INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

PNEUMONIA CLASSIFICATION FROM CHEST X-RAY IMAGES USING A CUSTOM CONVOLUTIONAL NEURAL NETWORK AND MOBILENETV2

Hammad Akram[1], Muhammad Luqman[2], Ayesha Maryam[3], Israr Hanif[4], Faisal Shahzad[*5]

[1,3,4,*5]*Department of Computer Science, Bahauddin Zakariya University, Multan 60800, Pakistan*
[2]*Faculty of Computer Science and Information Technology, Multan University of Science and Technology, Multan 60000, Pakistan*

[1]hammad.akram.cs@gmail.com, [2]muhammadluqmankhan7860@gmail.com,
[3]ayeshamaryam2133@gmail.com, [4]aoaisrar@bzu.edu.pk, [*5]faisal.shahzad.research@gmail.com

DOI: https://doi.org/10.5281/zenodo.24031593

Keywords
Pneumonia; Chest X-ray; CNN; MobileNetV2; Transfer Learning; Medical Image Classification.

Article History
Received: 05 August 2026
Accepted: 19 October 2026
Published: 04 November 2026

Copyright @Author
Corresponding Author: *
Faisal Shahzad

Abstract
This study compares a custom convolutional neural network (CNN) trained from scratch with an ImageNet-pretrained MobileNetV2 model for binary classification of chest X-ray images into NORMAL and PNEUMONIA classes. The experimental setup included 5,216 training images, 16 validation images, and 624 test images. Images were resized to 224 × 224 pixels; custom CNN training used rotation, zoom, and horizontal-flip augmentation. Both models were trained for five epochs using class weights. On the test set, the custom CNN achieved 86.86% accuracy, 0.8615 precision, 0.9410 recall, and 0.4210 loss. MobileNetV2 achieved 85.42% accuracy, 0.8201 precision, 0.9821 recall, 0.5149 loss, and an ROC-AUC of 0.9576. The study selected the custom CNN using an accuracy-based rule. Its confusion matrix contained 175 true NORMAL predictions, 59 false positives, 23 false negatives, and 367 true PNEUMONIA predictions. MobileNetV2 had higher pneumonia recall, showing a trade-off between overall accuracy and sensitivity. The very small validation set and undocumented dataset provenance limit generalizability. These experimental models are not clinical diagnostic devices.

1. INTRODUCTION

Pneumonia is a respiratory infection that may produce visible changes in chest radiographs. Deep learning methods, particularly convolutional neural networks, can learn image representations for classification tasks. A custom CNN learns task-specific filters from training data, whereas transfer learning reuses representations learned from a larger dataset. MobileNetV2 is a computationally efficient convolutional architecture that can be adapted by adding a task-specific classification head. The objective was to compare a custom CNN and an ImageNet-pretrained MobileNetV2 model on the same binary chest X-ray classification task. Comparison was based on test accuracy, precision, recall, and loss; ROC-AUC was available for MobileNetV2.

1.1 Literature Review

Deep learning has become an important approach for computer-aided interpretation of chest radiographs. Convolutional neural networks (CNNs) learn visual features directly from image data, reducing the need for manually designed

image descriptors. Prior research has investigated both custom CNN architectures and transfer-learning methods for identifying pneumonia-related patterns in chest X-rays. The following studies provide the background for the present comparison.

1.1.1 Deep learning on pediatric chest X-rays

Kermany et al. (2018) demonstrated the use of image-based deep learning for medical diagnosis, including the classification of pediatric chest X-ray images into normal and pneumonia categories. Their work helped establish a widely used benchmark for investigating automated pneumonia classification and illustrated the potential of transfer learning in medical imaging. The study is relevant to the present work because it demonstrates how learned image features can support a binary normal-versus-pneumonia task. However, results obtained on a particular dataset should not automatically be assumed to generalize to other hospitals, populations, or imaging conditions [5].

1.1.2 CNN-based pneumonia detection

Rajpurkar et al. (2017) introduced CheXNet, a 121-layer convolutional neural network trained on the ChestX-ray14 dataset to detect pneumonia and other thoracic findings. Convolutional neural networks (CNNs) learn visual features directly from image data, reducing the need for manually designed image descriptors [2]. Their study reported strong pneumonia detection performance and compared the model with radiologists on an annotated test set. This work illustrates the capacity of deeper CNN architectures to learn clinically relevant patterns from large collections of chest radiographs. In contrast, the present study investigates a much smaller custom CNN with two convolutional blocks, making the architecture simpler and computationally lighter, although the results are not directly comparable because the datasets and evaluation protocols differ [6].

1.1.3 Transfer learning and MobileNetV2

Sandler et al. (2018) proposed MobileNetV2, an efficient neural network architecture designed for mobile and resource-constrained applications. Its inverted residual blocks and linear bottlenecks reduce computational requirements while retaining useful representational capacity. Although the original paper evaluated MobileNetV2 on general computer-vision benchmarks rather than pneumonia diagnosis specifically, its efficiency makes the architecture a reasonable candidate for transfer learning in medical-image classification. In this study, an ImageNet-pretrained MobileNetV2 feature extractor was frozen and combined with a binary classification head [1].

1.1.4 Generalization, bias, and evaluation challenges

Zech et al. (2018) examined how pneumonia-detection models generalize across different hospital systems. They found that performance could decline on external data and showed that models may learn institution-specific signals in addition to disease-related features. Their findings highlight the importance of external validation and careful dataset construction when evaluating medical-image classifiers [8].

Li et al. (2020) conducted a systematic review and meta-analysis of deep-learning methods for pneumonia detection from chest X-rays. Although the included studies reported encouraging diagnostic performance overall, the authors emphasized methodological concerns that need to be addressed before clinical translation. A broader survey by Siddiqi and Javaid (2024) likewise discussed the expanding use of deep-learning architectures and the continuing challenges associated with datasets, evaluation, and reliable deployment. These reviews support the view that accuracy on a single benchmark is not sufficient evidence of clinical readiness [9], [10].

1.1.5 Research gap and relevance of the present study

Dataset development has been another important contribution to this field. Wang et al. introduced

ChestX-ray8, a large hospital-scale dataset with report-derived labels for multiple thoracic diseases [13]. Irvin et al. presented CheXpert, a large chest-radiograph dataset that explicitly models uncertainty in report-derived labels and includes expert comparison [11]. Johnson et al. described MIMIC-CXR, a de-identified collection of radiographs paired with free-text reports, enabling research on image and report analysis [12]. These resources illustrate both the scale of available data and the challenges of label uncertainty and dataset shift [11], [12], [13].

The reviewed literature shows that CNNs can learn useful radiographic features, while transfer-learning architectures can provide efficient feature extraction. At the same time, reported performance depends on dataset composition, class balance, labeling quality, and validation design. The present study contributes a controlled comparison of a compact custom CNN and a frozen MobileNetV2-based model on the same binary classification task. It also reports class-wise precision and recall, confusion matrices, and ROC-AUC where available. The study should be interpreted as an experimental comparison rather than a clinical validation: the validation split contains only 16 images, the dataset provenance and patient-level separation require confirmation, and no independent external test set was used. Large-scale resources such as ChestX-ray8, CheXpert, and MIMIC-CXR have also supported research on automated chest-radiograph interpretation while highlighting the importance of label quality, patient-level data structure, and representative evaluation [11–13]. Methods such as Grad-CAM can provide visual explanations of CNN predictions, although heatmaps do not by themselves establish clinical correctness [14]. More recent prospective-study analyses have examined model performance against multimodal reference diagnoses, reinforcing the need for clinically meaningful validation [15]. Related chest-radiograph research in tuberculosis has also shown the value of pretrained CNNs and augmentation, although that task is distinct from pneumonia classification [11], [12], [13], [14], [15], [16].

2. Materials and Methods

2.1 Dataset and class distribution

The experiment located the dataset in Google Drive and loaded two directory classes. The class mapping was NORMAL = 0 and PNEUMONIA = 1. The reported counts are shown in Table 1. The available study materials do not identify the dataset's formal name, source publication, license, patient-level split procedure, or acquisition sites; these details cannot be inferred from the supplied output.

Table 1: Distribution of NORMAL and PNEUMONIA Images Across Training, Validation, and Testing Sets.

Split	NORMAL	PNEUMONIA	Total
Training	1341	3875	5216
Validation	8	8	16
Testing	234	390	624
Overall	1583	4273	5856

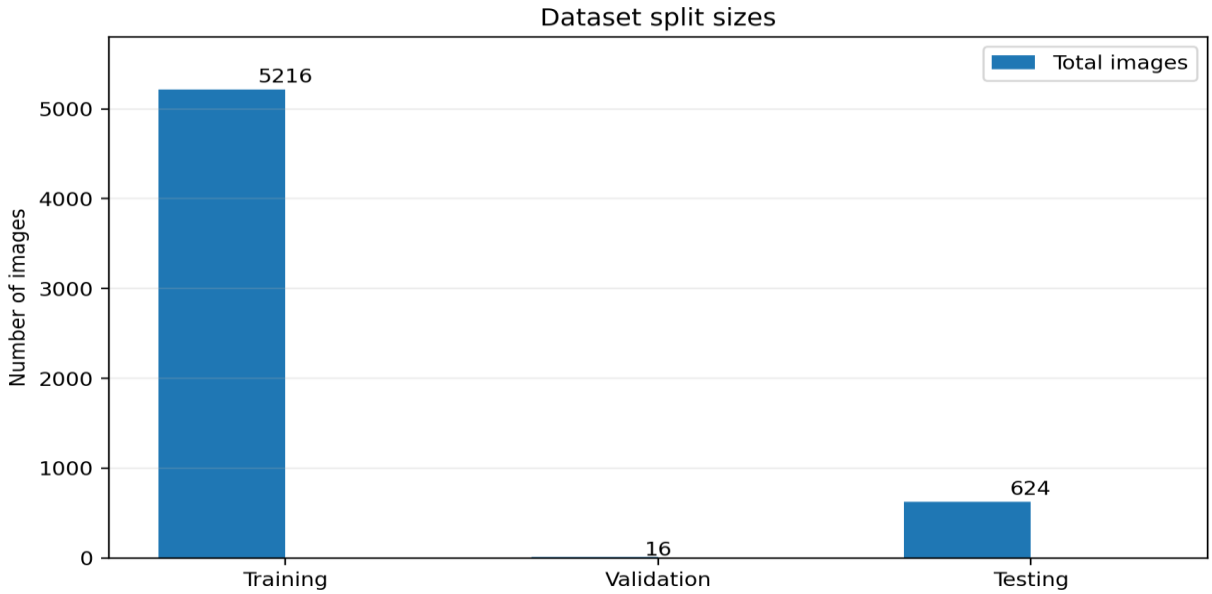


Figure 1. Total number of images in each dataset split.

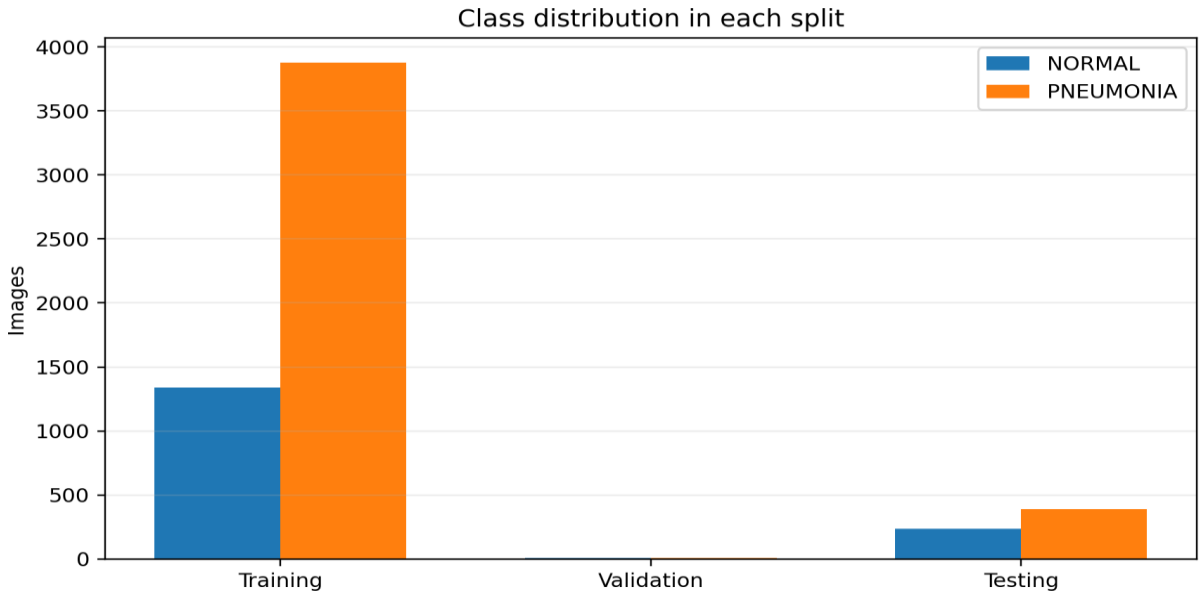


Figure 2. NORMAL and PNEUMONIA class counts by split.

The training set was imbalanced. Class weights calculated from the training labels were 1.9448 for NORMAL and 0.6730 for PNEUMONIA. The validation split contained only 16 images, which is too small for stable model selection. Class imbalance is a recognized challenge in CNN training, and its effects depend on the sampling and loss-weighting strategy [7].

2.2 Preprocessing and augmentation

Images were resized to 224 × 224 pixels and loaded in batches of 32. The custom CNN pipeline rescaled pixel values by 1/255. Training augmentation used rotation_range = 15, zoom_range = 0.2, and horizontal_flip = True. Validation and test images were rescaled without

augmentation. The MobileNetV2 pipeline used its architecture-specific preprocess_input function. Validation and test generators used shuffle=False. A random seed of 42 was set.

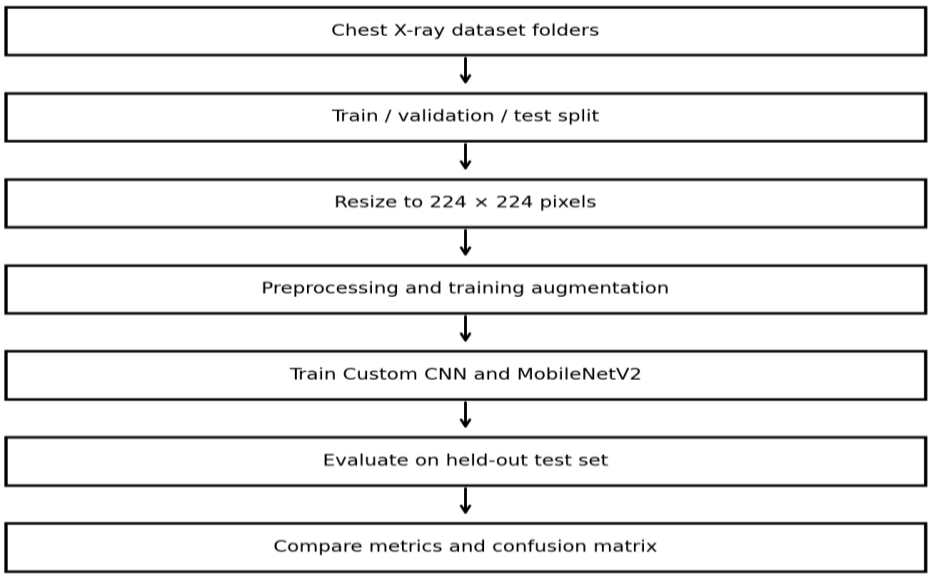
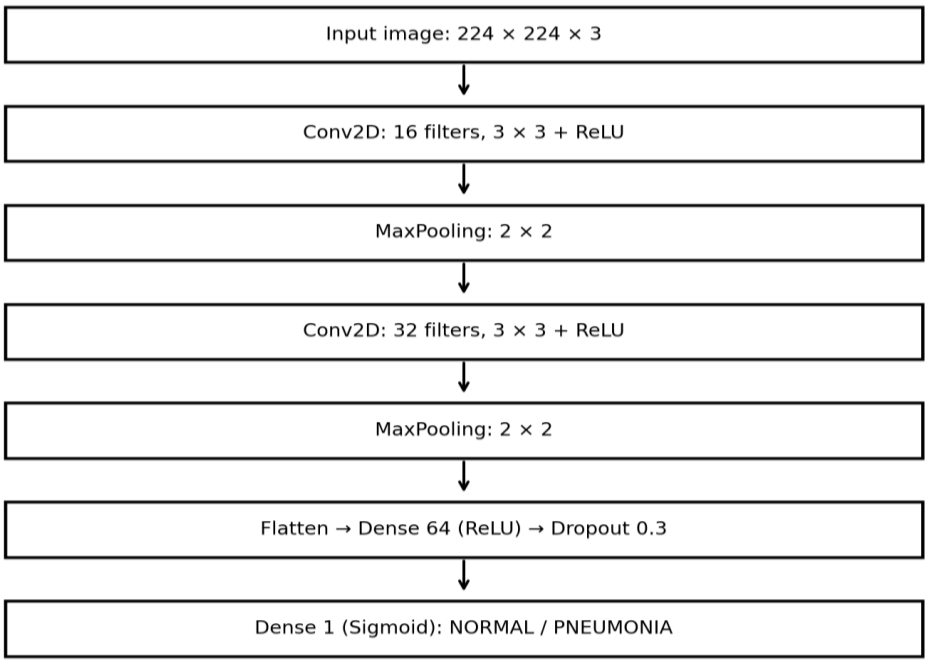


Figure 3. Experimental workflow used in this study.

2.3 Custom CNN architecture

The custom CNN accepted 224 × 224 × 3 images and consisted of Conv2D with 16 filters and ReLU, 2 × 2 max pooling, Conv2D with 32 filters and ReLU, another 2 × 2 max pooling, flattening, a 64-unit ReLU dense layer, dropout at 0.3, and a one-unit sigmoid output. It used Adam (learning rate 0.001), binary cross-entropy, accuracy, precision, and recall [4]. The model summary reported 5,977,185 trainable parameters.



Advanced Research

Figure 4. Layer sequence of the custom CNN model.

2.4 MobileNetV2 transfer-learning architecture

MobileNetV2 was initialized with ImageNet weights and include_top=False. The pretrained convolutional base was frozen. The added head consisted of global average pooling, a 64-unit ReLU dense layer, dropout at 0.3, and a sigmoid output. It used Adam (learning rate 0.001), binary cross-entropy, accuracy, precision, and recall. The model summary reported 2,340,033 total parameters, 82,049 trainable, and 2,257,984 non-trainable parameters.

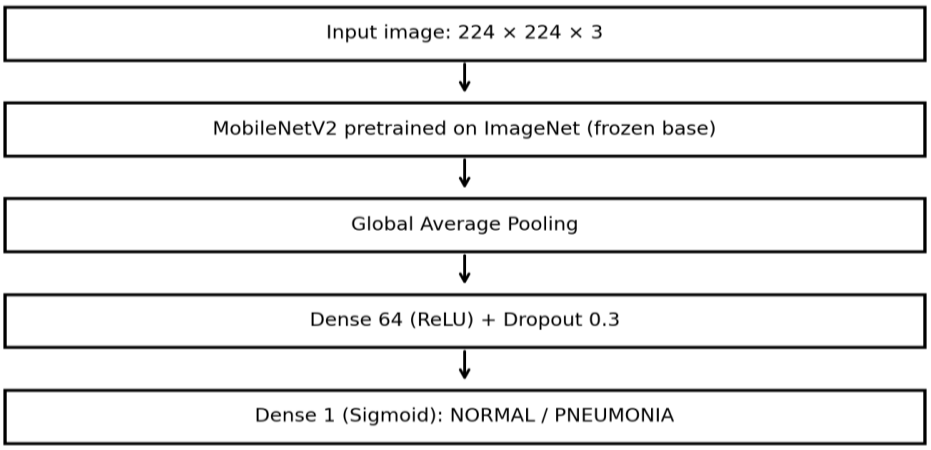


Figure 5. MobileNetV2 transfer-learning model structure.

2.5 Training and evaluation

Both models were trained for five epochs using class weights. The held-out test split contained 624 images. The classification threshold was 0.5. Confusion matrices and classification reports were generated with scikit-learn [3]. MobileNetV2 ROC-AUC was computed from predicted probabilities. The study selected the model with the higher test accuracy as the best model (Table 2).

Table 2: Training Configurations and Hyperparameters of the Custom CNN and MobileNetV2 Models.

Training setting	Custom CNN	MobileNetV2
Input size	224 × 224 × 3	224 × 224 × 3
Batch size	32	32
Epochs	5	5
Optimizer	Adam, 0.001	Adam, 0.001
Loss	Binary cross-entropy	Binary cross-entropy
Preprocessing	Rescale + augmentation	MobileNetV2 preprocess_input
Base network	Trained from scratch	ImageNet pretrained; frozen

3. Results

3.1 Overall performance

Table 3: Performance Comparison of the Custom CNN and MobileNetV2 Models on the Test Dataset.

Metric	Custom CNN	MobileNetV2
Test accuracy	86.86%	85.42%
Test loss	0.4210	0.5149
Test precision	0.8615	0.8201
Test recall	0.9410	0.9821
ROC-AUC	Not reported	0.9576

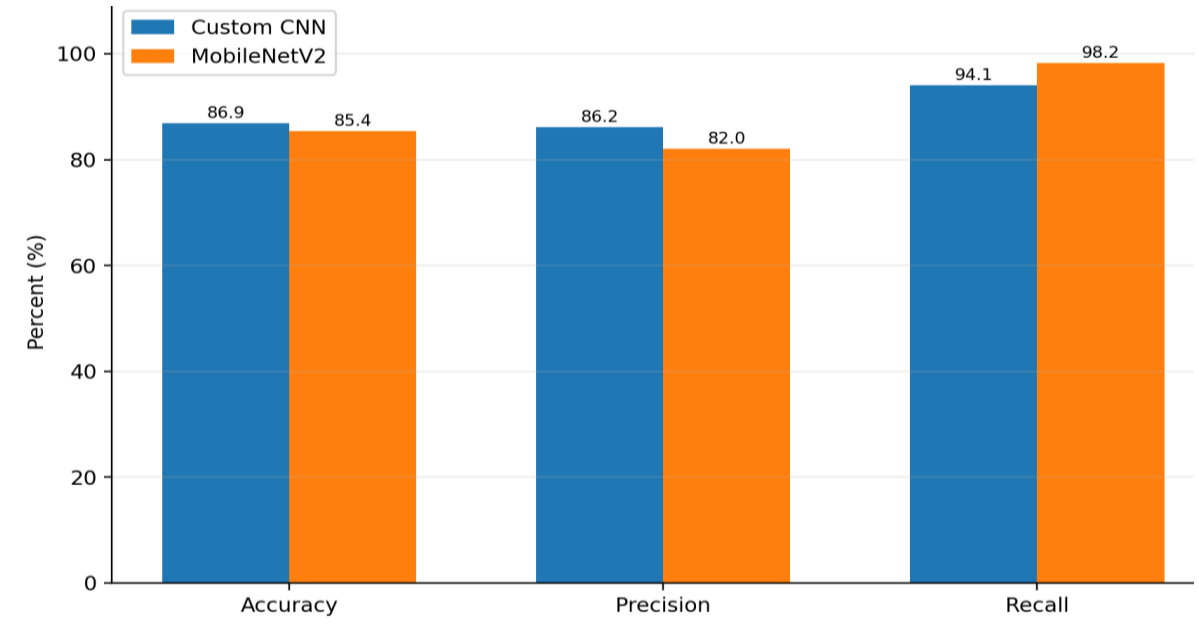


Figure 6. Test accuracy, precision, and recall comparison. Values are percentages.

The custom CNN had 1.44 percentage points higher test accuracy and lower loss. MobileNetV2 had higher recall, indicating fewer pneumonia-positive images missed at the selected threshold. ROC-AUC of 0.9576 was reported for MobileNetV2; a custom CNN AUC was not included (Table 3).

3.2 Training curves

Figures 7–10 reproduce the epoch-level training and validation metrics recorded during model training. Validation curves should be interpreted cautiously because there were only 16 validation images.

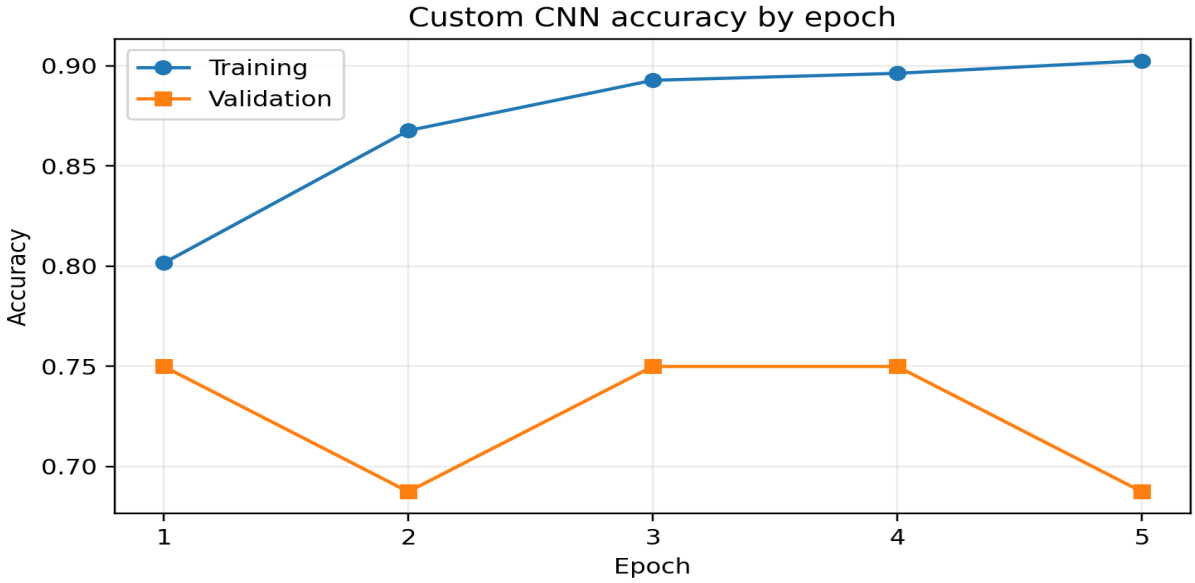


Figure 7. Custom CNN training and validation accuracy.

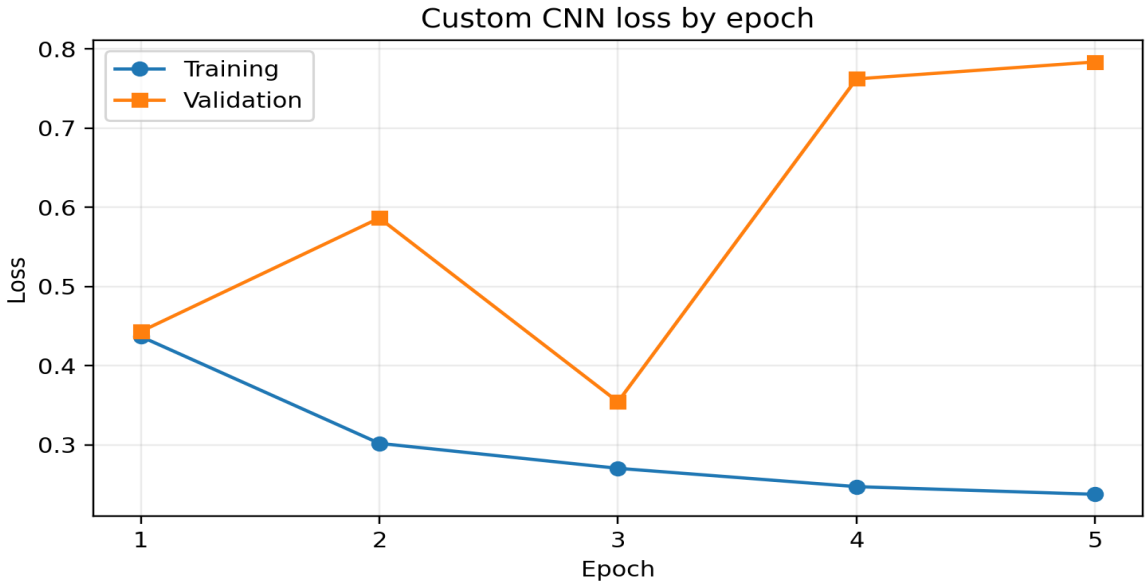


Figure 8. Custom CNN training and validation loss.

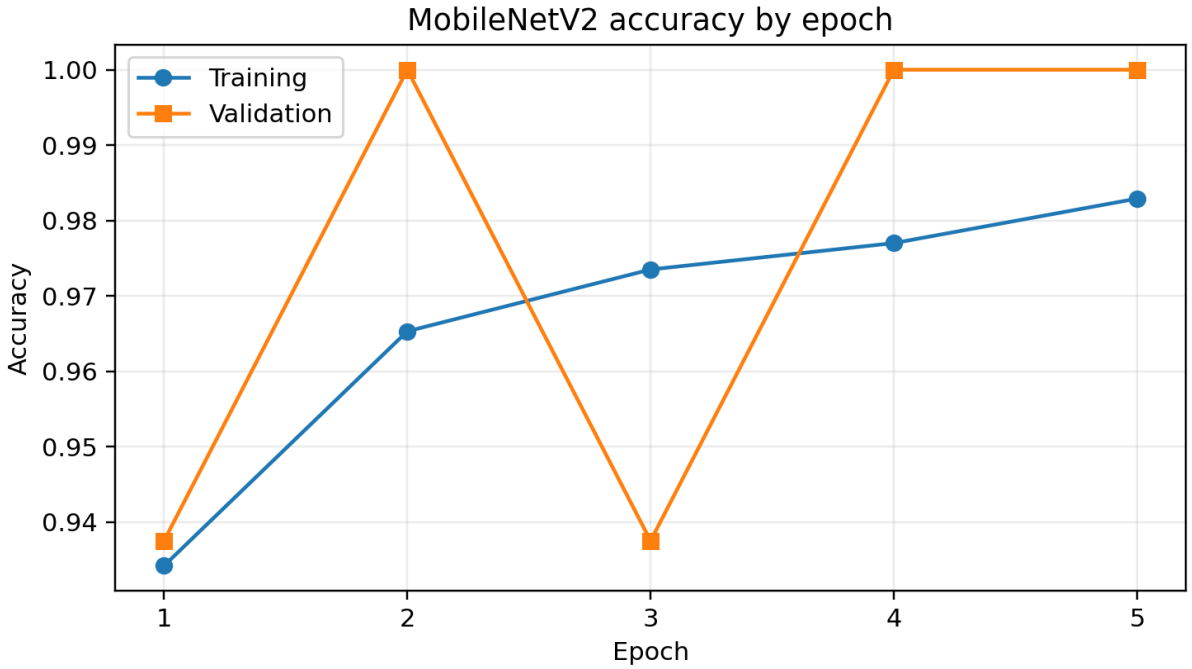


Figure 9. MobileNetV2 training and validation accuracy.

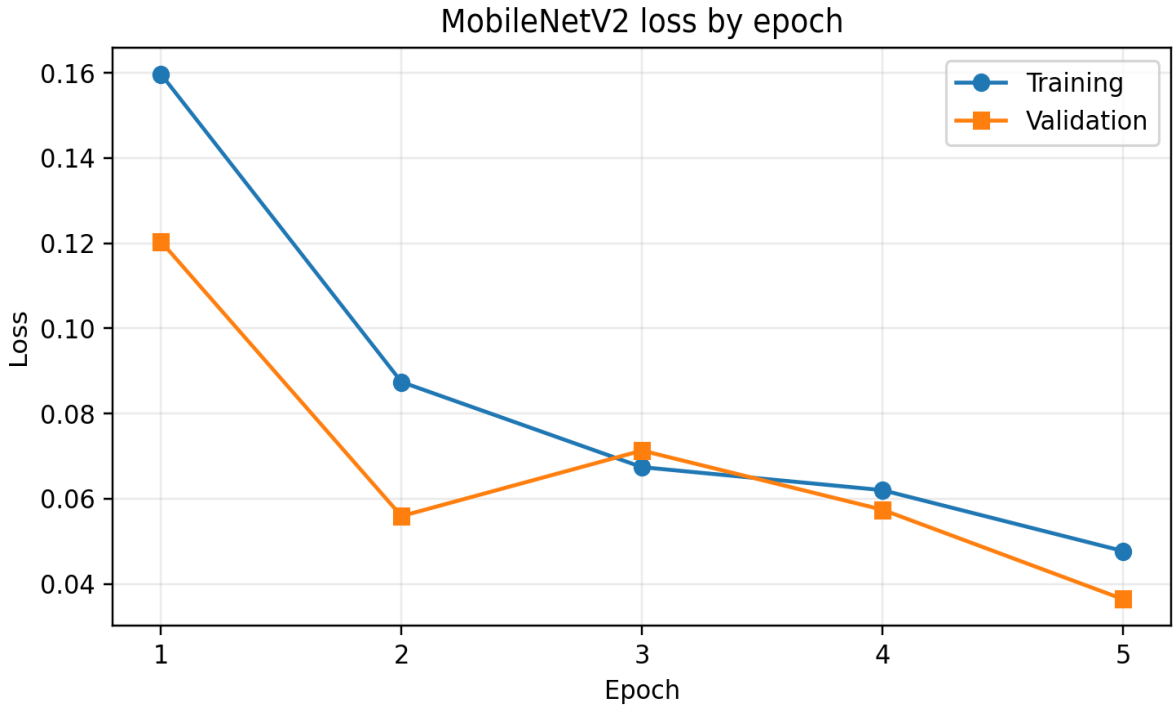


Figure 10. MobileNetV2 training and validation loss.

3.3 Custom CNN confusion matrix and class-wise report

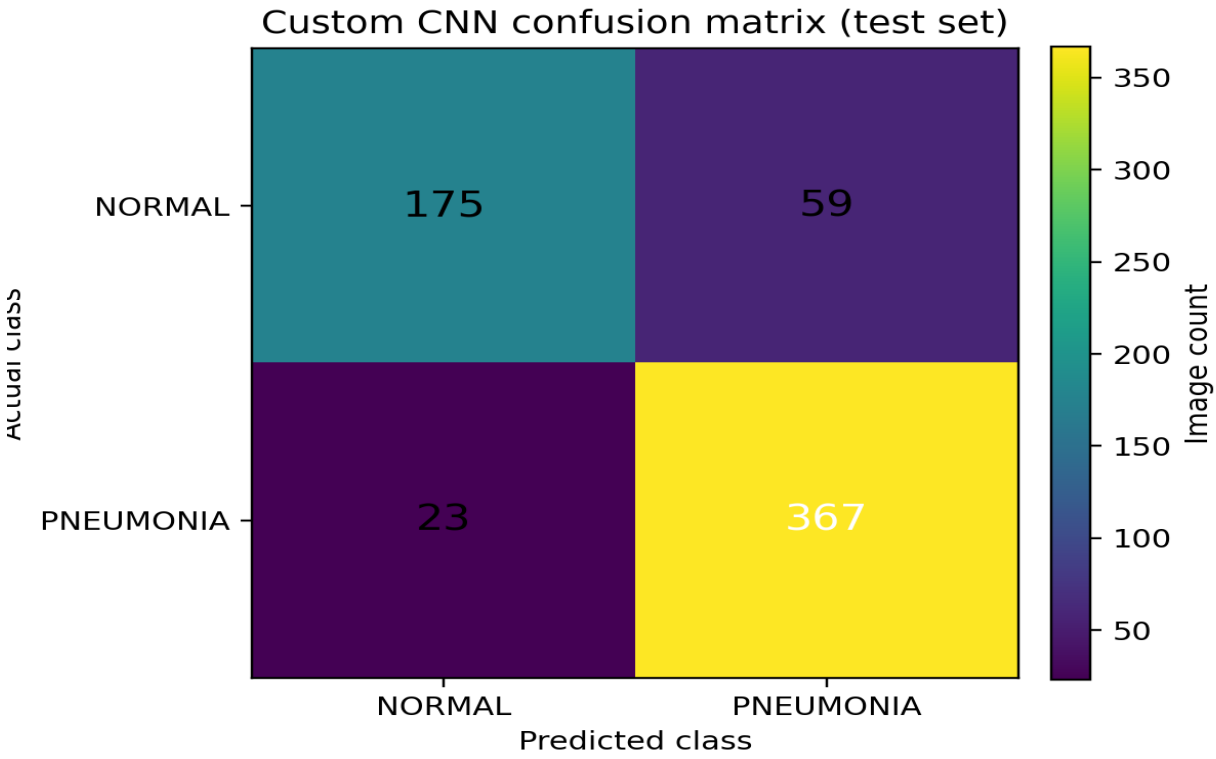


Figure 11. Custom CNN confusion matrix on the 624-image test set. Rows are actual classes, and columns are predicted classes.

Table 4: Confusion Matrix and Class-Wise Performance Metrics of the Custom CNN Model on the Test Dataset.

Actual class	Predicted NORMAL	Predicted PNEUMONIA	Support
NORMAL	175	59	234
PNEUMONIA	23	367	390

Class	Precision	Recall	F1-score	Support
NORMAL	0.88	0.75	0.81	234
PNEUMONIA	0.86	0.94	0.90	390
Macro average	0.87	0.84	0.85	624
Weighted average	0.87	0.87	0.87	624

The model correctly classified 175 NORMAL and 367 PNEUMONIA images. It produced 59 false positives and 23 false negatives. NORMAL recall was 0.75, and PNEUMONIA recall was 0.94 (Table 4).

3.4 MobileNetV2 class-wise report

Table 5: Class-Wise Classification Performance Metrics of the MobileNetV2 Model on the Test Dataset.

Class	Precision	Recall	F1-score	Support
NORMAL	0.96	0.64	0.77	234
PNEUMONIA	0.82	0.98	0.89	390
Macro average	0.89	0.81	0.83	624
Weighted average	0.87	0.85	0.85	624

The model summary reported MobileNetV2 ROC-AUC = 0.9576. It did not preserve ROC curve coordinates in the text output; therefore, this manuscript reports the AUC value rather than drawing an estimated ROC curve (Table 5).

3.5 Example single-image prediction

An unseen chest X-ray image (IM-0117-0001.jpeg), separate from the images used in the reported test-set evaluation, was passed to the trained model. The model predicted NORMAL with a pneumonia score of 0.03%. This is an illustrative single-image inference, not a clinical performance measure.

4. Discussion

The custom CNN was selected as the best model under the study's test-accuracy criterion (86.86%). However, MobileNetV2 achieved higher pneumonia recall (98.21% versus 94.10%) and an AUC of 0.9576. This highlights the difference between maximizing overall accuracy and reducing false-negative pneumonia predictions. The preferred operating point depends on the intended application and the relative costs of false negatives and false positives. The results do not establish clinical suitability.

The custom CNN confusion matrix shows 23 false negatives and 59 false positives. These errors demonstrate why accuracy alone is insufficient for medical-image classification. Class-wise recall, precision, F1-score, threshold selection, and independent validation are important. MobileNetV2 validation accuracy is particularly uncertain because the validation set contained only 16 cases.

MobileNetV2 confusion matrix

5. Limitations

- The validation set contained only 16 images, inadequate for robust model selection or threshold tuning.
- The dataset name, source, license, patient demographics, and image acquisition details were not provided.
- Patient-level separation between train, validation, and test sets was not documented.
- Both models were trained for only five epochs, and no systematic hyperparameter search was reported.
- ROC-AUC was available only for MobileNetV2.
- No external validation, confidence intervals, calibration analysis, or subgroup evaluation was reported.
- A single uploaded-image prediction does not demonstrate diagnostic reliability.

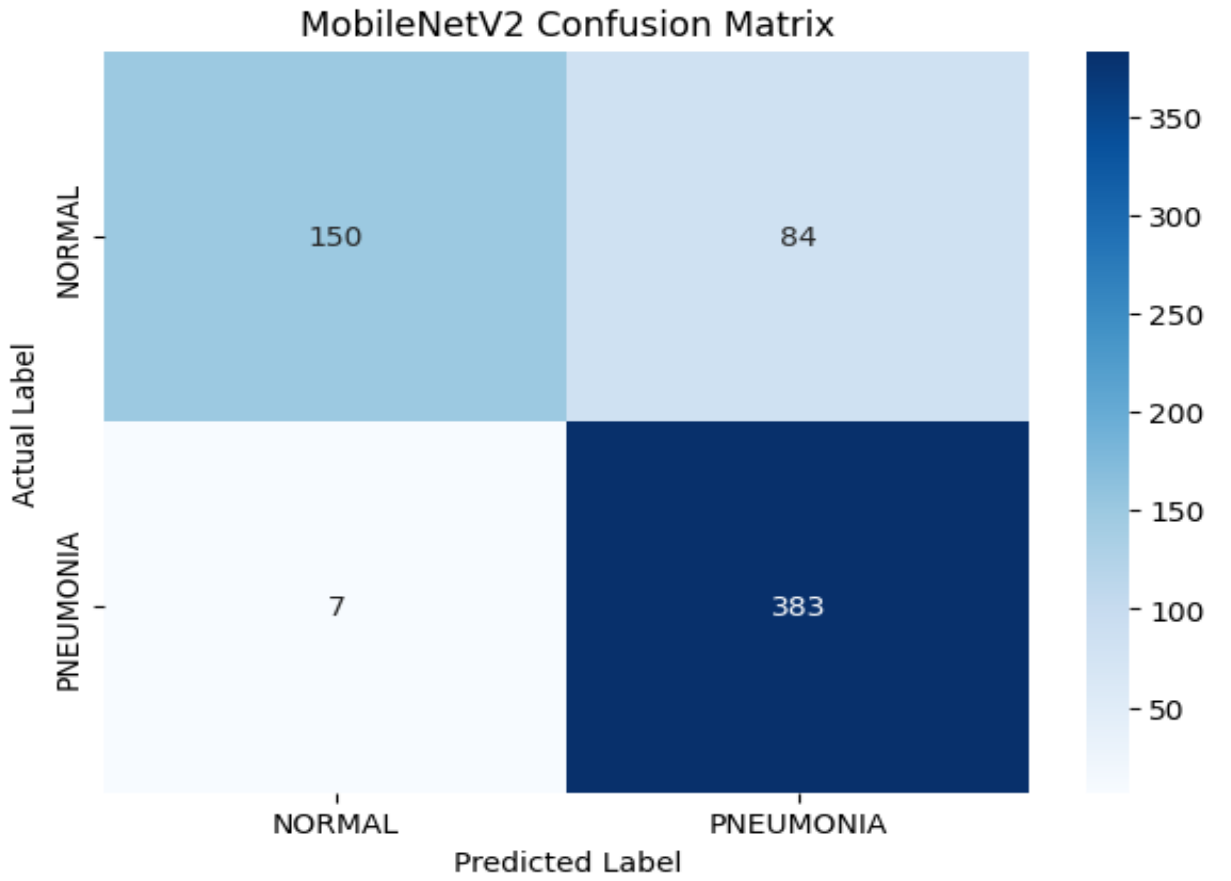


Figure 12. MobileNetV2 confusion matrix generated during model evaluation on the 624-image test set.

The corresponding experiment classification report gives NORMAL precision 0.96, recall 0.64, and F1-score 0.77; PNEUMONIA precision 0.82, recall 0.98, and F1-score 0.89.

MobileNetV2 ROC curve

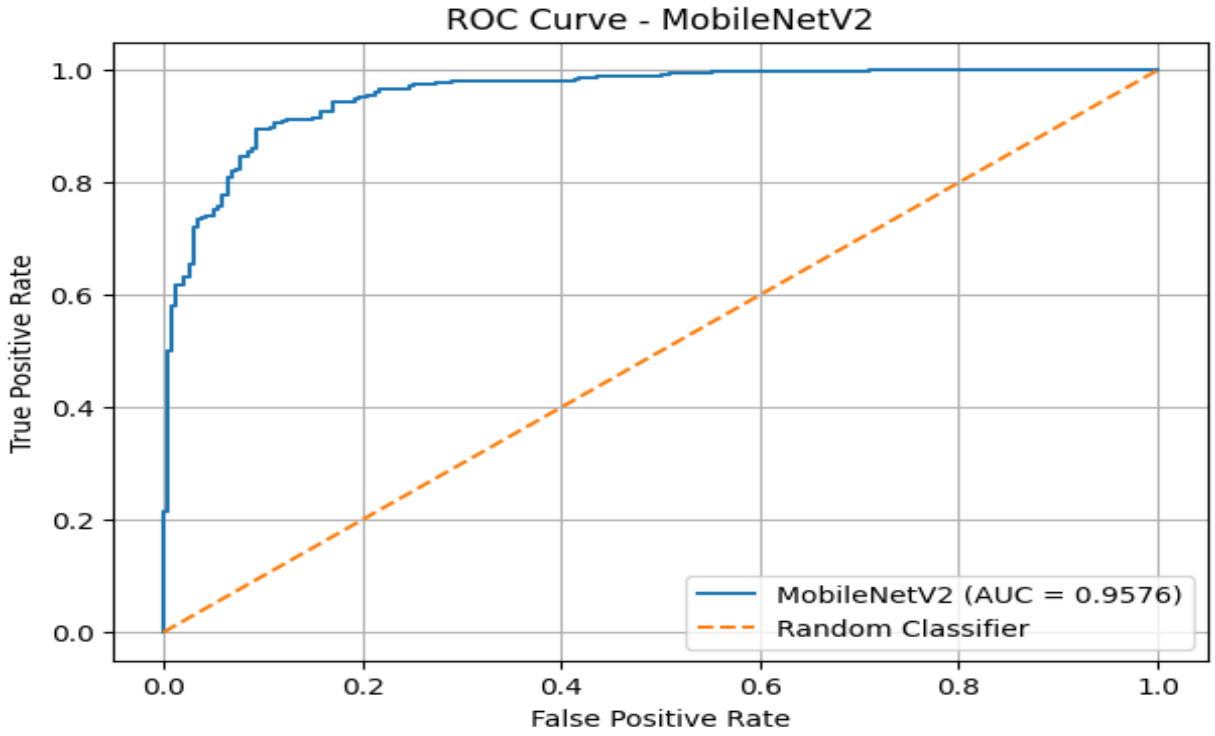


Figure 13. Original MobileNetV2 receiver operating characteristic (ROC) curve; AUC = 0.9576

The MobileNetV2 model achieved a ROC-AUC of 0.9576. The corresponding ROC curve is presented in Figure 13.

MobileNetV2 training curves

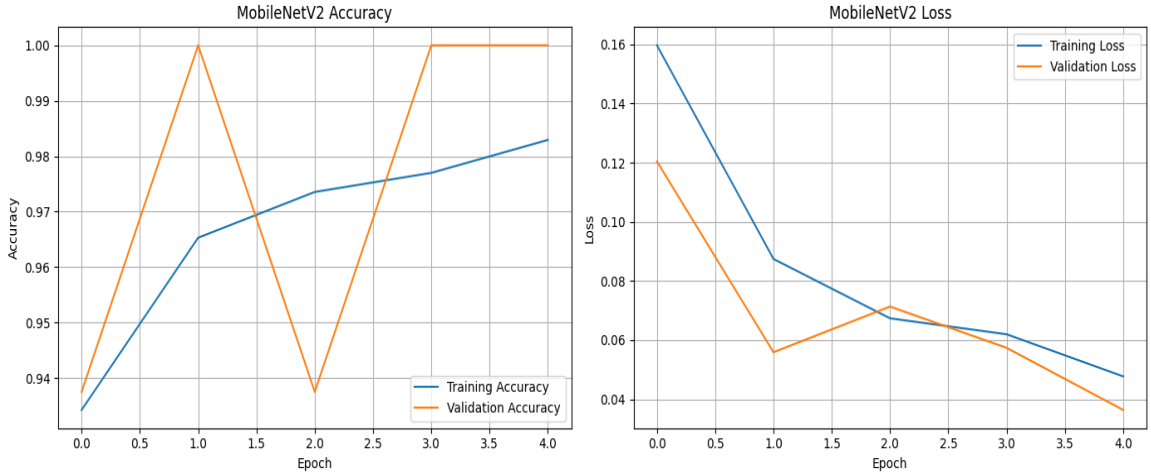


Figure 14. Original MobileNetV2 training and validation accuracy/loss plots from the training run.

Model accuracy comparison

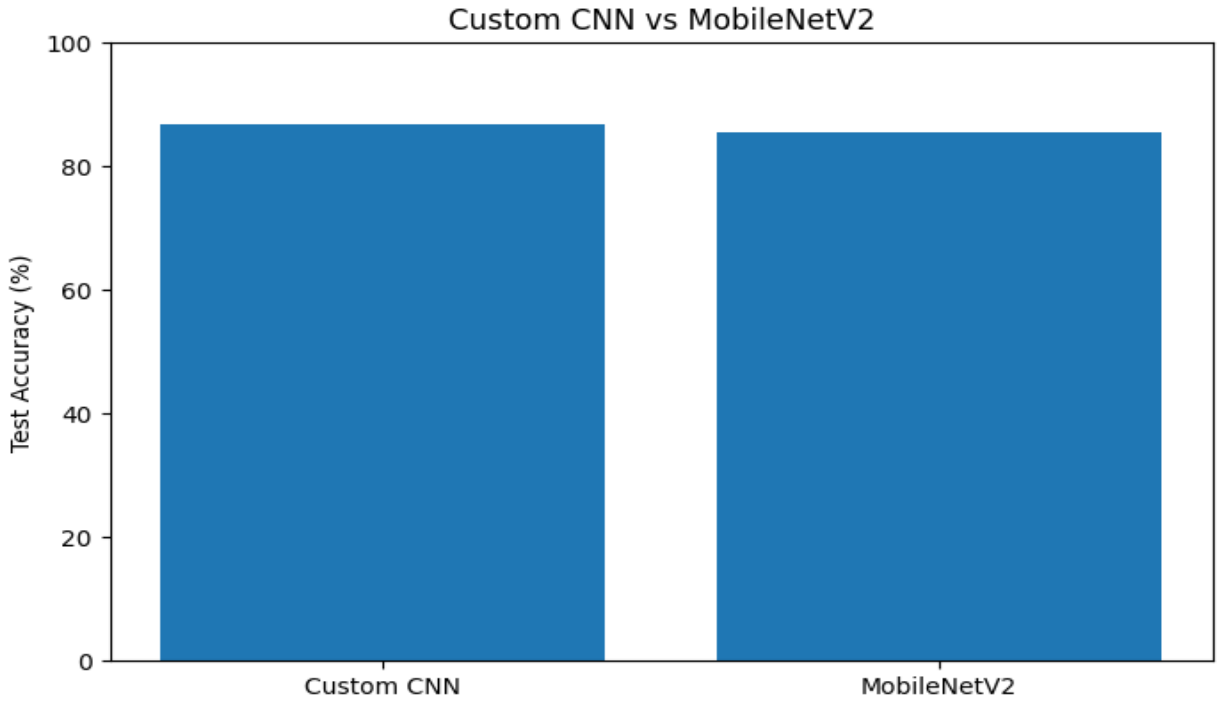


Figure 15. Experimental bar chart comparing Custom CNN and MobileNetV2 test accuracy.

Selected Best-Model Confusion Matrix

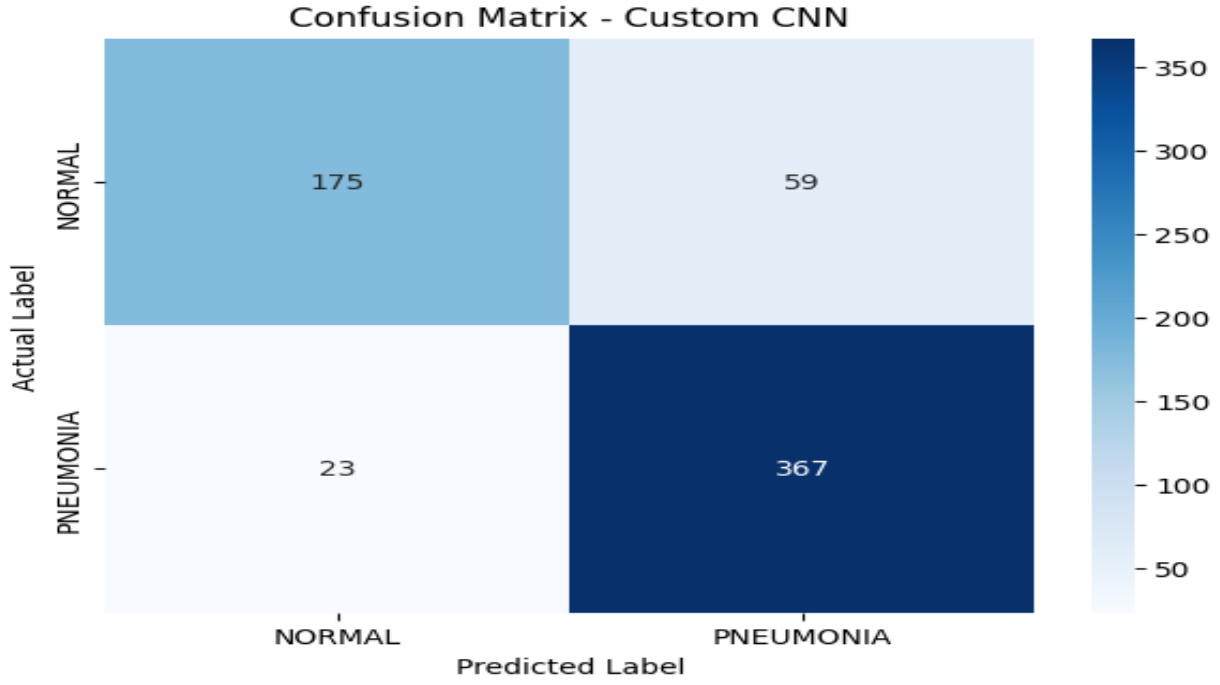


Figure 16. Experimental output for the selected best model (Custom CNN).

The selected model was chosen by comparing test accuracy. Custom CNN achieved 86.86%, compared with 85.42% for MobileNetV2. The confusion matrix is [[175, 59], [23, 367]], with

rows representing actual classes and columns representing predicted classes.

Unseen Image Prediction

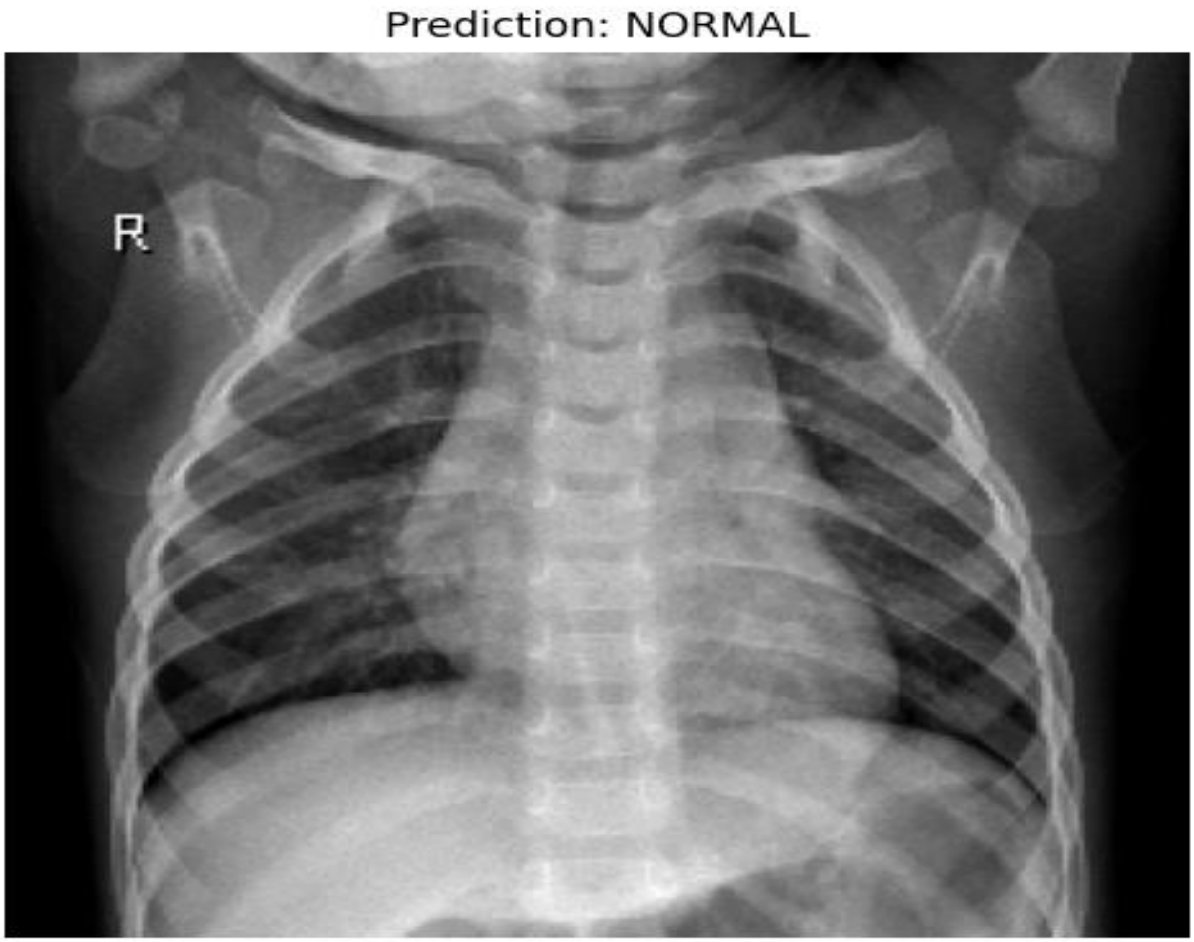


Figure 17. Unseen chest X-ray image prediction result: the model classified the image as NORMAL with a pneumonia score of 0.03%.

This is a single illustrative prediction and must not be interpreted as a clinical validation result.

Complete Test-Set Confusion-Matrix Counts

Table 6: Comparison of Confusion Matrix Results for the Custom CNN and MobileNetV2 Models on the Test Dataset.

Model	Actual NORMAL → Predicted PNEUMONIA	Actual PNEUMONIA → Predicted NORMAL	Correct predictions
Custom CNN	59	23	175 NORMAL + 367 PNEUMONIA
MobileNetV2	84	7	150 NORMAL + 383 PNEUMONIA

Note: MobileNetV2 counts are derived from the original confusion-matrix image and classification report outputs. The displayed matrix is retained above as the authoritative visual. The class labels follow the study mapping: NORMAL = 0 and PNEUMONIA = 1 (Table 6).

6. Conclusion

This experiment compared a custom CNN and MobileNetV2 for binary classification of chest X-ray images. The custom CNN achieved the higher test accuracy (86.86%) and was selected by the study's accuracy-based rule. MobileNetV2 achieved higher pneumonia recall (98.21%) and an ROC-AUC of 0.9576. These preliminary results indicate a trade-off between overall accuracy and sensitivity. Further work should use a sufficiently large validation cohort, patient-independent splits, external testing, and transparent dataset documentation. Neither model should be used to diagnose or rule out pneumonia without rigorous clinical validation.

REFERENCES

[1] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, 2018. doi: 10.1109/CVPR.2018.00474.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998. doi: 10.1109/5.726791.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[4] F. Chollet et al., "Keras," 2015. Available: https://keras.io/.

[5] D. S. Kermany 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.

[6] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv:1711.05225, 2017.

[7] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," Neural Networks, vol. 106, pp. 249–259, 2018. doi: 10.1016/j.neunet.2018.07.011.

[8] J. R. Zech et al., "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," PLOS Medicine, vol. 15, no. 11, e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

[9] Y. Li, Z. Zhang, C. Dai, Q. Dong, and S. Badrigilan, "Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images: A systematic review and meta-analysis," Computers in Biology and Medicine, vol. 123, 103898, 2020. doi: 10.1016/j.compbiomed.2020.103898.

[10] 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, 176, 2024. doi: 10.3390/jimaging10080176.

[11] J. Irvin et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

[12] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, article 317, 2019. doi: 10.1038/s41597-019-0322-0.

[13] X. Wang et al., "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2097–2106. doi: 10.1109/CVPR.2017.369.

[14] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[15] J. Hofmeister et al., "Validating the accuracy of deep learning for the diagnosis of pneumonia on chest x-ray against a robust multimodal reference diagnosis: a post hoc analysis of two prospective studies," European Radiology Experimental, 2024. doi: 10.1186/s41747-023-00416-y.

[16] P. Lakhani and B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," Radiology, vol. 284, no. 2, pp. 574–582, 2017. doi: 10.1148/radiol.2017162326.

International Journal of Advanced Research