DOI: https://doi.org/10.70656/ijcse.v3i01.857

Comparative Evaluation of MobileNetV2 and Vision Transformer for Pneumonia Classification on Chest X-Ray Images: Fine-Tuning, Focal Loss, and Decision-Threshold Analysis

Ali Novian
Department of Informatics Computer Science of Faculty Universitas Amikom Purwokerto
Purwokerto, Indonesia
alinovian.edu@gmail.com

Ma`dan Shomsomi
Department of Informatics Computer Science of Faculty Universitas Amikom Purwokerto
Purwokerto, Indonesia
madanshomsomi2@gmail.com

Widhaksa Triawan
Department of Informatics Computer Science of Faculty Universitas Amikom Purwokerto
Purwokerto, Indonesia
widhaksatriawan7@gmail.com

Hendra Marcos
Department of Informatics Computer Science of Faculty Universitas Amikom Purwokerto
Purwokerto, Indonesia
hendra.marcos@amikompurwokerto.ac.id

Abstract*—Pneumonia remains a major respiratory infection that imposes a substantial health burden and requires rapid and accurate detection. This study evaluates five deep-learning configurations for binary classification of chest X-ray (CXR) images into Normal and Pneumonia classes: MobileNetV2 with a frozen backbone, fine-tuned MobileNetV2, fine-tuned MobileNetV2 with Focal Loss, Vision Transformer (ViT) without full fine-tuning, and fine-tuned ViT. The public dataset used in the experiments contains 5,856 images, comprising 5,216 training images, 16 validation images, and 624 test images. All metrics in this revised version were recalculated consistently from the experimental confusion matrices. ViT without full fine-tuning achieved the best overall performance at a threshold of 0.5, with 92.63% accuracy, 98.72% sensitivity, 82.48% specificity, a 94.36% F1-score, 90.60% balanced accuracy, and a Matthews correlation coefficient (MCC) of 0.845. MobileNetV2 with Focal Loss achieved 92.15% accuracy, 98.97% sensitivity, 80.77% specificity, and a 94.03% F1-score, providing slightly higher sensitivity with only four false-negative cases. In contrast, fine-tuned ViT achieved 100% sensitivity but only 57.69% specificity at the 0.5 threshold, indicating a shift in the predicted probability distribution and imbalanced generalization. Post hoc threshold analysis showed that changing the decision threshold can improve the error trade-off; however, it was not used to select the primary model because the threshold sweep was evaluated on the test set. These findings demonstrate that fine-tuning strategy and loss function affect model error characteristics differently, while screening-oriented evaluation should consider sensitivity and specificity together with accuracy

Keywords-Chest X-Ray, Pneumonia, MobileNetV2, Vision Transformer, Fine-Tuning, Focal Loss, Decision Threshold

I. INTRODUCTION

Lower respiratory tract infections remain an important cause of global morbidity and mortality. The Global Burden of Disease 2021 analysis estimated approximately 344 million episodes of lower respiratory infections and 2.18 million deaths in 2021, including about 502,000 deaths among children younger than five years [1]. This burden reinforces the need for rapid and affordable detection strategies that can support triage and screening, particularly in healthcare settings with limited clinical resources.

Chest X-ray (CXR) is a widely used and relatively accessible radiographic modality. The pediatric dataset published by Kermany et al. and subsequently distributed through public repositories has become a frequently used benchmark for developing image-based pneumonia classification methods [2]. The dataset comprises anterior-posterior CXR images from pediatric patients approximately one to five years of age and underwent quality control and expert label assessment. These characteristics make it useful for benchmarking while simultaneously limiting direct generalization to adult populations and other institutions.

In convolutional neural network (CNN)-based approaches, transfer learning enables visual representations learned from ImageNet to be reused for medical classification tasks with more limited datasets. MobileNetV2 employs inverted residual blocks and linear bottlenecks, making it relatively efficient compared with larger CNN architectures [3]. Transfer-learning benchmark studies on pneumonia CXR images have also shown that backbone freezing and fine-tuning strategies can produce different performance behavior across architectures [6].

Vision Transformer (ViT) offers a different modeling mechanism by processing images as sequences of patches and using self-attention to capture global dependencies [4]. Transformer-based approaches have shown competitive results for CXR classification, including IEViT, which evaluated multiple chest radiography datasets [7], and the study by Singh et al., which reported the potential of ViT for pneumonia detection [8]. More recent developments also suggest that integrating local CNN features with the global context captured by transformers can improve pneumonia classification, as demonstrated by DSSViT in 2025 [11].

In addition to architecture, two other factors influence classification performance. First, class imbalance and

differences in sample difficulty may bias a model toward the dominant class. Focal Loss modifies cross-entropy by reducing the contribution of samples that are already easy to classify while assigning relatively greater emphasis to hard examples [5]. Second, a fixed probability threshold of 0.5 does not necessarily provide the best trade-off between sensitivity and specificity. In a screening context, false negatives and false positives have different consequences; therefore, evaluation should extend beyond accuracy alone.

This study aims to perform a comparative evaluation of five MobileNetV2 and ViT configurations on the same pneumonia CXR dataset. The main contributions are: (1) comparing frozen and fine-tuned models using the same test set; (2) evaluating the additional effect of Focal Loss on MobileNetV2; (3) recalculating all metrics directly from the confusion matrices to ensure consistency; and (4) exploratorily analyzing performance sensitivity to changes in the decision threshold. The reporting follows transparency principles for artificial intelligence research in medical imaging, including a clear distinction between primary results and post hoc analyses [9].

II. METHODOLOGY

A. Study Design and Dataset

This study is a retrospective experiment based on a public dataset for binary CXR classification. The dataset used was Chest X-Ray Images (Pneumonia), originating from the pediatric collection of Guangzhou Women and Children's Medical Center and distributed publicly [2]. The dataset copy used in the experiments contained 5,856 images in three partitions: training, validation, and test. The class distribution is shown in Table 1.

Table 1. Distribution of the Dataset Used

Partition	Normal	Pneumonia	Total
Training	1,341	3,875	5,216
Validation	8	8	16
Test	234	390	624
Total	1,583	4,273	5,856

The original validation set contained only 16 images. Because this partition was used in the initial experiments to monitor val_loss and select checkpoints, its very small size is treated as a methodological limitation and is discussed explicitly in the Limitations subsection. The 624-image test set was retained as the primary evaluation basis for all models.

B. Preprocessing and Data Augmentation

All images were resized to 224 × 224 pixels. In the original pipeline, pixel values were normalized using a 1/255 scale. Augmentation was applied only to the training set and included rotation, translation, zoom, and horizontal flipping. To reduce class imbalance, Normal images in the training set were augmented until their effective number approached the number of Pneumonia images. The validation and test sets received no stochastic augmentation.

Figure 1. Examples of Normal, bacterial pneumonia, and viral pneumonia images from the dataset used.

C. Model Configurations

Five model configurations were evaluated. Three configurations used ImageNet-pretrained MobileNetV2 as the CNN backbone. The classification head consisted of GlobalAveragePooling2D, a Dense layer with 128 units and ReLU activation, and a single sigmoid output unit. The baseline model used a learning rate of 3 × 104 with Binary Cross-Entropy (BCE). During fine-tuning, the last 100 MobileNetV2 layers were unfrozen and the learning rate was reduced to 1 × 105.

The other two configurations used a Vision Transformer implemented through keras_cv.models.VisionTransformer with 224 × 224 input images. The experiments compared a pretrained configuration without full fine-tuning and a configuration after fine-tuning. Internal ViT parameters that were not fully documented in the original manuscript follow the experimental notebook provided in the code repository. This limitation in ViT configuration documentation is retained transparently and represents an aspect that should be improved in future replications.

Table 2. Eperimental Configurations

ID	Configuration	Loss	LR
M1	MobileNetV2 frozen	BCE	3e-4
M2	MobileNetV2 fine-tuned	BCE	1e-5
M3	MobileNetV2 fine-tuned + Focal	Foca l	1e-5
M4	ViT pretrained / no full FT	BCE	Run-specific
M5	ViT fine-tuned	BCE	Run-specific

D. Focal Loss

For M3, Focal Loss was applied with gamma = 2.0 and alpha = 0.75. In general, Focal Loss is defined as [5]:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t)$$

Because the training stream had already been balanced through augmentation, the primary interpretation of Focal Loss in this experiment is not solely as a class-imbalance treatment, but also as a mechanism for increasing emphasis on samples that are difficult to classify.

E. Training Procedure

The dataset was loaded using image_dataset_from_directory with a batch size of 32. Models were trained for a maximum of 50 epochs. EarlyStopping and ModelCheckpoint were used to retain the best model according to val_loss. For fine-tuned MobileNetV2,

pretrained weights were adapted using a lower learning rate. All configurations were then evaluated on the same test set, ensuring that the confusion-matrix comparisons used identical sample counts.

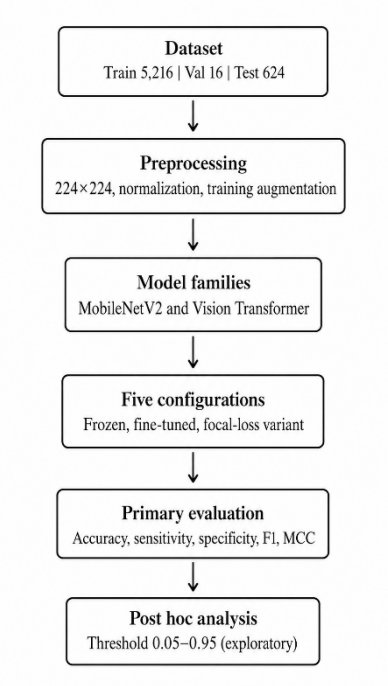

Figure 2. Experimental and evaluation workflow

F. Evaluation Metrics

For the positive Pneumonia class, the confusion matrix was defined using true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The main metrics were calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

Negative predictive value (NPV), balanced accuracy, and Matthews correlation coefficient (MCC) were also calculated. For accuracy, sensitivity, and specificity, 95% confidence intervals were estimated using Wilson score intervals based on the event counts in each confusion matrix. AUROC and AUPRC were not reconstructed because individual prediction probabilities were unavailable in the retained experimental artifacts.

G. Decision-Threshold Analysis

The sigmoid output was classified as Pneumonia when probability p was greater than or equal to threshold t. A threshold of 0.5 was used for all primary comparisons. In addition, a post hoc analysis was performed over 19 thresholds from 0.05 to 0.95 in increments of 0.05. Because the threshold sweep in the original experiment used the test set, these results are presented only as a sensitivity analysis and were not used to select the primary model or to claim a clinically optimal threshold. Future studies should select the decision threshold using a validation set or a separate calibration set.

III. Result

A. Primary Performance at Threshold 0.5

All metrics in the revised version were recalculated from the confusion matrices to eliminate inconsistencies among the narrative, tables, and visualizations. ViT pretrained without full fine-tuning (M4) achieved the highest accuracy, 92.63% (95% CI: 90.31%-94.43%). The model also achieved an F1-score of 94.36%, balanced accuracy of 90.60%, and MCC of 0.845. The difference from MobileNetV2 + Focal Loss was small, only 0.48 percentage points in accuracy; therefore, this result should not be interpreted as a statistically significant difference without paired prediction-level testing.

Table 3. Test-Set Performance at Threshold 0.5

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1
M1 MNV2 Frozen	86.22% (83.29-88.70)	85.38% (81.53-88.55)	87.61% (82.77-91.23)	88.56%
M2 MNV2 FT	91.99% (89.59-93.87)	98.97% (97.39-99.60)	80.34% (74.78-84.93)	93.92%
M3 MNV2 FT+FL	92.15% (89.77-94.01)	98.97% (97.39-99.60)	80.77% (75.24-85.31)	94.03%
M4 ViT Frozen	92.63% (90.31-94.43)	98.72% (97.03-99.45)	82.48% (77.09-86.81)	94.36%
M5 ViT FT	84.13% (81.06-86.79)	100% (99.02-100)	57.69% (51.29-63.85)	88.74%

Table 4. Additional Metrics and Error Distribution

Model	Precision	NPV	Balanced Acc.	MCC	FP	FN
M1	91.99%	78.24%	86.50%	0.716	29	57
M2	89.35%	97.92%	89.66%	0.832	46	4
M3	89.56%	97.93%	89.87%	0.835	45	4
M4	90.38%	97.47%	90.60%	0.845	41	5
M5	79.75%	100%	78.85%	0.678	99	0

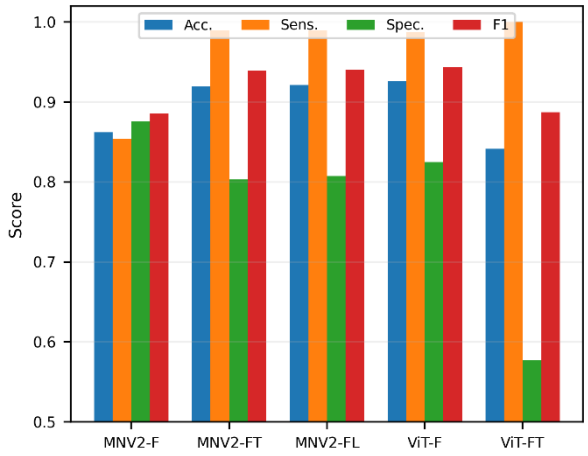

Figure 3. Comparison of accuracy, sensitivity, specificity, and F1-score across the five models at a threshold of 0.5. MNV2: MobileNetV2; FT: fine-tuned; FL: Focal Loss.

B. Effects of Fine-Tuning and Focal Loss on MobileNetV2

Frozen MobileNetV2 (M1) produced 205 TN, 29 FP, 57 FN, and 333 TP. Fine-tuning (M2) sharply reduced FN from 57 to 4, increasing sensitivity from 85.38% to 98.97%. However, this improvement was accompanied by a decrease in specificity from 87.61% to 80.34% because FP increased from 29 to 46. Thus, fine-tuning shifted the model toward more aggressive pneumonia detection.

Adding Focal Loss to the fine-tuned configuration (M3) maintained FN at four cases while reducing FP from 46 to 45. The effect was small but consistent: accuracy increased from 91.99% to 92.15%, specificity from 80.34% to 80.77%, F1-score from 93.92% to 94.03%, and MCC from 0.832 to 0.835. Because the training set had already been balanced through augmentation, this modest improvement is more reasonably interpreted as an effect of down-weighting easy examples than as the primary solution to class imbalance.

C. Effect of Fine-Tuning on Vision Transformer

ViT pretrained without full fine-tuning (M4) showed the best balance between detecting pneumonia and preserving recognition of the Normal class. Its confusion matrix contained 193 TN, 41 FP, 5 FN, and 385 TP. In contrast, fine-tuned ViT (M5) at a threshold of 0.5 produced 135 TN, 99 FP, 0 FN, and 390 TP. Sensitivity reached 100%, but specificity decreased to 57.69% and MCC fell to 0.678. This pattern indicates that fine-tuning shifted the score distribution toward the Pneumonia class, making the default threshold of 0.5 imbalanced.

These findings are insufficient to conclude that overfitting occurred based on test accuracy alone. The combination of reduced specificity and changes in threshold response is more appropriately interpreted as unstable generalization and possible probability miscalibration. Calibration-curve analysis or expected calibration error would be required to test this hypothesis directly.

D. Post Hoc Decision-Threshold Analysis

reduced false positives but began to increase false negatives at high thresholds. At a threshold of 0.50, FP = 41, FN = 5, and F1 = 0.944. At a threshold of 0.95, FP decreased to 23 and F1 increased to 0.960, but FN increased to 9. Arithmetically, the 0.95 configuration yielded 94.87% accuracy, 97.69% sensitivity, and 90.17% specificity. Because the 0.95 threshold was selected after examining the test set, these values are exploratory results rather than unbiased estimates of generalization.

For fine-tuned ViT, a threshold of 0.50 produced 99 FP and 0 FN, whereas a threshold of 0.95 produced 44 FP and 5 FN with an F1-score of 0.940. This large shift further suggests that the main issue with M5 is not only feature representation but also the location of the decision boundary in the output-probability space.

Table 5. Summary of Post Hoc Threshold Analysis

Model	Threshold	FP	FN	Sensitivity	Specificity	F1
ViT Frozen	0.50	41	5	98.72%	82.48%	0.944
ViT Frozen	0.75	33	5	98.72%	85.90%	0.953
ViT Frozen	0.85	28	5	98.72%	88.03%	0.959
ViT Frozen	0.95	23	9	97.69%	90.17%	0.960
ViT FT	0.50	99	0	100%	57.69%	0.887
ViT FT	0.75	75	1	99.74%	67.95%	0.911
ViT FT	0.85	60	2	99.49%	74.36%	0.926
ViT FT	0.95	44	5	98.72%	81.20%	0.940
ViT FT	0.50	99	0	100%	57.69%	0.887
ViT FT	0.75	75	1	99.74%	67.95%	0.911
ViT FT	0.85	60	2	99.49%	74.36%	0.926
ViT FT	0.95	44	5	98.72%	81.20%	0.940

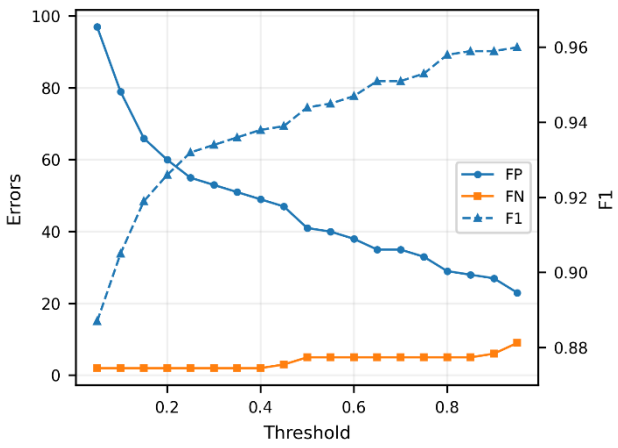

Figure 4. Trade-off among false positives, false negatives, and F1-score across thresholds for pretrained ViT. The analysis is post hoc.

IV. DISCUSSION

A. Main Findings

This study shows that the model with the highest accuracy is not necessarily the model with the highest sensitivity. Pretrained ViT (M4) provided the most balanced overall performance based on accuracy, balanced accuracy, F1-score, and MCC. In contrast, MobileNetV2 + Focal Loss (M3) achieved slightly higher sensitivity and slightly fewer false negatives. For a screening scenario that strongly prioritizes reducing missed pneumonia cases, M3 may therefore be a reasonable candidate. However, for a more balanced sensitivity-specificity profile at the standard threshold, M4 performed better.

These results are consistent with the literature showing that CNNs and transformers provide different representational biases. CNNs exploit local structure and translational inductive bias, whereas ViT can model global relationships among image patches [3], [4]. Okolo et al. showed that enhanced ViT representations can provide high F1 performance across multiple CXR datasets [7], while Singh et al. reported 97.61% accuracy, 95% sensitivity, and 98% specificity using a ViT framework for pneumonia detection [8]. The present study achieved lower specificity, indicating that model configuration, data splitting, preprocessing, and fine-tuning strategy substantially influence results across studies.

B. Fine-Tuning Does Not Always Improve Generalization

Fine-tuning MobileNetV2 provided a substantial gain in sensitivity, but the trade-off in specificity was also evident. This is consistent with transfer-learning principles: adapting pretrained features can improve alignment with the target domain, while the number of unfrozen layers, learning rate, and validation-set size influence stability [6]. For ViT, fine-tuning instead produced predictions strongly biased toward Pneumonia at the 0.5 threshold. Transformers are known to depend heavily on pretraining strategy and regularization when the target dataset is relatively limited [4].

More recent studies have used more complex transformer modifications. DSSViT, for example, combines DenseNet121, squeeze-excitation, and adaptive fusion and reported 97.69% accuracy [11]. Vi-GeN 1.0 uses WGAN-GP for augmentation and a ViT classifier, reporting 97.3% accuracy, 98.1% sensitivity, 96.5% specificity, and an AUC of 0.985 [12]. These numerical comparisons are not head-to-head because the sampling protocols, preprocessing procedures, and model-selection strategies differ; rather, the variation underscores the importance of a uniform evaluation pipeline.

C. Focal Loss Provides a Small but Consistent Improvement

Focal Loss in M3 increased accuracy by only about 0.16 percentage points compared with M2. This small effect is understandable because the training set had already been balanced through augmentation. Under these conditions, the focal mechanism primarily acts by down-weighting easy samples rather than directly correcting class proportions [5]. Future studies should separate these two factors through a stricter ablation design: original imbalance + BCE, original imbalance + Focal Loss, balanced augmentation + BCE, and balanced augmentation + Focal Loss.

D. Threshold, Calibration, and Relevance for Screening

The threshold analysis demonstrates that a higher F1-score does not necessarily correspond to fewer false negatives. For M4, increasing the threshold to 0.95 raised F1 from 0.944 to 0.960 but increased FN from 5 to 9. In screening applications, the increase in missed cases may be more important than the improvement in F1. Therefore, the threshold should be selected according to a prespecified clinical objective, for example by maximizing specificity subject to a minimum sensitivity constraint of 95% or 98%.

More importantly, the thresholds in this study were evaluated post hoc on the test set. Using the test set to select a threshold would yield an optimistically biased estimate of performance. A subsequent protocol should select the threshold using a separate validation/calibration set and then apply it once to the test set. In addition, temperature scaling, Platt scaling, or isotonic regression could be evaluated to ensure that predicted probabilities are adequately calibrated before use in decision support.

E. Comparison with Previous Studies

Table 6. Comparison Eith Representative Studies

Study	Approach	Accuracy	Sensitivity	Specificity
Szepesi & Szilagyi, 2022 [10]	Custom CNN	97.2%	97.3%	-
Singh et al., 2024 [8]	Vision Transformer	97.61 %	95%	98%
Huang, 2025 [11]	DSSViT	97.69 %	-	-
Dash et al., 2025 [12]	Vi-GeN 1.0	97.3%	98.1%	96.5%
Aljuaid et al., 2026 [14]	Uniform CNN benchmark (VGG)	97%	-	-
This study	ViT pretrained	92.63 %	98.72%	82.48 %

Table 6 shows that several previous studies reported higher accuracy. However, cross-study comparisons should be interpreted cautiously because of differences in data partitioning, dataset size, class balancing, model complexity, and potential use of augmented data. Aljuaid et al. emphasized the need for a uniform pipeline to enable fairer model comparisons [14]. Therefore, the results of this study are more appropriately interpreted as an internal comparison among five configurations evaluated on the same test set rather than as a state-of-the-art claim.

F. Limitations

This study has several important limitations. First, the validation set contained only 16 images; consequently, early stopping and checkpoint selection may have high variance. Second, the dataset originated from a single pediatric center

and a limited age range, so generalization to adult populations, other radiographic devices, and other hospitals cannot be assumed. Third, the internal ViT documentation in the original manuscript did not record all architectural hyperparameters; future replications should explicitly report patch size, embedding dimension, depth, number of attention heads, pretrained checkpoint, trainable parameters, and random seed, in accordance with the CLAIM 2024 reporting principles [9].

Fourth, the threshold analysis was performed on the test set and is therefore exploratory only. Fifth, the retained experimental outputs did not store individual probabilities required to calculate AUROC, AUPRC, calibration error, bootstrap confidence intervals for non-binomial metrics, and paired McNemar tests between models. Sixth, the study did not include external validation, subgroup analysis, explainability, or comparison with radiologist assessments. Accordingly, the findings should not be interpreted as evidence of readiness for clinical implementation.

V. CONCLUSION

Evaluation of the five deep-learning configurations showed that pretrained ViT without full fine-tuning provided the best balance of performance at a threshold of 0.5, with 92.63% accuracy, 98.72% sensitivity, 82.48% specificity, a 94.36% F1-score, 90.60% balanced accuracy, and an MCC of 0.845. Fine-tuned MobileNetV2 with Focal Loss achieved the highest sensitivity among models that still maintained specificity above 80%, reaching 98.97% sensitivity with four false negatives. Fine-tuning MobileNetV2 substantially increased sensitivity relative to the frozen backbone, whereas fine-tuning ViT resulted in low specificity at the default threshold, indicating that transformer adaptation strategies require more careful control.

Post hoc analysis showed that the decision threshold can substantially change the FP-FN trade-off; however, a threshold selected from the test set cannot be used as an unbiased estimate of final performance. Future research should use an adequately sized validation set, validation- or calibration-based threshold selection, repeated runs or cross-validation, external testing, calibration analysis, and complete reporting of the ViT configuration. With these improvements, comparisons between MobileNetV2 and ViT can provide stronger evidence for the development of CXR-based pneumonia screening support systems without claiming clinical readiness before external and prospective validation.

ACKNOWLEDGMENT

The authors thank the Faculty of Computer Science, Universitas Amikom Purwokerto, for the academic support and research facilities provided for this study. The authors also acknowledge the contributors of the publicly available Chest X-Ray Images (Pneumonia) dataset on Kaggle for supporting the experimental evaluation presented in this work.

REFERENCES

[1] GBD 2021 Lower Respiratory Infections and Antimicrobial Resistance Collaborators, "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," Lancet Infect. Dis., vol. 24, no. 9, pp. 974-1002, 2024, doi: 10.1016/S1473-3099(24)00176-2.

[2] D. S. Kermany, M. Goldbaum, W. Cai, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122-1131.e9, 2018, doi: 10.1016/j.cell.2018.02.010.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510-4520.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980-2988.

[6] D. Avola, A. Bacciu, L. Cinque, A. Fagioli, M. R. Marini, and R. Taiello, "Study on transfer learning capabilities for pneumonia classification in chest-x-rays images," Comput. Methods Programs Biomed., vol. 221, Art. no. 106833, 2022, doi: 10.1016/j.cmpb.2022.106833.

[7] G. I. Okolo, S. Katsigiannis, and N. Ramzan, "IEViT: An enhanced vision transformer architecture for chest X-ray image classification," Comput. Methods Programs Biomed., vol. 226, Art. no. 107141, 2022, doi: 10.1016/j.cmpb.2022.107141.

[8] 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, Art. no. 2487, 2024, doi: 10.1038/s41598-024-52703-2.

[9] A. S. Tejani, M. E. Klontzas, A. A. Gatti, et al., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update," Radiol. Artif. Intell., vol. 6, no. 4, 2024, doi: 10.1148/ryai.240300.

[10] P. Szepesi and L. Szilagyi, "Detection of pneumonia using convolutional neural networks and deep learning," Biocybern. Biomed. Eng., vol. 42, no. 3, pp. 1012-1022, 2022, doi: 10.1016/j.bbe.2022.08.001.

[11] J. Huang, "DSSViT: Multi-scale adaptive fusion Vision Transformer with dense feature reuse for robust pneumonia detection in chest radiography," Int. J. Imaging Syst. Technol., vol. 35, no. 3, Art. no. e70127, 2025, doi: 10.1002/ima.70127.

[12] A. Dash, S. Panigrahi, D. S. K. Nayak, A. P. Dash, and T. Swarnkar, "Vi-GeN 1.0: GAN-Augmented Vision Transformer pipeline for diversified pneumonia classification," J. Transform. Technol. Sustain. Dev., vol. 9, Art. no. 14, 2025, doi: 10.1007/s41314-025-00081-6.

[13] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," Comput. Biol. Med., vol. 152, Art. no. 106391, 2023.

[14] H. Aljuaid, H. H. A. Adlan, B. Alkebsi, B. S. Alfurhood, A. Liotta, and L. Cavallaro, "An experimental comparison of deep learning models for pneumonia classification from chest X-ray images," Biomed. Signal Process. Control, vol. 112, Art. no. 108742, 2026, doi: 10.1016/j.bspc.2025.108742.