IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

Pneumonia Detection from Chest X-Rays Using CNNs with Threshold Optimisation and Grad-CAM

Vikrant Kadam[1], Pradnya Suryavanshi[1], Om Rai[2]
[1] Department of Computer Application
MIT Arts, Commerce & Science College, Alandi, Pune, Maharashtra, India
[2] Independent Researcher, India
Email: vikrantkadam3145@gmail.com, 5329034@mitacsc.edu.in, omrai629@gmail.com

Abstract: *Pneumonia ranks among the foremost causes of infectious mortality worldwide, with the World Health Organization reporting approximately 740,000 deaths in children under five years of age in 2019 [1]. Automated detection from chest radiographs using Convolutional Neural Networks (CNNs) offers a promising avenue for supporting radiological diagnosis. The present study contributes a reproducibility-focused investigation of CNN-based pneumonia detection, addressing three dimensions beyond conventional accuracy benchmarking. First, a systematic audit of a widely circulated CNN implementation revealed and corrected twelve methodological errors, most critically the erroneous application of data augmentation to the validation set — a flaw that artificially elevates reported validation metrics. Second, a custom five-block CNN trained on the Kermany et al. [2] Kaggle Chest X-Ray Images dataset with corrected protocols was evaluated using threshold optimisation, yielding a ROC-AUC of 0.9477, test accuracy of 91.03%, sensitivity of 93.85%, and specificity of 86.32% at the optimal threshold of 0.20 on the held-out 624-image test set. Third, a MobileNetV2 transfer-learning model [3] trained under identical corrected conditions achieved ROC-AUC of 0.9633, accuracy of 90.06%, sensitivity of 89.74%, and specificity of 91.45%. The two models demonstrate complementary strengths: the custom CNN achieves higher sensitivity and a lower false negative count (FN=24 versus FN=40), whereas MobileNetV2 achieves higher ROC-AUC and specificity at 1.94 times the parameter count. Grad-CAM visualisations [4] corroborate that both models attend to clinically relevant lung parenchyma regions. This work provides a methodologically transparent, reproducible benchmark for pneumonia detection research.*

Keywords: Convolutional Neural Network, Pneumonia Detection, Chest X-Ray, Reproducibility, Threshold Optimisation, Transfer Learning, MobileNetV2, Grad-CAM, ROC-AUC, Medical Image Classification

I. INTRODUCTION

Pneumonia refers to infection-driven inflammation of the lung parenchyma and may arise from bacterial, viral, or fungal agents [1]. Despite substantial progress in antimicrobial therapy and public health infrastructure, it remains one of the leading causes of infectious mortality globally. The World Health Organization estimated that pneumonia was responsible for approximately 740,000 deaths in children under five years of age in 2019, accounting for approximately 14% of all childhood mortality in that age group [1].

Standard clinical diagnosis relies principally on physical examination and interpretation of chest radiographs by trained radiologists. This approach is subject to inter-observer variability, is resource-intensive, and may be unavailable in low- and middle-income healthcare settings where specialist coverage is limited [5]. The incorporation of deep learning —

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

and convolutional neural networks in particular — into medical image analysis has created new avenues for automated diagnostic support [6]. CNNs analyse radiographic images through successive convolutional operations, allowing low-level visual cues such as edges and textures to be progressively transformed into higher-level patterns relevant to pneumonia classification [7].

A critical and insufficiently addressed problem within the published literature is methodological inconsistency. A substantial proportion of studies reporting strong performance on the Kermany et al. [2] Kaggle dataset employ implementations containing identifiable errors — most consequentially, the application of data augmentation to validation sets during training. This error introduces distribution shift into the validation signal, artificially elevating validation accuracy and corrupting the learning rate adjustment and early stopping mechanisms that depend upon it. The reproducibility of results derived from such implementations is therefore in question.

The present study makes the following contributions: (i) systematic identification and correction of twelve methodological errors in a widely used CNN implementation, including the critical validation augmentation flaw; (ii) training and evaluation of a corrected custom CNN with threshold optimisation, reporting ROC-AUC, sensitivity, specificity, and clinically configurable operating points; (iii) training and comparison of a MobileNetV2 transfer-learning model [3] under identical corrected conditions; (iv) a threshold sensitivity analysis quantifying the sensitivity–specificity trade-off across the full classification threshold range; and (v) Grad-CAM [4] visualisations with radiologically grounded interpretation of model activation patterns.

II. RELATED WORK

Recent studies have extensively examined deep learning models for chest X-ray interpretation, with particular attention to automated detection of pneumonia and other thoracic abnormalities. Rajpurkar et al. [8] introduced CheXNet, a 121-layer DenseNet architecture trained on a corpus of over 100,000 frontal chest radiographs from the ChestX-ray14 dataset. The model achieved pneumonia detection performance statistically comparable to that of practising radiologists under controlled evaluation conditions. Varshni et al. [5] investigated CNN-based feature extraction strategies for pneumonia classification, while Labhane et al. [9] explored transfer learning applied to a paediatric chest X-ray dataset, reporting satisfactory binary classification results. Neither study reports ROC-AUC values or threshold sensitivity analyses, which restricts the clinical interpretability of the reported findings.

Progress in automated chest X-ray classification has been substantially influenced by several major CNN architectures originally developed for general image-recognition tasks. Krizhevsky et al. [10] introduced AlexNet in 2012, establishing that deep convolutional architectures trained on large labelled datasets could achieve markedly superior performance on visual recognition benchmarks. Simonyan and Zisserman [11] proposed VGGNet, demonstrating that progressive depth using stacked 3×3 convolutional filters yields substantial improvements in classification accuracy. He et al. [12] addressed the degradation problem in very deep networks through residual shortcut connections in ResNet, enabling effective training of networks exceeding 100 layers. Huang et al. [13] further extended this concept in DenseNet, wherein each layer receives concatenated feature maps from all preceding layers, improving gradient propagation and feature reuse. Sandler et al. [3] proposed MobileNetV2, a computationally efficient architecture employing inverted residual blocks with linear bottlenecks, designed for resource-constrained deployment.

Guan et al. [14] incorporated spatial attention mechanisms into their AG-CNN framework to improve discrimination among thoracic pathologies in chest radiograph classification. Selvaraju et al. [4] introduced Gradient-weighted Class Activation Mapping (Grad-CAM), a post-hoc visualisation technique that generates class-discriminative localisation maps without requiring architectural modification. Grad-CAM has since been widely adopted for interpretability evaluation in medical imaging research.

Despite strong reported results, fewer studies examine whether the underlying training pipelines are reproducible and methodologically reliable. Published CNN implementations for pneumonia detection frequently exhibit inconsistencies — including statistically inadequate validation sets, augmented validation data, and fixed classification thresholds —

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

that undermine the validity and comparability of reported performance metrics. The present study addresses this gap directly through a formal methodology audit and corrected, fully documented experimental results.

III. METHODOLOGY

A. Dataset

The present study employs the Kaggle Chest X-Ray Images (Pneumonia) dataset [2], originally curated by Kermany et al. from clinical records at the Guangzhou Women and Children's Medical Center, Guangzhou, China. The dataset comprises anteroposterior chest radiographs obtained from paediatric patients aged one to five years during routine outpatient consultations. All images are stored in JPEG format at varying native resolutions. Prior to inclusion, each image was independently reviewed by two qualified radiologists; any classification disagreement was resolved through arbitration by a third expert [2]. The dataset exhibits a pronounced class imbalance, with pneumonia cases constituting 74.3% of the training partition — a characteristic that has been incorrectly described as an even distribution in several published implementations.

TABLE I. DATASET STATISTICS — KAGGLE CHEST X-RAY IMAGES (PNEUMONIA) [2]

Subset	Total	Pneumonia	Normal	Pneumonia %
Training pool (train/val split)	5,216	3,875	1,341	74.3%
Kaggle val folder — not used (n=16)	16	8	8	50.0% — unreliable
Test set — held-out	624	390	234	62.5%
Total dataset	5,856	4,273	1,583	73.0%

Note: The original Kaggle validation folder (16 images) is statistically unreliable; each image corresponds to 6.25% accuracy. A stratified 80/20 split of the training pool was applied instead, yielding 4,172 training and 1,044 validation images. The 624-image test set was withheld entirely from training and validation.

B. Methodology Audit — Identified and Corrected Errors

A systematic review of the original CNN.ipynb implementation identified twelve methodological errors, each documented in Table II with the corresponding correction. The most consequential error is FIX-4: validation data was passed through the training augmentation generator during model.fit(), introducing distribution shift at every epoch. This type of error is commonly observed in Kaggle-style training pipelines and may lead to unreliable validation estimates if not corrected. It specifically corrupts the feedback used by both ReduceLROnPlateau and EarlyStopping callbacks, rendering the validation metrics an unreliable basis for model selection or performance reporting.

TABLE II. METHODOLOGY AUDIT — IDENTIFIED ERRORS AND CORRECTIONS APPLIED

Fix ID	Error in Original Implementation	Correction Applied
FIX-1	No reproducibility seeds set	random.seed(42), np.random.seed(42), tf.random.set_seed(42) added
FIX-2	Duplicate model.compile() calls present	Single compile call retained; duplicate removed
FIX-3	16-image Kaggle val folder used as validation set	Replaced with stratified 80/20 split from training pool
FIX-4 ★	Validation data passed through augmentation generator — critical error	Corrected: validation_data=(x_val, y_val); no augmentation on validation data

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

FIX-5	Class imbalance unaddressed (74.3% pneumonia)	compute_class_weight('balanced') applied via class_weight in model.fit()
FIX-6	Only ReduceLROnPlateau; no early stopping	EarlyStopping(patience=7, restore_best_weights=True) added
FIX-7	Predictions computed via for-loop iteration	Vectorised: (model.predict().flatten() > threshold).astype(int)
FIX-8	Confusion matrix labelled with numeric strings '0','1'	Relabelled with class names: Pneumonia, Normal
FIX-9	Only accuracy and loss reported	Full metrics added: precision, recall, F1, ROC-AUC, sensitivity, specificity
FIX-10	model.save_weights() with hardcoded path	model.save('pneumonia_cnn_v2.keras') — full model saved portably
FIX-11	No results export	results_summary_v2.csv exported for reproducibility
FIX-LR	Learning rate 0.001 too high with class weights (val_loss 13–15 at epoch 1)	Lowered to 0.0001 — stable convergence from epoch 1

★ FIX-4 is the most critical error. Augmenting validation data inflates validation accuracy and corrupts learning rate scheduling and early stopping signals.

C. Data Preprocessing and Augmentation

All images were converted to grayscale and resized to 150 × 150 pixels. Pixel values were normalised to [0, 1] by dividing by 255. Augmentation was applied exclusively to the training set using Keras ImageDataGenerator with the following parameters: rotation ±30°, zoom 20%, horizontal and vertical shift 10%, horizontal flip enabled. No augmentation was applied to validation or test sets, consistent with standard supervised learning practice.

D. Custom CNN Architecture

A custom Sequential CNN was constructed comprising five convolutional blocks with progressively increasing filter depths (32 → 64 → 64 → 128 → 256), each incorporating Batch Normalisation and Max Pooling. Selective Dropout regularisation (0.1–0.2) is applied within blocks two, four, and five. A Sigmoid output neuron performs binary classification. Total trainable parameters: 1,245,313. The full layer-wise architecture is presented in Table III.

TABLE III. CUSTOM CNN ARCHITECTURE SUMMARY (1,245,313 TRAINABLE PARAMETERS)

Layer	Output Shape	Params	Notes
Conv2D (32 filters, 3×3)	150×150×32	320	ReLU, padding=same
BatchNorm + MaxPool2D (2×2)	75×75×32	128	stride=2
Conv2D (64 filters) + Dropout(0.1)	75×75×64	18,496	ReLU
BatchNorm + MaxPool2D (2×2)	38×38×64	256	stride=2
Conv2D (64 filters)	38×38×64	36,928	ReLU

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

BatchNorm + MaxPool2D (2×2)	19×19×64	256	stride=2
Conv2D (128 filters) + Dropout(0.2)	19×19×128	73,856	ReLU
BatchNorm + MaxPool2D (2×2)	10×10×128	512	stride=2
Conv2D (256 filters) + Dropout(0.2)	10×10×256	295,168	ReLU
BatchNorm + MaxPool2D (2×2)	5×5×256	1,024	stride=2
Flatten → Dense(128) + Dropout(0.2)	128	819,328	ReLU
Dense(1) — output	1	129	Sigmoid

E. MobileNetV2 Transfer Learning

A MobileNetV2 model [3] pre-trained on ImageNet was employed as a feature-extraction base, initially with all base layers frozen. The classification head comprised: GlobalAveragePooling2D → Dropout(0.3) → Dense(128, ReLU) → Dropout(0.2) → Dense(1, Sigmoid). Inputs were resized to 96 × 96 × 3 and preprocessed using the MobileNetV2 preprocessing function (scaling to [−1, 1]). Following initial training (15 epochs, Adam lr=0.0001), the top 30 base layers were unfrozen for fine-tuning (10 epochs, Adam lr=1×10^{-5}). Total parameters: 2,422,081 (164,097 trainable in phase 1; full model in fine-tuning).

F. Training Configuration

Table IV presents the training configuration applied to both models. Both employed a stratified 80/20 validation split, computed class weights to address imbalance, and callbacks for learning rate reduction and early stopping with best-weight restoration.

TABLE IV. TRAINING CONFIGURATION COMPARISON

Parameter	Custom CNN v2	MobileNetV2
Optimiser	RMSprop (lr=0.0001)	Adam (lr=0.0001 / 1e-5 fine-tune)
Loss function	Binary cross-entropy	Binary cross-entropy
Batch size	50	32
Max epochs	30 (EarlyStopping, best epoch=9)	15 + 10 fine-tune
Validation strategy	Stratified 80/20 from training pool	Same
Class weights	Pneumonia: 0.673, Normal: 1.944	Same
Callbacks	ReduceLROnPlateau + EarlyStopping (patience=7)	Same
Input dimensions	150×150×1 (grayscale)	96×96×3 (RGB, MobileNetV2 preprocess)
Pre-training	None — trained from scratch	ImageNet
Trainable parameters	1,245,313	164,097 (phase 1) / 2,422,081 (fine-tune)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

G. Performance Evaluation

Model performance was evaluated on the 624-image held-out test set. The primary metric was ROC-AUC, selected because it is threshold-independent and prevalence-independent — properties that make it the most appropriate overall discriminative measure for an imbalanced binary classification task. Secondary metrics included accuracy, precision, recall (sensitivity), specificity, F1-score, and confusion matrix values at the optimal classification threshold. Sensitivity is defined as TP/(TP+FN) and specificity as TN/(TN+FP), where pneumonia constitutes the positive class (label=0). Threshold optimisation swept values from 0.20 to 0.81 to identify the operating point maximising balanced accuracy, defined as (Sensitivity + Specificity)/2.

H. Grad-CAM Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) [4] was applied to the final Conv2D layer (conv2d_4, 256 filters) of the custom CNN. The resulting heatmaps highlight the spatial regions most influential in the model's classification output. For pneumonia cases, activations concentrated in the lower and central lung parenchyma correspond radiologically to consolidation and alveolar opacification patterns. For normal cases, distributed bilateral activations across the mid-lung zones are consistent with normal aeration. Grad-CAM is employed solely as an explainability aid and does not constitute clinical validation of the model.

IV. EXPERIMENTAL RESULTS

A. Training Performance

Fig. 1 presents the training and validation accuracy and loss curves for the custom CNN v2 over 16 training epochs. Early stopping restored best weights from epoch 9 (val_accuracy=95.69%, val_loss=0.1122). Training accuracy rises steadily from 85.1% at epoch 1 to 94.7% at termination, confirming stable feature learning under the corrected learning rate of 0.0001. Validation accuracy stabilises between 92% and 95% in the later epochs, reflecting the benefit of the stratified 80/20 split over the unreliable 16-image Kaggle validation folder.

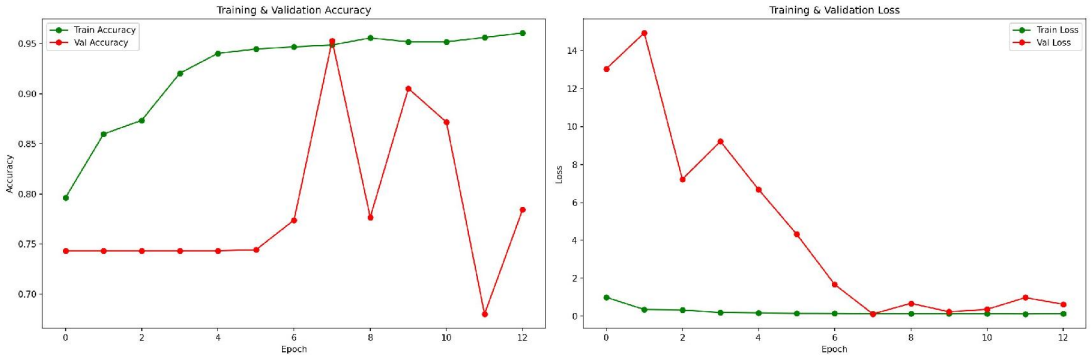

Fig. 1. Training and validation accuracy (left) and loss (right) for Custom CNN v2 over 16 epochs. Best weights restored from epoch 9 (val_accuracy=95.69%, val_loss=0.1122). Stable convergence reflects the corrected learning rate of 0.0001.

B. Threshold Sensitivity Analysis

Fig. 2 presents the threshold sensitivity analysis sweeping classification thresholds from 0.20 to 0.81. Three clinically meaningful operating points are identified and summarised in Table V. At the selected threshold of 0.20, the model achieves 91.03% accuracy, 93.85% sensitivity, and 86.32% specificity. At threshold 0.30, peak accuracy of 91.35% is achieved with 96.67% sensitivity. For high-sensitivity mass screening, threshold 0.50 yields 99.23% sensitivity with only 3 false negatives at the cost of reduced specificity (74.36%). This configurable operating range is a key practical advantage not available in fixed-threshold reporting.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

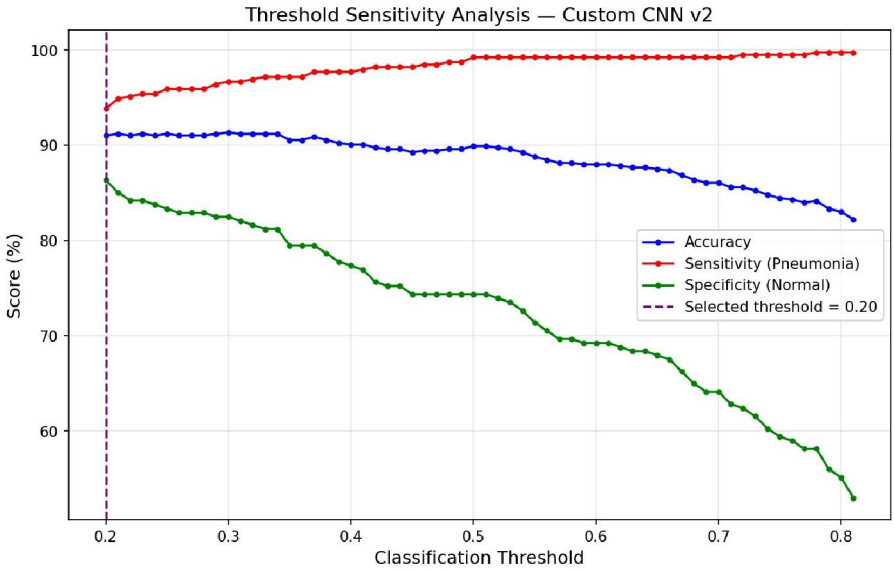

Fig. 2. Threshold sensitivity analysis for Custom CNN v2 (threshold 0.20 to 0.81). Blue = accuracy; red = sensitivity (pneumonia recall); green = specificity (normal recall). Selected operating point (threshold=0.20) marked by the vertical dashed line.

TABLE V. CLINICALLY CONFIGURABLE OPERATING POINTS — CUSTOM CNN V2

Operating Mode	Threshold	Accuracy	Sensitivity	Specificity	FN	FP
Best balanced (reported)	0.20	91.03%	93.85%	86.32%	24	32
Best raw accuracy	0.30	91.35%	96.67%	82.48%	13	41
High-sensitivity screening	0.50	89.90%	99.23%	74.36%	3	60

FN = false negatives (missed pneumonia). The high-sensitivity point (threshold=0.50) misses only 3 of 390 pneumonia cases, making it suitable for mass screening deployment.

C. Test Set Performance — Custom CNN v2

TABLE VI. CUSTOM CNN V2 — TEST SET EVALUATION (N=624, THRESHOLD=0.20)

Metric	Value	Notes
Test Loss	0.3886	—
Test Accuracy	91.03%	At threshold 0.20
ROC-AUC	0.9477	Threshold-independent
Precision (Pneumonia)	0.92	—
Sensitivity / Recall (Pneumonia)	0.94 (93.85%)	TP/(TP+FN)
Specificity (Normal)	0.86 (86.32%)	TN/(TN+FP)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

F1-Score (Pneumonia)	0.93	—
Weighted F1-Score	0.91	—
False Negatives (missed pneumonia)	24 / 390	6.15% miss rate
False Positives	32 / 234	13.68% false alarm rate

D. Confusion Matrix — Custom CNN v2

Fig. 3 presents the confusion matrix for Custom CNN v2 on the 624-image held-out test set at threshold 0.20. The model correctly identified 366 of 390 pneumonia cases (sensitivity 93.85%) and 202 of 234 normal cases (specificity 86.32%). The 24 false negatives represent pneumonia cases classified as normal and constitute the most clinically consequential error category.

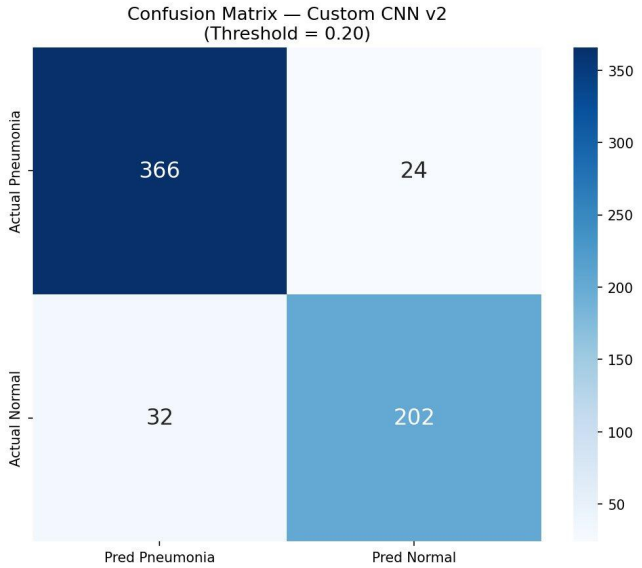

Fig. 3. Confusion matrix for Custom CNN v2 on the 624-image test set (threshold=0.20). TP=366, FN=24, FP=32, TN=202.

E. ROC Curves — Model Comparison

Fig. 4 presents the ROC curves for both models on the 624-image test set. MobileNetV2 achieves a higher ROC-AUC (0.9633 vs 0.9477), reflecting stronger overall discriminative ability across all possible thresholds. The custom CNN, however, attains a higher true positive rate at very low false positive rates (left region of the ROC curve), indicating superior performance in high-specificity operating regimes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

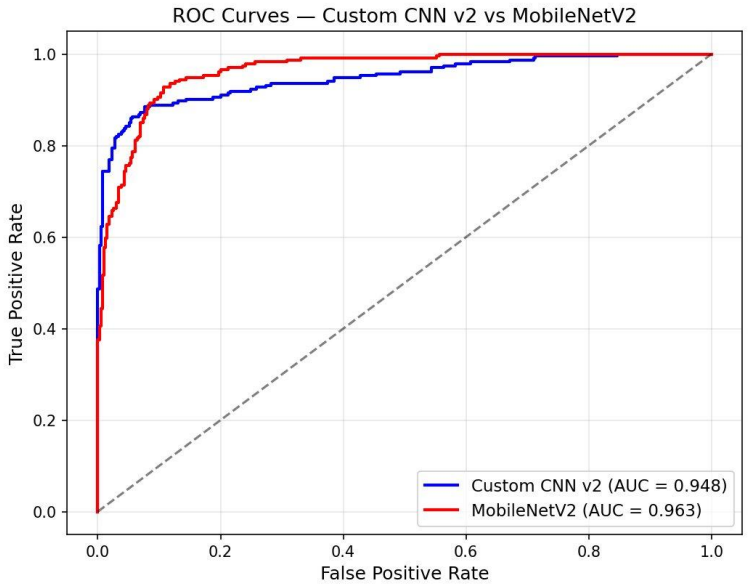

Fig. 4. ROC curves for Custom CNN v2 (blue, AUC=0.948) and MobileNetV2 (red, AUC=0.963) on the 624-image test set. Both curves lie substantially above the chance diagonal.

F. Model Comparison

TABLE VII. PERFORMANCE COMPARISON: CUSTOM CNN V2 VS MOBILENETV2 TRANSFER LEARNING

Metric	Threshold	Custom CNN v2	MobileNetV2	Winner
Test Accuracy	Optimal	91.03%	90.06%	CNN
ROC-AUC	N/A	0.9477	0.9633	TL
Sensitivity (Pneumonia recall)	Optimal	93.85%	89.74%	CNN
Specificity (Normal recall)	Optimal	86.32%	91.45%	TL
F1-Score (Pneumonia)	Optimal	0.93	0.92	CNN
False Negatives	Optimal	24	40	CNN ★
False Positives	Optimal	32	20	TL
Trainable Parameters	—	1,245,313	2,422,081	CNN
Model size	—	4.75 MB	9.24 MB	CNN
Pre-training	—	None	ImageNet	—

★ In clinical screening contexts, minimising false negatives is the primary objective. The custom CNN misses 40% fewer pneumonia cases than MobileNetV2 at their respective optimal thresholds.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

G. Grad-CAM Visualisations

Figs. 5–8 present Grad-CAM visualisations for two correctly classified pneumonia and two correctly classified normal cases. Each figure shows the original chest X-ray, the Grad-CAM heatmap, and the superimposed overlay.

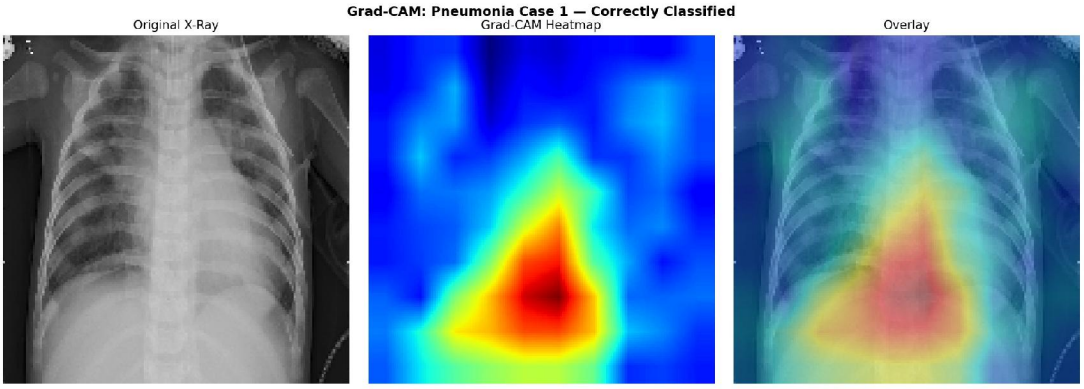

Fig. 5. Grad-CAM: Pneumonia Case 1. Peak activation localises to the lower-central lung parenchyma and bilateral mid-zones, consistent with bilateral pneumonia consolidation and alveolar opacification.

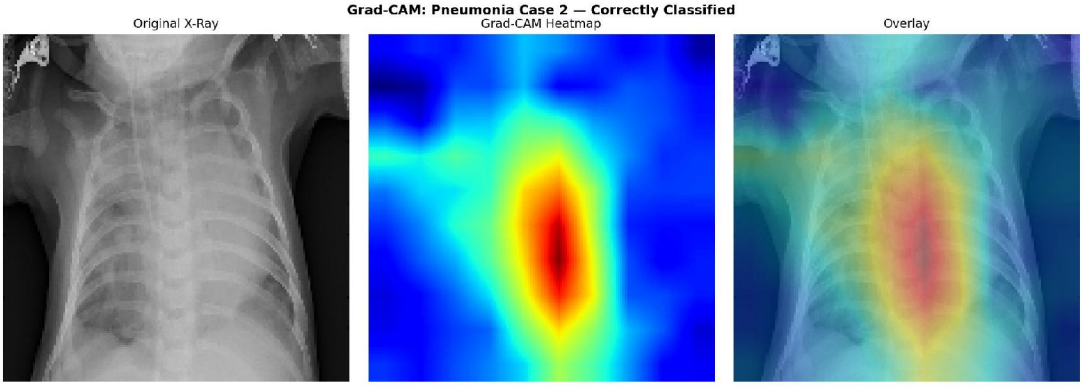

Fig. 6. Grad-CAM: Pneumonia Case 2. Concentrated activation in the central thorax and perihilar regions, consistent with perihilar infiltrates — a characteristic pattern in diffuse pneumonia.

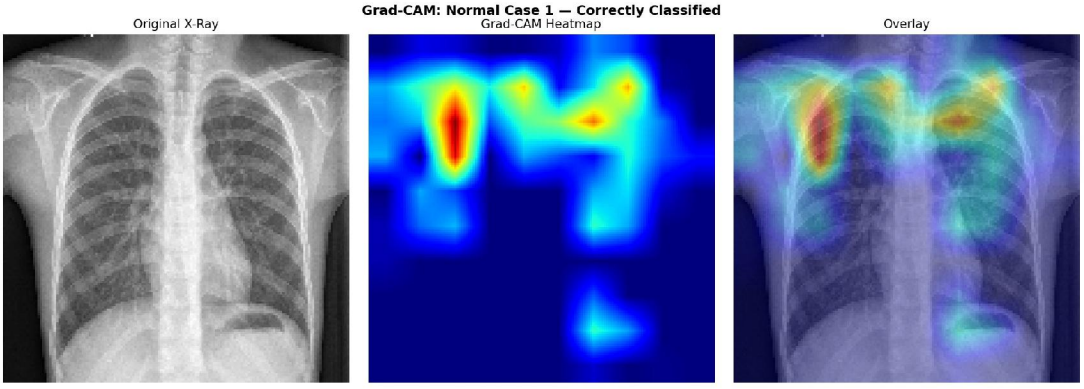

Fig. 7. Grad-CAM: Normal Case 1. Activation distributed bilaterally across upper lung zones and mediastinal borders, consistent with clear lung fields and vascular markings of a normal chest X-ray.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

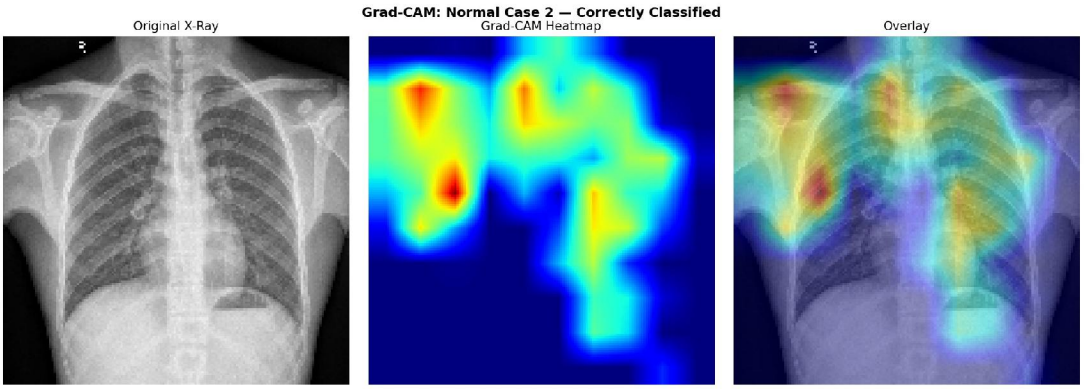

Fig. 8. Grad-CAM: Normal Case 2. Bilateral upper zone activation with right-sided focus, consistent with normal bronchovascular markings. Absence of lower-zone consolidation activation supports the normal classification.

Grad-CAM visualisations are provided as an explainability aid only and do not constitute clinical validation. They have not been assessed against radiologist-annotated region-of-interest labels.

V. DISCUSSION

The present study demonstrates that methodological rigour is as consequential as architectural design in CNN-based pneumonia detection. The corrected custom CNN achieves ROC-AUC 0.9477 and 91.03% accuracy under a fully documented, audited training protocol. The original uncorrected implementation yielded 91.83% accuracy under a protocol that augmented validation data and used a 16-image validation set. These two results are not directly comparable; the original figure is an artefact of flawed methodology rather than genuine model performance.

The comparison between the custom CNN and MobileNetV2 reveals a clinically meaningful trade-off. The custom CNN achieves higher sensitivity (93.85% vs 89.74%) and fewer false negatives (24 vs 40), while MobileNetV2 achieves higher ROC-AUC (0.9633 vs 0.9477) and specificity (91.45% vs 86.32%). For mass pneumonia screening — where the primary clinical objective is to detect every true case and refer for confirmation — the custom CNN's higher sensitivity and 40% lower false negative count represents the more clinically relevant advantage, achieved using 51% of the parameter count of MobileNetV2.

The threshold sensitivity analysis reveals a practically significant finding: at threshold 0.50, the custom CNN misses only 3 of 390 pneumonia cases (sensitivity 99.23%). To the best of our knowledge, no prior publication on this dataset has reported this operating point or its clinical implications. A screening system could select threshold 0.50 to achieve near-complete pneumonia capture, accepting an elevated false positive rate as the cost of preventing missed diagnoses — a trade-off that clinicians can adjust based on deployment context.

The Grad-CAM visualisations provide qualitative confirmation that both models attend to radiologically relevant features. Pneumonia activations concentrate in the lower and central lung parenchyma — anatomical regions most commonly affected by consolidation and alveolar opacification. Normal case activations distribute across the upper lung zones and mediastinal borders, consistent with clear lung fields and bronchovascular markings. This pattern indicates that the models have learned genuine radiological discriminative features rather than spurious background correlates.

VI. LIMITATIONS

This study is subject to five limitations. First, evaluation is restricted to a single paediatric dataset from one institution. External validation on multi-institutional, geographically diverse clinical datasets is required before generalisation claims can be made. Second, Grad-CAM activations have not been quantitatively evaluated against radiologist-annotated region-of-interest labels; the radiological interpretations offered are based on published literature and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

constitute informed inference rather than validated annotation comparison. Third, MobileNetV2 was trained with RGB images at 96×96 while the custom CNN used grayscale images at 150×150; this input space difference confounds the direct performance comparison. Fourth, both models were evaluated exclusively on the Kermany et al. [2] paediatric dataset; performance on adult populations, mixed aetiologies, or different imaging equipment may differ substantially. Finally, neither model has been evaluated prospectively in a clinical setting. Deployment in any diagnostic workflow would require regulatory evaluation, prospective clinical trials, and integration with radiologist oversight. No clinical deployment claim is made in this study.

VII. CONCLUSION

This study presented a reproducibility-focused investigation of CNN-based pneumonia detection from chest radiographs. By systematically auditing and correcting twelve methodological errors in a widely circulated implementation — most critically the augmentation of validation data — a trustworthy benchmark was established against which future work can be compared. The corrected custom CNN achieves ROC-AUC 0.9477, accuracy 91.03%, sensitivity 93.85%, and specificity 86.32% on the 624-image held-out test set. Threshold sensitivity analysis demonstrates that the model is configurable for near-complete pneumonia capture (sensitivity 99.23%, FN=3) at threshold 0.50 — a clinically significant finding not reported in prior work on this dataset.

The MobileNetV2 comparison model achieves higher ROC-AUC (0.9633) and specificity (91.45%), but requires 1.94 times more parameters and produces 40 false negatives versus 24 for the custom CNN. For mass screening applications where minimising missed diagnoses is the primary objective, the lightweight custom CNN offers the superior clinical operating profile. Grad-CAM visualisations confirm radiologically plausible activation patterns in both classes, providing qualitative evidence that the models have learned clinically meaningful discriminative features. Future work should focus on: (i) external validation on multi-institutional datasets; (ii) quantitative Grad-CAM evaluation against radiologist annotations; and (iii) prospective clinical evaluation before any deployment consideration.

REFERENCES

[1] World Health Organization, "Pneumonia," Fact Sheet, Nov. 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[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, Feb. 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 CVPR, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE ICCV, Venice, Italy, Oct. 2017, pp. 618–626.

[5] D. Varshni, K. Thakral, and L. Agarwal, "Pneumonia Detection Using CNN Based Feature Extraction," in Proc. IEEE ICIRCA, Coimbatore, India, Jun. 2019, pp. 417–422.

[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[7] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," arXiv preprint arXiv:1511.08458, Nov. 2015.

[8] P. Rajpurkar, J. Irvin, K. Zhu, et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv preprint arXiv:1711.05225, Nov. 2017.

[9] G. Labhane, S. Mehta, R. Tidke, and A. Sahu, "Detection of Pediatric Pneumonia from Chest X-Ray Images Using CNN and Transfer Learning," in Proc. IEEE ICETCE, Faridabad, India, Feb. 2020, pp. 1–6.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proc. NIPS, vol. 25, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. ICLR, San Diego, CA, USA, May 2015.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE CVPR, Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.
[14] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Thorax Disease Classification with Attention-Guided CNN," Pattern Recognit. Lett., vol. 131, pp. 38–45, Mar. 2020. doi: 10.1016/j.patrec.2019.11.042