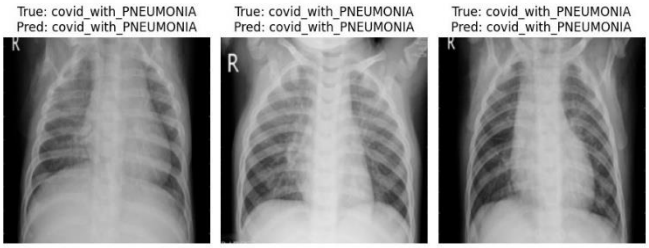
An important limitation of this study is the absence of a direct CNN baseline experiment on the same dataset. Although the theoretical motivation for using ViT is based on its global attention mechanism, the present study did not empirically compare the ViT model against CNN architectures under identical training and testing conditions. Therefore, the findings should not be used to claim that ViT outperforms CNNs in this dataset. A direct comparison with CNN models such as ResNet, DenseNet, EfficientNet, or other established chest X-ray classification backbones would be necessary to validate any claim of architectural superiority.

Another limitation concerns validation stability. Although the model achieved strong performance on the held-out test subset, the test set remained relatively small. Reviewer concerns regarding small-sample evaluation are methodologically important because limited test partitions may produce optimistic or unstable performance estimates. Stratified k-fold cross-validation would provide a more reliable estimate by evaluating the model across multiple train-test partitions and reporting mean and standard deviation values [10]. Future studies should therefore conduct at least five-fold stratified cross-validation and, ideally, external validation using images from different sources.

The visual classification examples in Figure 6 provide additional qualitative insight into the model's prediction behavior. Panel (a) shows false prediction examples involving COVID without pneumonia images that were incorrectly classified as COVID with pneumonia. These examples suggest that the model may be sensitive to visual patterns that resemble pneumonic manifestations, even when the actual class is non-pneumonic COVID. Panel (b) shows true positive examples of COVID with pneumonia images that were correctly classified. These examples illustrate the ability of the ViT model to identify positive cases across varying radiographic appearances. However, qualitative visualization alone is insufficient to explain model decision-making. Future studies should incorporate explainable AI methods, such as attention visualization, saliency maps, Grad-CAM-like approaches, or transformer attention rollout, to better understand which image regions influence model predictions [21], [22].



(a)



(b)

Figure 6. Visualization Results of the ViT Model:
(a) False Prediction (covid_without_pneumonia), (b) True Prediction (covid_with_pneumonia)

Overall, the results indicate that the ViT transfer learning model achieved strong performance in classifying COVID with pneumonia and COVID without pneumonia images. The absence of false negative predictions suggests high sensitivity for detecting positive cases, while the presence of three false positives indicates that specificity can still be improved. The model's performance is promising, but the results remain preliminary because of the limited dataset size, COVID-specific class structure, absence of external validation, absence of radiologist-verified additional ground truth, and lack of direct comparison with CNN baselines. Future research should include larger multi-site datasets, stratified k-fold cross-validation, external testing, CNN baseline comparisons, and explainability analysis to improve robustness, transparency, and clinical applicability [23], [24], [25].

CONCLUSION

The present study evaluated a Vision Transformer (ViT) transfer learning model for binary classification of COVID-19 chest radiography images into COVID with pneumonia and COVID without pneumonia classes. The model achieved an accuracy of 96.30%, precision of 0.93, recall of 1.00, F1-score of 0.97, and ROC-AUC of 1.00 on 81 testing images. These values were derived from a confusion matrix consisting of 43 true positives, 3 false positives, 35 true negatives, and 0 false negatives, indicating that the model demonstrated strong performance in distinguishing between the two diagnostic categories.

The absence of false negative predictions indicates that the model was highly sensitive in identifying COVID with pneumonia cases in the testing subset. However, the presence of three false positive predictions shows that the model still requires further refinement to improve specificity and reduce incorrect positive classifications. Therefore, the results should be interpreted as promising but preliminary, particularly because the dataset size was limited and external validation was not conducted.

This study should also be understood as an evaluation of ViT-based transfer learning rather than a fully implemented ViT-based data augmentation approach. Since no ViT-driven synthetic image generation was applied, future research should develop and test a clearly defined augmentation strategy. In addition, this study did not include a direct comparison with CNN-based models; therefore, no claim of superiority over CNN architectures is made. Future studies should involve larger multi-site datasets, stratified k-fold cross-validation, external testing, CNN baseline comparisons, and explainability analysis to strengthen model robustness, transparency, and clinical applicability.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to Universitas Negeri Makassar for providing academic support during the completion of this study. The authors also acknowledge all parties who contributed indirectly to the preparation of this research, particularly those who supported the data processing, model development, and manuscript preparation stages.

AUTHOR CONTRIBUTION STATEMENT

VA, NA, MR, AR, AN, and SR contributed to the conceptualization of the study, data preparation, model implementation, and manuscript development. VA, NA, and MR were responsible for dataset

preprocessing, model training, and performance evaluation. AR, AN, and SR contributed to literature review, result interpretation, manuscript refinement, and final editing. All authors reviewed and approved the final version of the manuscript.

AI DISCLOSURE STATEMENT

The authors used ChatGPT during the preparation of this manuscript to support language refinement, sentence restructuring, and academic writing improvement. After using the tool, the authors carefully reviewed, verified, and edited the content to ensure accuracy, originality, and alignment with the research objectives. The authors take full responsibility for the final content of this publication.

REFERENCES

[1] H. M. S. S. Herath, H. M. K. K. M. B. Herath, N. Madusanka, and B.-I. Lee, "A systematic review of medical image quality assessment," Journal of Imaging, vol. 11, no. 4, Art. no. 100, 2025, doi: 10.3390/jimaging11040100

[2] E. Goceri, "Medical image data augmentation: Techniques, comparisons and interpretations," Artificial Intelligence Review, vol. 56, pp. 12561-12605, 2023, doi: 10.1007/s10462-023-10453-z

[3] E. Khan, M. Z. U. Rehman, F. Ahmed, F. A. Alfouzan, N. M. Alzahrani, and J. Ahmad, "Chest X-ray classification for the detection of COVID-19 using deep learning techniques," Sensors, vol. 22, no. 3, Art. no. 1211, 2022, doi: 10.3390/s22031211

[4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. International Conference on Learning Representations, 2021, arXiv:2010.11929

[5] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," Medical Image Analysis, vol. 88, Art. no. 102802, 2023, doi: 10.1016/j.media.2023.102802

[6] M. Usman, T. Zia, and A. Tariq, "Analyzing transfer learning of vision transformers for interpreting chest radiography," Journal of Digital Imaging, vol. 35, no. 6, pp. 1445-1462, 2022, doi: 10.1007/s10278-022-00666-z

[7] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," Scientific Reports, vol. 14, Art. no. 2487, 2024, doi: 10.1038/s41598-024-52703-2

[8] A. S. Tejani et al., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update," Radiology: Artificial Intelligence, vol. 6, no. 4, Art. no. e240300, 2024, doi: 10.1148/ryai.240300

[9] R. P. Rahman, "Covid w/wo Pneumonia Chest Xray," Kaggle Dataset. [Online]. Available: https://www.kaggle.com/rashikrahmanpritom/covid-wwo-pneumonia-chest-xray. Accessed: Jul. 1, 2026

[10] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A guide to cross-validation for artificial intelligence in medical imaging," Radiology: Artificial Intelligence, vol. 5, no. 4, Art. no. e220232, 2023, doi: 10.1148/ryai.220232

[11] R. Wightman et al., "rwightman/pytorch-image-models: v0.6.11 release," Zenodo, 2022, doi: 10.5281/zenodo.7140899

[12] F. S. Nahm, "Receiver operating characteristic curve: Overview and practical use for clinicians," Korean Journal of Anesthesiology, vol. 75, no. 1, pp. 25-36, 2022, doi: 10.4097/kja.21209

[13] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, "The area under the receiver operating characteristic curve as a performance metric for binary classifiers: A review and practical guidance," Ecological Modelling, vol. 505, Art. no. 110878, 2024, doi: 10.1016/j.ecolmodel.2024.110878

[14] S. Roumeliotis, K. Tsiakiris, and P. Dounousi, "ROC curve analysis: A useful statistic multi-tool in the research and clinical practice toolkit," International Urology and Nephrology, vol. 56, pp. 3657-3667, 2024, doi: 10.1007/s11255-024-04022-8

[15] J. Ko, S. Park, and H. G. Woo, "Optimization of vision transformer-based detection of lung diseases from chest X-ray images," BMC Medical Informatics and Decision Making, vol. 24, Art. no. 191, 2024, doi: 10.1186/s12911-024-02591-3

[16] T. Chen, I. Philippi, Q. B. Phan, L. Nguyen, N. T. Bui, C. daCunha, and T. T. Nguyen, "A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images," Healthcare Analytics, vol. 5, Art. no. 100332, 2024, doi: 10.1016/j.health.2024.100332

[17] F. Alshanketi, A. Alharbi, M. Kuruvilla, V. Mahzoon, S. T. Siddiqui, N. Rana, and A. Tahir, "Pneumonia detection from chest X-ray images using deep learning and transfer learning for imbalanced datasets," Journal of Imaging Informatics in Medicine, vol. 38, no. 4, pp. 2021-2040, 2025, doi: 10.1007/s10278-024-01334-0

[18] R. Siddiqi and S. Javaid, "Deep learning for pneumonia detection in chest X-ray images: A comprehensive survey," Journal of Imaging, vol. 10, no. 8, Art. no. 176, 2024, doi: 10.3390/jimaging10080176

[19] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378

[20] S. Sharma and K. Guleria, "A systematic literature review on deep learning approaches for pneumonia detection using chest X-ray images," Multimedia Tools and Applications, vol. 83, no. 8, pp. 24101-24151, 2024, doi: 10.1007/s11042-023-16419-1

[21] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," Medical Image Analysis, vol. 79, Art. no. 102470, 2022, doi: 10.1016/j.media.2022.102470

[22] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," ACM Computing Surveys, vol. 56, no. 4, Art. no. 85, pp. 1-41, 2023, doi: 10.1145/3625287

[23] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," Computers in Biology and Medicine, vol. 152, Art. no. 106391, 2023, doi: 10.1016/j.compbiomed.2022.106391

[24] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: A review," Journal of Imaging, vol. 9, no. 4, Art. no. 81, 2023, doi: 10.3390/jimaging9040081

[25] K. Lekadir et al., "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, Art. no. e081554, 2025, doi: 10.1136/bmj-2024-081554