JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

EXPLAINABLE DEEP LEARNING FOR PNEUMONIA DETECTION FROM CHEST X-RAYS

Hadiya Ali [1], Hamza Gulshad [2], Imran Ullah [3], Maha Asghar [4], Muhammad Adnan [5]

Affiliations

[1] Department of Computer Science, Abbottabad University of Science and Technology, Pakistan

[2] Department of Information Technology, University of Haripur, Pakistan

[3] Department of Computer Science, Hazara University Dhodial, Mansehra, Pakistan

[4] Department of Computer Science, University of Lahore, Sargodha Campus, Pakistan

[5] Faculty of Computing, Abasyn University, Pakistan

Corresponding Author's Email

[3] imranullah.ajm@gmail.com

License:

CC BY NC SA

ABSTRACT

Pneumonia is a leading cause of morbidity and mortality worldwide, and timely diagnosis from chest X-ray imaging is essential for effective treatment. Manual interpretation of radiographs is time-consuming and depends on the availability of experienced radiologists, motivating automated diagnostic support. Deep learning achieves strong accuracy in medical image classification, but its opacity limits clinical trust, since practitioners require justification for each prediction. This paper presents an explainable deep learning framework for pneumonia detection from chest X-rays that combines transfer learning with Gradient-weighted Class Activation Mapping (Grad-CAM). We fine-tune two pre-trained convolutional neural network backbones, MobileNetV2 and VGG16, to distinguish pneumonia from normal cases on a dataset of 5,232 training and 624 test images, and compare them across accuracy, precision, recall, F1-score, and area under the curve. The MobileNetV2 model achieves the best performance with 88.30 percent accuracy, an F1-score of 0.901, a precision of 0.959, and an area under the curve of 0.957, outperforming VGG16. Grad-CAM then produces visual heatmaps that localise the lung regions most responsible for each prediction, allowing validation of the model's reasoning against regions of radiological opacity. The framework delivers accurate diagnoses alongside interpretable visual explanations that localise pathological regions, supporting radiologist trust and the practical deployment of automated screening in clinical workflows.

Keywords: Pneumonia Detection, Chest X-ray, Explainable AI, Grad-CAM, Transfer Learning, Convolutional Neural Network, Deep Learning, Medical Imaging.

I. INTRODUCTION

Pneumonia is an acute respiratory infection that inflames the air sacs of the lungs and remains one of the leading causes of death worldwide, particularly among children and the elderly [1]. Early, accurate diagnosis is critical because timely treatment substantially improves outcomes, while delayed or missed diagnosis can be fatal [2]. Chest X-ray imaging is the most common and accessible modality for diagnosing pneumonia, as it is inexpensive, fast, and widely available [3]. However, interpreting chest radiographs requires considerable expertise, and the areas of opacity that indicate pneumonia can be subtle and easily confused with other conditions [4]. In many regions the number of trained radiologists is insufficient to meet demand, leading to diagnostic delays and inter-observer variability [5].

These challenges have motivated the development of automated diagnostic systems that can assist clinicians by screening chest X-rays and flagging suspected pneumonia [6]. Deep learning, and in particular the convolutional neural network (CNN), has transformed medical image analysis by learning discriminative features directly from images without manual feature engineering [7]. CNNs have achieved radiologist-level performance on several chest X-ray classification tasks, demonstrating the potential of automated screening

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

to reduce workload and improve consistency [8]. Training a deep CNN from scratch, however, requires very large labelled datasets that are rarely available in medical imaging, where annotation is costly and requires expert knowledge [9].

Transfer learning addresses this limitation by reusing a network pre-trained on a large general-purpose image dataset and fine-tuning it on the target medical task [10]. The pre-trained network has already learned generic visual features such as edges, textures, and shapes, which transfer effectively to medical images and enable strong performance even with modest amounts of labelled data [11]. Transfer learning with architectures such as VGG, ResNet, and the lightweight MobileNet family has become the standard approach for chest X-ray classification, achieving high accuracy with limited training data and computation [12].

Despite their accuracy, deep learning models are widely regarded as black boxes, because their predictions arise from millions of parameters with no inherent explanation [13]. In medicine this opacity is a serious barrier to adoption: a clinician cannot act on a prediction without understanding its basis, both to trust the system and to detect errors, and regulatory frameworks increasingly require that automated decisions be explainable [14], [15]. For image classification, Gradient-weighted Class Activation Mapping (Grad-CAM) has become a standard explanation method, producing a heatmap that highlights the regions of the image most responsible for the prediction [16]. Applied to chest X-rays, Grad-CAM localises the lung regions the model attends to, allowing its reasoning to be compared against the areas of radiological opacity that a radiologist would examine, and thereby turning an opaque score into a visually interpretable diagnosis.

This paper presents an explainable deep learning framework for pneumonia detection from chest X-rays that unites transfer learning with Grad-CAM explanation. The contributions are as follows. First, two pre-trained CNN backbones, MobileNetV2 and VGG16, are fine-tuned for pneumonia classification and rigorously compared across five metrics, with class weighting to address the class imbalance of the dataset. Second, Grad-CAM is applied to the best model to produce visual explanations that localise the lung regions driving each prediction, and these are examined for agreement with the expected radiological findings. Third, the framework is evaluated on a standard chest X-ray benchmark and shown to achieve high accuracy together with clinically meaningful visual explanations. The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the methodology. Section 4 reports the results and discussion. Section 5 concludes.

II. LITERATURE REVIEW

This section reviews deep learning for chest X-ray analysis, transfer learning in medical imaging, and explainability methods for medical deep learning.

A. Deep Learning for Chest X-Ray Analysis

Deep learning has become the dominant approach to automated chest X-ray interpretation. Rajpurkar et al. introduced CheXNet, a 121-layer dense convolutional network trained on the large ChestX-ray14 dataset, and reported pneumonia detection performance exceeding that of practising radiologists, which demonstrated the clinical potential of deep learning for this task [8]. Wang et al. released the ChestX-ray14 benchmark and showed that deep CNNs could localise and classify multiple thoracic diseases from radiographs [17]. Kermany et al. assembled the widely used pediatric chest X-ray pneumonia dataset and demonstrated that transfer learning achieves high accuracy in distinguishing pneumonia from normal cases, establishing the benchmark used in much subsequent work including the present study [18].

Numerous studies have since applied and refined CNN architectures for pneumonia detection. Rajaraman et al. evaluated customised and pre-trained CNNs for pediatric pneumonia and analysed the effect of model depth and data augmentation [19]. Stephen et al. trained a CNN from scratch with extensive augmentation and achieved competitive accuracy, while noting the data requirements of training without pre-training [20]. A consistent finding across this literature is that pre-trained networks fine-tuned on the target data outperform networks trained from scratch when labelled data is limited, which is the normal situation in medical imaging.

B. Transfer Learning in Medical Imaging

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Transfer learning has become the standard paradigm for medical image classification precisely because annotated medical datasets are small relative to the millions of images used to train modern networks [9], [10]. By initialising with weights learned on the ImageNet dataset, a network inherits generic low-level and mid-level visual features that transfer well to radiographs, so only the higher layers and a new classification head need to be adapted to the medical task [11]. This reduces both the data and the computation required and typically improves generalisation.

A range of pre-trained architectures has been applied to chest X-rays. Deep networks such as VGG16 and VGG19 offer strong feature-extraction capacity and have been widely used for pneumonia detection [12], [21]. Residual networks such as ResNet mitigate the vanishing-gradient problem and enable very deep models [22]. Lightweight architectures such as MobileNet, which use depthwise separable convolutions to reduce parameters and computation, are attractive for deployment in resource-constrained clinical settings while retaining competitive accuracy [23]. Comparative studies have found that no single backbone is universally best, and that lightweight models can match or exceed heavier ones on specific tasks, which motivates the direct comparison of MobileNetV2 and VGG16 undertaken in this work.

C. Explainability in Medical Deep Learning

The opacity of deep networks is a particularly acute problem in medicine, where decisions carry high stakes and must be justified [13], [14]. This has driven the application of explainable artificial intelligence to medical imaging, so that predictions can be accompanied by human-interpretable evidence [15]. For image classification, class activation mapping methods are the most widely used, because they produce a spatial heatmap indicating which image regions support the prediction, which aligns naturally with how clinicians reason about images [24].

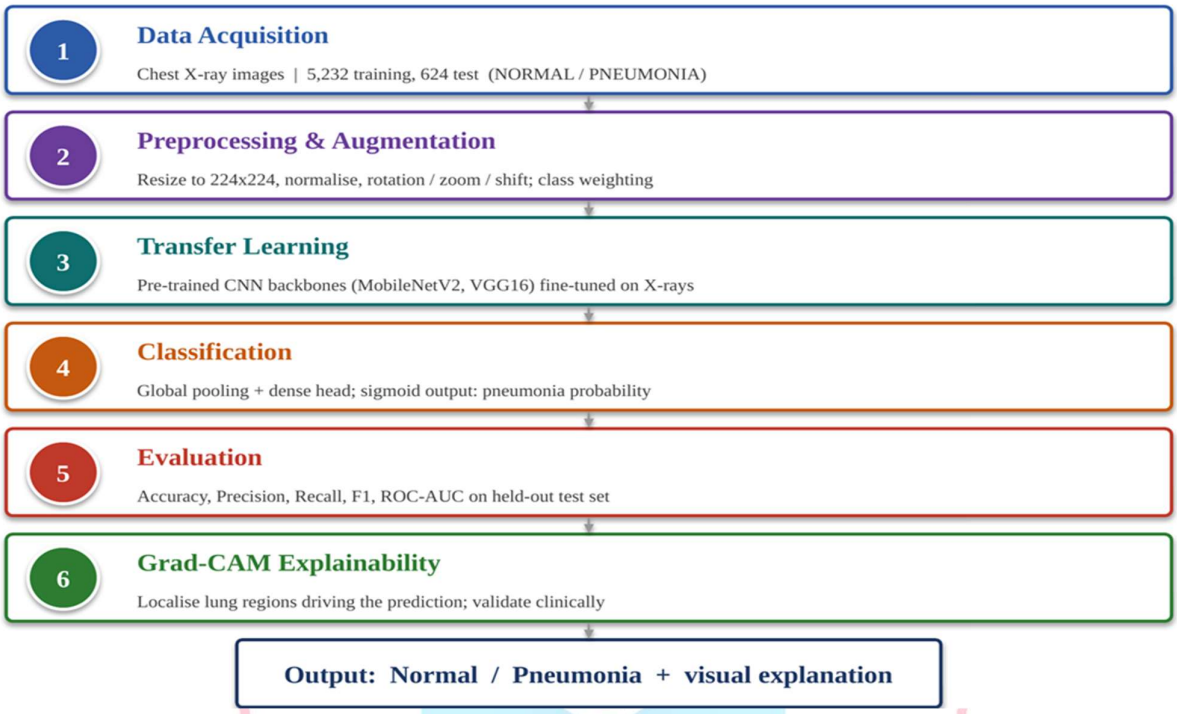
Grad-CAM, introduced by Selvaraju et al. [16], generalised class activation mapping to any CNN architecture by using the gradients of the target output with respect to the final convolutional feature maps to weight those maps into a class-discriminative localisation heatmap. Grad-CAM has been extensively applied to chest X-ray models to verify that they attend to clinically relevant regions rather than spurious artefacts, and to build clinician trust [25]. Studies have shown that a model can achieve high accuracy while attending to irrelevant regions such as image markers or borders, which Grad-CAM readily exposes, underscoring the importance of explanation alongside accuracy. The present work adopts Grad-CAM to explain the fine-tuned pneumonia detector and to confirm that it localises the lung regions associated with the disease, uniting accurate transfer-learning classification with visual interpretability in a single framework.

III. PROPOSED METHODOLOGY

The proposed framework classifies chest X-rays as normal or pneumonia using transfer learning and explains each prediction with Grad-CAM. Fig. 1 presents the overall framework, Fig. 2 the transfer-learning architecture with the Grad-CAM branch, and Fig. 3 the Grad-CAM computation pipeline.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

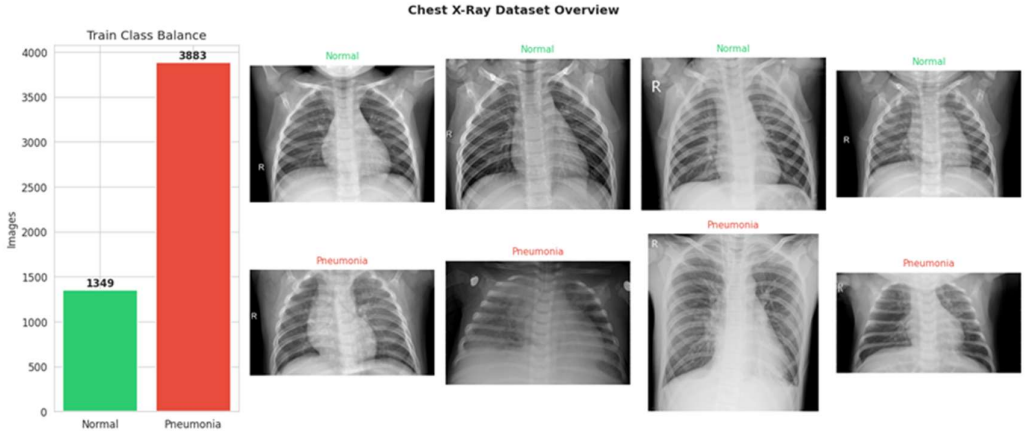
Fig. 1: PROPOSED EXPLAINABLE DEEP LEARNING FRAMEWORK FOR PNEUMONIA DETECTION, FROM DATA ACQUISITION AND AUGMENTATION THROUGH TRANSFER-LEARNING CLASSIFICATION AND EVALUATION TO GRAD-CAM EXPLANATION OF THE DIAGNOSIS.



A. Dataset and Preprocessing

The study uses the chest X-ray pneumonia dataset [18], which contains frontal chest radiographs labelled as normal or pneumonia. The images are organised into training, validation, and test folders; because the official validation folder contains only a handful of images, the training and validation images are merged and re-split, giving a training pool of 5,232 images and a held-out test set of 624 images, of which 234 are normal and 390 are pneumonia. The dataset is imbalanced, with pneumonia cases roughly three times as frequent as normal cases in the training pool, which is addressed through class weighting during training. All images are resized to 224 by 224 pixels and normalised, and the training images are augmented with small random rotations, zooms, and shifts to improve generalisation. Fig. 2 shows the class distribution and representative images.

Fig. 2: CHEST X-RAY DATASET OVERVIEW. LEFT: CLASS BALANCE OF THE TRAINING POOL. RIGHT: REPRESENTATIVE NORMAL AND PNEUMONIA RADIOGRAPHS.



JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

B. Transfer Learning Models

Two pre-trained CNN backbones are used as feature extractors, as shown in Fig. 3. MobileNetV2 [23] is a lightweight architecture built from depthwise separable convolutions and inverted residual blocks, and VGG16 [21] is a deeper architecture of stacked 3 by 3 convolutions. Both are initialised with weights pre-trained on ImageNet and their convolutional base is frozen, so that the learned generic features are reused. A classification head is appended, consisting of a global average pooling layer, a dense layer of 128 units with rectified linear activation, dropout for regularisation, and a single sigmoid output that gives the pneumonia probability. For an input image x, the model computes

$$p(pneumonia \mid x) = \sigma(w^T GAP(f_base(x)) + b) \quad (1)$$

where f_base is the frozen convolutional backbone, GAP is global average pooling, and σ is the sigmoid function. The network is trained to minimise the class-weighted binary cross-entropy loss

$$L = -(1/N) \Sigma [\omega_1 y_i \log p_i + \omega_0 (1-y_i) \log(1-p_i)] \quad (2)$$

where y_i is the true label, p_i the predicted probability, and ω_0, ω_1 are class weights that compensate for the class imbalance. Table 1 lists the configuration.

Fig. 3: TRANSFER-LEARNING ARCHITECTURE WITH GRAD-CAM EXPLANATION. A FROZEN PRE-TRAINED CNN BACKBONE EXTRACTS FEATURES THAT A LIGHTWEIGHT HEAD CLASSIFIES, AND GRAD-CAM USES THE BACKBONE FEATURE MAPS AND THE OUTPUT GRADIENT TO EXPLAIN THE PREDICTION.

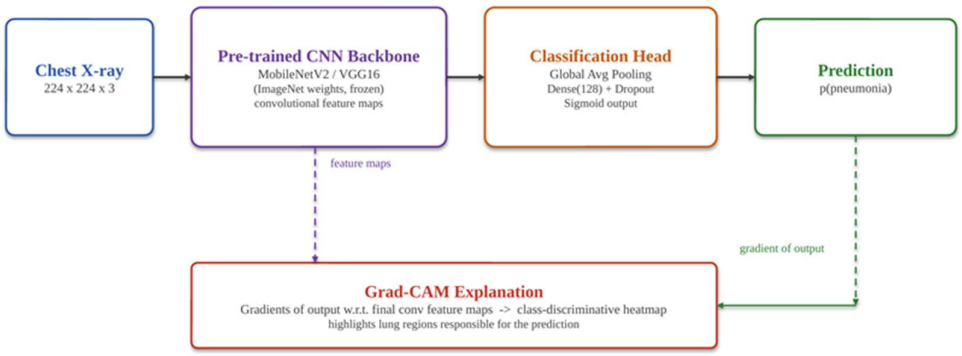


Table 1: TRANSFER LEARNING AND EXPERIMENTAL CONFIGURATION.

Parameter	Value
Backbones	MobileNetV2, VGG16 (ImageNet)
Input size	224 x 224 x 3
Classification head	GAP + Dense(128) + Dropout
Output	Sigmoid (pneumonia probability)
Optimiser	Adam (learning rate 1 x 10^{-4})
Loss	Class-weighted binary cross-entropy
Epochs / batch size	10 (early stopping) / 32
Augmentation	Rotation, zoom, shift
Train pool / test	**5,232 / 624 images**

C. Grad-CAM Explainability

To explain the predictions, Grad-CAM [16] is applied to the best model. Grad-CAM produces a class-discriminative localisation heatmap by weighting the final convolutional feature maps with the gradient of the output. Let A_k denote the k-th feature map of the final convolutional layer and y the model output. The

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

importance weight of each feature map is the global average of the gradient of the output with respect to that map

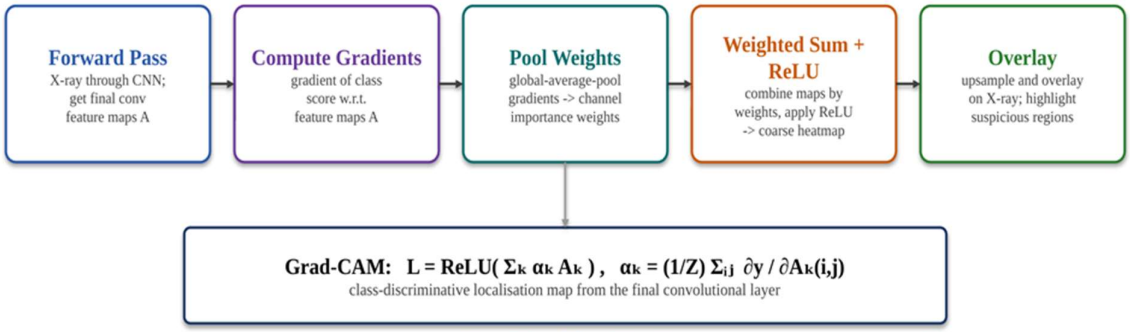
$$\alpha_k = (1/Z)\, \Sigma_i \Sigma_j \; \partial y / \partial A_k(i,j) \qquad (3)$$

where Z is the number of spatial locations. The heatmap is the rectified weighted sum of the feature maps

$$L = \mathrm{ReLU}(\Sigma_k \alpha_k A_k) \qquad (4)$$

The rectification retains only the features with a positive influence on the class of interest. The heatmap is upsampled to the input resolution and overlaid on the X-ray, highlighting the regions most responsible for the prediction. Fig. 4 summarises the computation, and Algorithm 1 gives the full procedure

Fig. 4: GRAD-CAM HEATMAP GENERATION PIPELINE, FROM THE FORWARD PASS AND GRADIENT COMPUTATION THROUGH WEIGHT POOLING AND WEIGHTED SUMMATION TO THE OVERLAY ON THE CHEST X-RAY.



Algorithm 1: Explainable Pneumonia Detection with Grad-CAM

Input: chest X-ray images with labels, pre-trained backbones
Output: pneumonia predictions and Grad-CAM heatmaps
1: merge train + val images; re-split into train / val
2: resize to 224x224, normalise, augment training images
3: compute class weights from label frequencies
4: for backbone in {MobileNetV2, VGG16} do
5: freeze ImageNet base; add GAP + Dense + sigmoid head
6: fine-tune on training set (class-weighted BCE, Adam)
7: evaluate on test set (Acc, Prec, Rec, F1, AUC)
8: best <- model with highest F1
9: A <- final conv feature maps of best; y <- output
10: alpha_k <- global-average-pool(dy / dA_k)
11: L <- ReLU(sum_k alpha_k A_k) // heatmap
12: upsample L and overlay on X-ray
13: return predictions, heatmaps

D. Evaluation Metrics

Model performance is measured by accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). The F1-score, the harmonic mean of precision and recall, is particularly informative under class imbalance

$$F1 = 2 \bullet (\mathrm{Precision} \bullet \mathrm{Recall}) / (\mathrm{Precision} + \mathrm{Recall}) \qquad (5)$$

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Recall is clinically important because a missed pneumonia case is more dangerous than a false alarm, while precision reflects the reliability of a positive prediction. The AUC assesses discrimination independently of the decision threshold.

IV. RESULTS AND DISCUSSION

The framework is evaluated on the held-out test set of 624 chest X-rays, comprising 234 normal and 390 pneumonia cases. Both backbones were fine-tuned with class weighting and early stopping. Results cover classification performance, training behaviour, and Grad-CAM explainability.

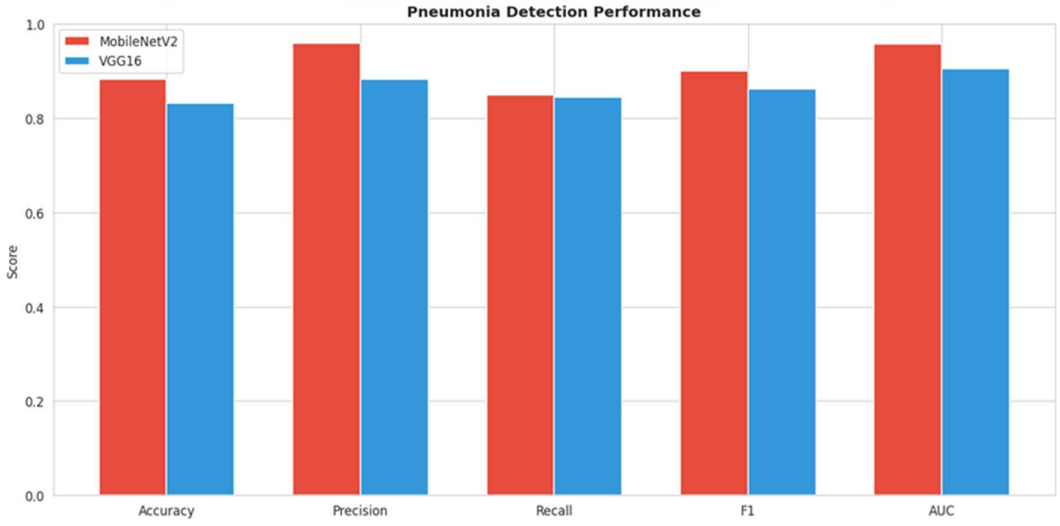
A. Classification Performance

Table 2 reports the performance of the two models. MobileNetV2 clearly outperforms VGG16, achieving 88.30 percent accuracy, a precision of 0.959, a recall of 0.849, an F1-score of 0.901, and an area under the curve of 0.957, compared with 83.17 percent accuracy and an F1-score of 0.862 for VGG16. The high precision of MobileNetV2 indicates that when it predicts pneumonia it is correct in almost 96 percent of cases, which is valuable for limiting unnecessary follow-up, while its recall of 0.849 means it identifies the large majority of true pneumonia cases. That the lightweight MobileNetV2 outperforms the much larger VGG16 is consistent with prior findings that compact architectures can match or exceed heavier ones on specific medical imaging tasks, and it is advantageous for deployment because MobileNetV2 is far more efficient. Figs. 5 and 6 present the metric comparison and the confusion matrices with ROC curves.

Table 2: PNEUMONIA DETECTION PERFORMANCE ON THE TEST SET.

Model	Acc (%)	Prec	Rec	F1	AUC
MobileNetV2	**88.30**	**0.959**	**0.849**	**0.901**	**0.957**
VGG16	83.17	0.882	0.844	0.862	0.905

Fig. 5: PERFORMANCE COMPARISON OF MOBILENETV2 AND VGG16 ACROSS ACCURACY, PRECISION, RECALL, F1-SCORE, AND AUC ON THE CHEST X-RAY TEST SET.



JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Fig. 6: Confusion matrices for the two models (left and centre) and ROC curves (right). MobileNetV2 attains the highest AUC of 0.957.

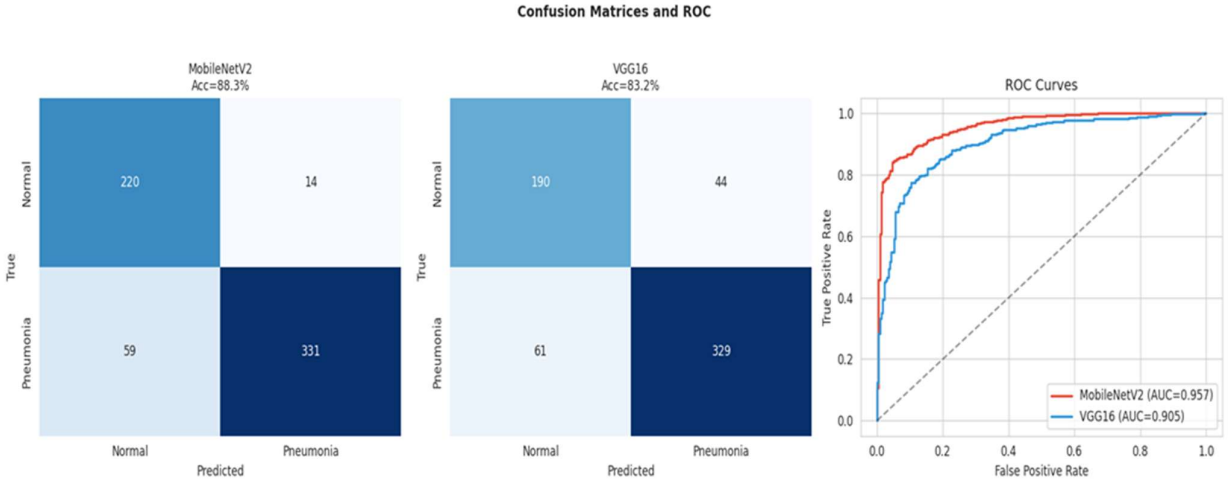


Table 3 reports the per-class results of the best model. The model attains a precision of 0.96 and a recall of 0.85 on the pneumonia class and a precision of 0.79 and a recall of 0.94 on the normal class. The high recall on the normal class indicates that few normal cases are misclassified as pneumonia, and the high precision on the pneumonia class confirms the reliability of positive predictions.

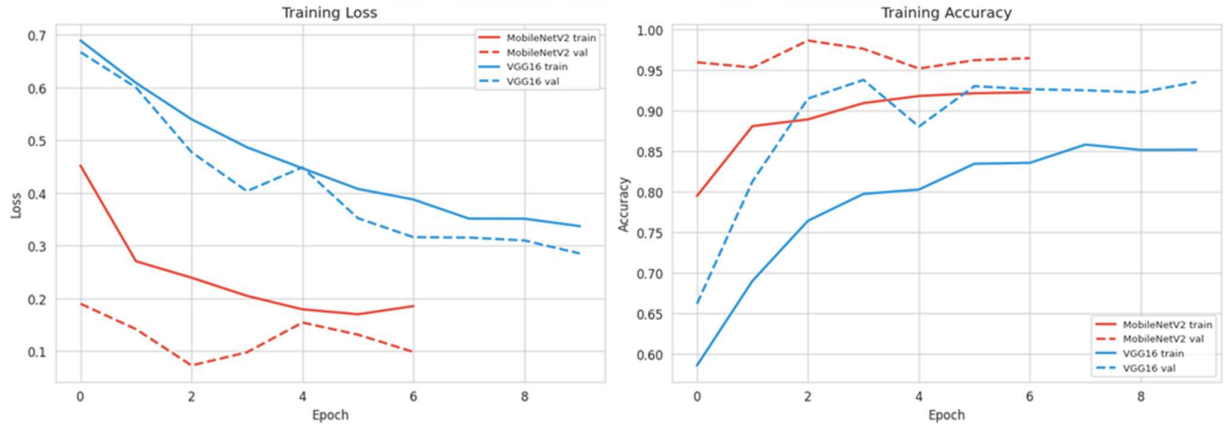
Table 3: PER-CLASS PERFORMANCE OF THE BEST MODEL (MOBILENETV2).

Class	Precision	Recall	F1-score
Normal	0.79	0.94	0.86
Pneumonia	0.96	0.85	0.90
Weighted avg	0.90	0.88	0.88

B. Training Behaviour

Fig. 7 shows the training and validation curves for both models. The losses decrease steadily and the validation accuracy rises and stabilises, indicating that the models learn effectively from the fine-tuning data without severe overfitting, which the augmentation and dropout help to prevent. The validation accuracy of the best model exceeds 93 percent during training, consistent with its strong test performance.

Fig. 7: TRAINING AND VALIDATION LOSS (LEFT) AND ACCURACY (RIGHT) FOR MOBILENETV2 AND VGG16 OVER THE FINE-TUNING EPOCHS.



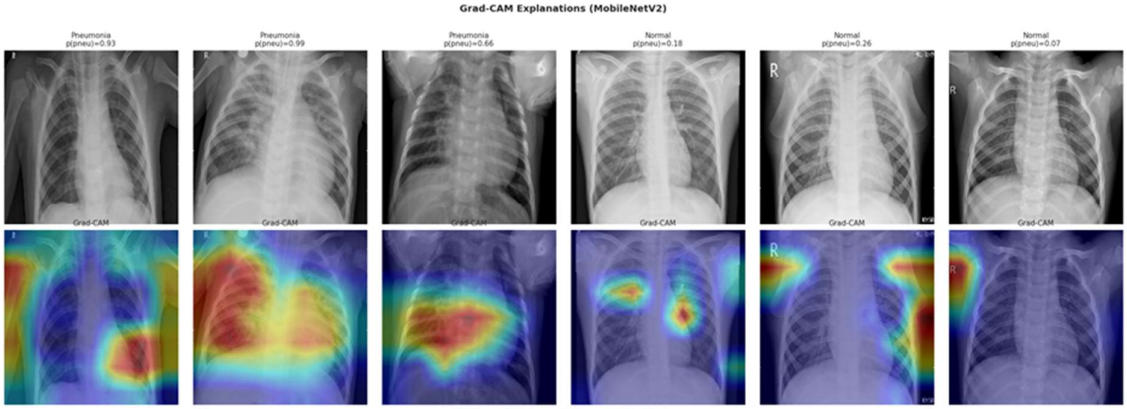
C. Grad-CAM Explainability

Grad-CAM was applied to the best model to explain its predictions. Fig. 8 shows the original radiographs in the top row and the corresponding Grad-CAM overlays in the bottom row for representative

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

pneumonia and normal cases. For the pneumonia cases, the heatmaps concentrate on the lung fields, highlighting the regions of opacity and consolidation that characterise the infection, which is where a radiologist would look. For the normal cases, the activation is more diffuse and of lower intensity, consistent with the absence of localised pathology. This agreement between the model attention and the expected radiological findings provides evidence that the model bases its decisions on clinically relevant image regions rather than on spurious artefacts, which is essential for building clinician trust. The per-prediction heatmaps also give a direct, visual justification that can accompany each automated diagnosis in a screening workflow.

Fig. 8: GRAD-CAM EXPLANATIONS FOR REPRESENTATIVE TEST IMAGES. TOP ROW: ORIGINAL CHEST X-rays with PREDICTED PNEUMONIA PROBABILITY. BOTTOM ROW: GRAD-CAM OVERLAYS LOCALISING THE REGIONS THAT DRIVE THE PREDICTION, CONCENTRATING ON THE LUNG FIELDS FOR PNEUMONIA CASES.



D. Discussion

The results show that transfer learning with a lightweight backbone provides accurate and, through Grad-CAM, interpretable pneumonia detection. MobileNetV2 achieves high accuracy and an area under the curve of 0.957 while being efficient enough for deployment in resource-constrained settings, and its Grad-CAM explanations localise clinically meaningful regions, addressing the twin requirements of accuracy and interpretability in medical AI. A limitation of this study is that the dataset consists of pediatric chest X-rays from a single source, so the generalisation of the models to adult populations and to images from different scanners and institutions remains to be established. A further consideration is that Grad-CAM provides a coarse localisation at the resolution of the final convolutional layer, so finer-grained explanation methods could offer more precise localisation. Finally, the recall on the pneumonia class, while high, leaves room for improvement, and in a clinical screening setting the decision threshold could be adjusted to increase recall at the cost of some precision, since missed cases are more dangerous than false alarms. These considerations motivate the directions outlined in the conclusion.

V. CONCLUSION AND FUTURE WORK

This paper presented an explainable deep learning framework for pneumonia detection from chest X-rays that combines transfer learning with Grad-CAM explanation. Two pre-trained backbones, MobileNetV2 and VGG16, were fine-tuned to classify radiographs as normal or pneumonia, with class weighting to address the dataset imbalance. The lightweight MobileNetV2 achieved the best performance, with 88.30 percent accuracy, a precision of 0.959, an F1-score of 0.901, and an area under the curve of 0.957, outperforming the much larger VGG16. Grad-CAM applied to the best model produced visual heatmaps that localised the lung regions responsible for each prediction, concentrating on the areas of opacity for pneumonia cases in agreement with expected radiological findings. The framework thus delivers accurate diagnosis together with the visual interpretability needed for clinician trust and practical deployment.

Several directions remain for future work. The most immediate is to validate the framework on larger and more diverse datasets spanning adult populations and multiple institutions and scanners, to confirm that the results generalise beyond the single-source pediatric data used here. A second direction is to fine-tune the deeper layers of the backbone, rather than freezing the entire base, which may further improve accuracy at the

cost of additional computation. A third direction is to explore finer-grained explanation methods and to quantitatively validate the heatmaps against radiologist annotations of the pathological regions. Finally, integrating the detector and its Grad-CAM explanations into a clinical decision support interface, presenting the predicted probability alongside the visual explanation, would support the practical adoption of explainable deep learning in radiology workflows.

REFERENCES

[1] World Health Organization, "Pneumonia in children," WHO Fact Sheet, 2022.
[2] T. Cherian et al., "Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies," Bulletin of the World Health Organization, vol. 83, no. 5, pp. 353–359, 2005.
[3] World Health Organization, Standardization of Interpretation of Chest Radiographs for the Diagnosis of Pneumonia in Children. Geneva: WHO, 2001.
[4] C. J. Williams et al., "Variability and accuracy in the interpretation of chest radiographs for paediatric pneumonia," Pediatric Radiology, vol. 43, no. 8, pp. 1017–1025, 2013.
[5] M. I. Neuman et al., "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children," Journal of Hospital Medicine, vol. 7, no. 4, pp. 294–298, 2012.
[6] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[8] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.
[9] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 3347–3357.
[10] H. C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285–1298, 2016.
[11] N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.
[12] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," Journal of Big Data, vol. 6, no. 1, p. 113, 2019.
[13] Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
[14] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," WIREs Data Mining and Knowledge Discovery, vol. 9, no. 4, p. e1312, 2019.
[15] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.
[16] 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 International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE CVPR, 2017, pp. 2097–2106.
[18] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122–1131, 2018.
[19] S. Rajaraman et al., "Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs," Applied Sciences, vol. 8, no. 10, p. 1715, 2018.
[20] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," Journal of Healthcare Engineering, vol. 2019, p. 4180949, 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International Conference on Learning Representations (ICLR), 2015.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.
[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE CVPR, 2018, pp. 4510–4520.
[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE CVPR, 2016, pp. 2921–2929.
[25] I. Ahmad et al., "Deep learning for chest X-ray diagnosis with explainable AI: A review," IEEE Access, vol. 10, pp. 78895–78917, 2022.