

AN EXPLAINABLE DEEP LEARNING APPROACH FOR PNEUMONIA DETECTION USING CHEST X-RAY IMAGES

¹Adyan Ahmad, ^{1,2}Didar Hussain, ^{*}²Hashim Ali, ¹Abdullah Sadiq, ¹Junaid Shah, ²Shahzad Alam

¹Department of Computer Science, Bacha Khan University Charsadda, KPK, Pakistan

²Department of Computer Science, Abdul Wali Khan University Mardan, KPK, Pakistan

DOI: <https://doi.org/10.5281/zenodo.22206870>

Keywords:

Pneumonia Detection, Chest X-Ray, Deep Learning, Explainable AI, Grad-CAM, Adversarial Training, Spatial Attention Mechanism, Mean Relevance Score, ResNet-50, Shortcut Learning, Layer-wise Relevance Propagation

Article History

Received: 01 Jan, 2026

Accepted: 22 Jan, 2026

Published: 25 Jan, 2026

Copyright @Author

Corresponding Author

*Hashim Ali

hashimali@awkum.edu.pk

Abstract

Pneumonia is still one of the leading causes of death in children under five, and the burden is worst in low-income regions where trained radiologists are scarce. Chest X-ray imaging is usually the first diagnostic step, and deep learning classifiers have gotten very good at reading these images, with benchmark accuracies now above 95%. The problem is that a lot of these models are not actually looking at the disease, they often key in on scanner artifacts or other spurious patterns instead of real pathology, a failure known as shortcut learning, which makes them risky to trust clinically no matter how high the reported accuracy is. This study compares four explainable AI methods side by side under identical conditions, Layer-wise Relevance Propagation (LRP), Grad-CAM, FGSM-based Adversarial Training, and a Spatial Attention Mechanism (SAM), all built on the same ResNet-50 backbone and tested on a public pediatric chest X-ray dataset of 5,856 labeled images. A stricter preprocessing pipeline, a sensitivity-targeted decision threshold, and a new composite Trade-off Score capturing accuracy and explanation quality together were also introduced. The results point to a genuine trade-off. LRP came out on top for interpretability, reaching a Mean Relevance Score of 0.9078 while holding accuracy at 93.91%, the same as the baseline. Grad-CAM gave the best classification numbers, 95.03% accuracy and 98.72% AUC-ROC, with a still-decent interpretability score of 0.7963. All four variants beat the previously reported baseline, including a 12.30-point jump in specificity for Grad-CAM, a 10.45-point gain in AUC-ROC for Adversarial Training, and a 0.5481 improvement in MRS for SAM. This shows accuracy and interpretability need not be a trade-off accepted blindly, giving clinician's real evidence to base that choice on.

1. Introduction

In the world whole, around 14% of all deaths in children under-5 are caused by pneumonia, and the problem is particularly severe in low income countries where there is a lack of expertise in diagnostic specialist services. Chest radiography is the first line screening test for pulmonary compromise as it is widely available and cost-effective, and provides visual information on consolidation patterns, infiltrates and fluid accumulation. In a well-equipped facility a trained radiologist can make a diagnosis in a few minutes, but this is not true in all facilities. In under resourced environment, one radiologist might be expected to read hundreds of radiographs in a single day, and wait hours when the diagnosis could have been clinically valuable while the patient was waiting for antibiotics and/or antiviral medication [1]. Such is also true for adults with impaired immune systems or underlying lung disease, as well as the elderly (who are particularly prone to death due to late diagnosis) [2]. This mismatch of the number of radiographers and radiological equipment caused interest in automated analysis of the CXR images. The initial machine learning methods used hand-crafted feature descriptors but pneumonia is highly variable between age groups, pathogen types and imaging devices which made it impossible to have a single descriptor that could reliably cover the variations. The emergence of convolutional neural networks (CNNs) marked a paradigm shift away from handcrafted feature engineering toward end-to-end representation learning directly from raw pixel input [3]. In the Kaggle Chest X-ray (CXR) Pneumonia dataset (5,856 labeled CXR from children in the Guangzhou Women and Children's Medical Centre), modern CNN designs often achieve an accuracy of 95% or higher and some recent architectures achieve an accuracy of 99% or more [4]. Classifying the problem it seems to be a solved problem but it is not. The levels of accuracy and clinical trustworthiness are not the same and that is the beginning of this study.

2. Literature Review

The interpretability backbone of this study consists of gradient-based saliency methods, such as Grad-CAM and LRP, which are also employed here due to their role in improving model robustness and saliency coherence, respectively, and are used, in adversarial training. A critical synthesis that justifies the controlled four-way comparison, which bridges the gap of literature standardization, is provided by the end of the chapter.

Manju et al. [2] introduced a new model namely VGG19 for multi-class classification of normal/bacterial/viral cases using transfer learning. The preprocessing techniques applied include wavelet denoising and elastic deformation while the interpretation technique used is Grad-CAM. It showed the trade-off between interpretability and accuracy, favoring utility and interpretability with the use of complex attention techniques. Ennab et al. [3] fused VGG19 and Grad-CAM for pneumonia CXR classification, achieving an accuracy of more than 90% and superb sensitivity and specificity. The most important result was the clinical actionability of the region-level explanation, which makes Grad-CAM a valuable baseline for interpretability than just a popular hack. To train an improved DenseNet201 on transfer learning, Hadavale et al. [5] performed an interpretive audit of a DenseNet model for the diagnosis of pneumonia in children on the 5,856-image dataset. Prior to the audit, pre-audit AUROC was 0.9682, and Grad-CAM++ and RISE demonstrated the use of non-pathological features. On NIH ChestX-ray14, specificity after retraining increased to 0.8675 and accuracy to 0.9247, but external validation AUROC fell to 0.6509. Colin et al. [6] set the first benchmark for pneumonia classification with controlled multi-method XAI via various interpretability techniques. Two of them had a higher spatial attribution quality (LPR with 0.85 MRS and 91% classification accuracy, where it attributed to the lung tissue most relevant to the diagnosis, compared to the other three methods). Priyatharsini et al. [7] combined CNNs, LSTMs, and attention layers for lung disease identification using CXRs and patient metadata. Tested on NIH ChestX-ray14, the model achieved 0.92 accuracy, 0.94 recall and an AUC of 0.95, outperforming simpler baselines, with Grad-CAM providing visual decision maps.

Patil et al. [8] introduced a CNN application to detect the lung infection in children from X-ray images for resource-limited environments. The model was successful in the detection of early symptom; it yielded the accuracy of 82% and sensitivity of 87%, which was suitable for early symptom detection in areas with limited

access to healthcare facilities. The first application of the interpretable Deformable Prototypical Part Network (D-ProtoPNet) to the diagnosis of pediatric pneumonia was provided by Yazdani et al. [9] with a Kaggle dataset of 5,856 CXR images. When evaluated with ResNet-50, D-ProtoPNet achieved 86% accuracy and 93% AUC, which is 3% higher than that of ProtoPNet, and offers prototype based explanations to assist the clinical decision making process. Mahfuz et al. [10] suggested a hybrid method by anomaly detection and multi-class classification of CXR into three classes: Normal, Bacterial and Viral. System performance was 82% overall accuracy and it employed heat maps to mark out diagnostically important areas, with uncertainty flags on for unknown inputs. Reddy et al. [11] constructed a binary CXR classifier based on the MobileNetV2 pre-trained on ImageNet along with the visual interpretation of the classification result using the Grad-CAM. The system was designed to be transparent and provide real-time diagnostic support, but it lacks a comprehensive set of performance metrics to make it comparable to other models. Prodhan et al. [12] used VICReg to self-supervised pre-train on unlabeled CXRs and then fine-tuned on a small labeled population to achieve around 91% accuracy on the test set, and were comparable to the supervised baselines on these datasets without using ImageNet weights. Grad-CAM visualizations verified that attention of the model was at clinically predictable regions of the lungs.

Otsuki et al. [13] solved a problem of the conventional LRP when it is applied to ResNet architectures that break the conservation constraint imposed on LRP. It was proposed that conservation be maintained during propagation with a Relevance Splitting mechanism at residual junction points. The improved LRP also showed improvements on two challenging datasets, namely the ImageNet and Caltech-UCSD Birds-200-2011. Neenu et al. [14] employed a ResNet-50 pipeline with the addition of Multiscale Curvelet Filtering with Directional Denoising to tackle the noise, late fusion of deep features with handcrafted LBP features to diversify the features, and a spatially concentrated attention module to interpretability. In a systematic review, studied by Akhila et al. [15] proposed a basic CNN, ensemble of ResNet-50, DenseNet121, VGG16, InceptionV3 and EfficientNet on the Kaggle dataset and NIH ChestX-ray14. The system showed 95% accuracy and an AUC-ROC score of 0.97, and with the help of explainability techniques, the clinical confidence in the system in a hospital ready design is enhanced. Ali et al. [16] compared six CNN architectures on 5856 CXRs in an experiment. The best performance of EfficientNetV2L was the highest accuracy, reaching 94.02%, showing it performed well and consistently without external validation for pneumonia detection. Chakraborty et al. [17] introduced a custom CNN with lightweight components, such as Conv2D with a small kernel, MaxPooling, trained on 5,856 CXRs. With less than 9.45 million parameters, the model reached 90.12% accuracy, and even surpassed heavier pre-trained systems on different metrics without sacrificing speed.

Dalvi et al. [18] implemented VGGNet16, ResNet-50, and InceptionV3 and trained them on a balanced dataset of CXR with augmentation, explaining the trained models using Grad-CAM. ResNet-50 performed best overall, while Grad-CAM was able to localize pneumonia-related regions in the lung for the physician interpretation. Hasnat et al. [19] used 26824 CXRs from the RSNA collection, and modified the final 100 layers of a ResNet-50 backbone with SE modules and then performed few-shot learning. The model performed well with an accuracy of 84% and AUC of 0.87, and LIME and Grad-CAM++ maps combined marked the regions in the lung that were important for the diagnosis. Li et al. [20] segmented pediatric CXRs into six regions of the lungs and the mediastinum using RA-UNet and classified them with CBAM-ResNet into 4,400 standard WHO-classified CXRs. The region-based input increased the accuracy by 0.82 and gained 3% AUC over the unsplit data and Grad-CAM gave better visual interpretability. On 5,863 CXRs, Erukude et al. [21] reported DenseNet121 with a better accuracy of 89.1% compared to ResNet-50 with an accuracy of 84.4%. Grad-CAM heatmaps showed that DenseNet121 was interested in areas of the image that were important for the disease, whereas ResNet-50 was interested in peripheral image areas that were not informative. Jihad et al. [22] tested Xception, EfficientNet B0, ResNet-50 and a Vision Transformer (ViT) on CXRs, interpreting the models with Grad-CAM, LIME, Integrated Gradients and Saliency Maps. ResNet-50 and EfficientNet B0 performed better with lower overfitting and ViT performed better with greater amount of data. Mahamud et al. [23] used a denoising autoencoder (DAE), CLAHE and Gamma Correction for pre-processing all CXRs that included COVID-19, pneumonia and tuberculosis. The model had an 85.30% accuracy, 99% precision and recall, and SHAP, LIME, Grad-CAM and Grad-CAM++ were applied to understand the reasoning patterns of the model.

Ashraf et al. [24] showed that the CNN-based pneumonia classifiers such as VGG16, VGG19 and a combination of CNN MobileNetV2 and DenseNet169 are susceptible to both FGSM and PGD adversarial attacks. A convolutional autoencoder defense was proposed, where the models achieve up to 16% accuracy improvement over the existing defenses when attacked by PGD. Mozhegova et al. [25] investigated the adversarial robustness of multimodal large models in healthcare applications. Experiments across image and text modalities found that multimodal models exhibit substantially greater robustness to adversarial attacks than single-modality models, supporting the integration of multiple modalities for more robust medical image analysis systems. Rafferty et al. [26] identified that common XAI methods applied to InceptionV3 for lung cancer detection produced explanations misaligned with radiologist assessments. ClinicXAI, an expert-driven concept bottleneck model developed in collaboration with a radiologist, was introduced to provide clinically meaningful explanations without sacrificing performance, and was shown to be orders of magnitude more robust to adversarial attacks than the baseline. Wong et al. [27] introduced the Synthetic-Augmented Adversarial Training (SADA) framework, combining FGSM-based adversarial training with Stable Diffusion-generated synthetic images using ConvNeXt architectures. SADA achieved an F1-score of 0.8437 under cross-model attacks on unseen data, representing a 17.40% improvement over standard adversarial training.

From the existing literature we conclude that the proposed models performed well for the detection of pneumonia however, they suffered from the following issues:

- i. They provide less indication of whether the prediction was caused by a lower-lobe consolidation or if it was caused by a scanner brand burned into the corner of the image.
- ii. Following the interpretability audit that identified and eliminated the shortcut patterns, specificity reached 86.75% and accuracy 92.47% [5]. The presence of high benchmark accuracy cannot be assumed to indicate clinical reliability in the absence of interpretability evidence.

This study aims to apply DL-based model with suitable interpretability techniques for pneumonia diagnosis using CXR images. The task will be accomplished with the help of the following objectives:

- i. To propose a DL-based model with different interpretability techniques.
- ii. Remove border artifacts using clever body cropping and symmetric border removal.
- iii. Assess the quality of spatial attribution for each method.
- iv. At the end compare the proposed model with the state of the art models.

3. Material and Methods

This study extends the ResNet-50 pipeline by improving four interpretability techniques, the preprocessing pipeline, and a sensitivity-targeted threshold selection method to facilitate interpretation and improve the diagnostic accuracy and transparency of the model. The design includes a robust framework of mitigations for shortcut learning, inverse-frequency class weighting, and a multi-metric evaluation structure that facilitates clinical trust through anatomically grounded model explanations.

The study rests on two crucial aspects: CNN-based DL architecture and the selection of interpretability methods, since a large number of interpretability methods exists that offer different information about model behavior. This study compares the same four interpretability techniques used by Colin et al. [6] Grad-CAM, Adversarial Training, SAM, and LRP but each is slightly modified through improved implementation decisions, using a shared underlying model so that any differences in saliency quality are attributable to the explanation method itself rather than the model, as recommended by Vyas et al. [28] for clinically valid comparisons of XAI. The experimental design used in this study is illustrated in Fig. 1.

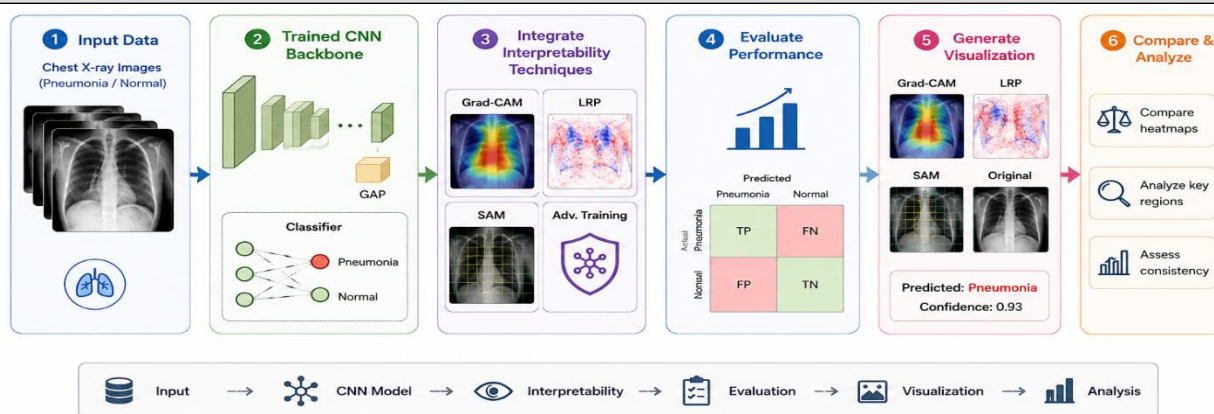


Fig. 1. Experimental Design Workflow

The workflow proceeds from the labeled chest X-ray images through the shared, trained ResNet-50 backbone, after which each of the four interpretability techniques is integrated in turn; the resulting classification and explanation outputs are then evaluated and visualised before being compared and analysed across all five configurations, forming the controlled comparison at the centre of this study.

3.1 Dataset Description

In this study, the CXR Images (Pneumonia, Normal) dataset which is publicly available on Kaggle (<https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>) was used. The dataset comprises 5,856 annotated pediatric chest radiograph images, in which 4,273 samples represent confirmed pneumonia cases and 1,583 are classified as normal, reflecting a real-world class imbalance of approximately 2.7:1 that also mirrors real-world clinical prevalence. Representative examples drawn from the three original label categories in the source collection, normal, viral pneumonia and bacterial pneumonia, are shown in Fig. 2, illustrating the subtle and often overlapping radiographic appearance that makes this classification task genuinely difficult even for a trained eye.

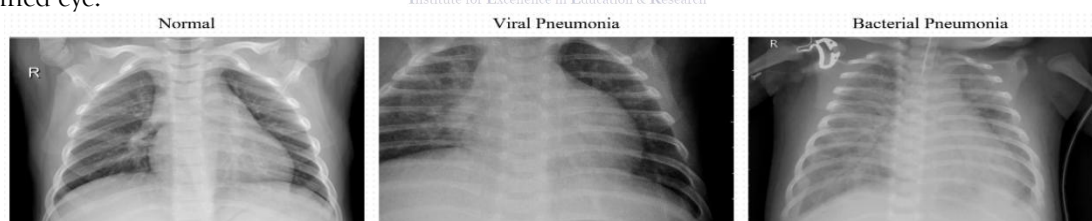


Fig. 2. Illustrative examples of CXRs from the Kaggle dataset

The normal radiograph shows clear lung fields with well-defined costophrenic angles, while the viral and bacterial pneumonia cases both show areas of increased opacity, though with different distribution and density. The original validation partition in this dataset was not large enough for consistent performance estimates (16 images total), so the training and validation folders were merged into a pool of 5,232 images (3,883 pneumonia and 1,349 normal) and then re-split at a 90:10, ratio using stratified sampling, ensuring both partitions preserved the same class distribution. This gave 4,709 training images (3,447 pneumonia, 1,262 normal) and 523 validation images (436, pneumonia and 87 normal). A separate test set (n = 624) was kept completely hidden from any preprocessing decision and threshold tuning, and was used only once for final evaluation to prevent data leakage.

3.2 Data Preprocessing

All images were resized to 224×224 with bilinear interpolation without distortion of the diagnostically relevant lung texture, and then normalized to (0, 1) by dividing the pixel intensities by 255. A smart body-crop applied to

each CXR removed 3% from around the patient silhouette, based on the mean channel value of the foreground pixels $> 15.0/255$, and resampled the image to the target resolution, and then a symmetric border-artifact removal (cropping 2% from each side to eliminate the scanner-imprinted text and orientation markers) also applied before passing to any model. The two routines are used in the ImageDataGenerator preprocessing function, guaranteeing that the preprocessing is always done before normalization. For augmentation Random rotations $\pm 10^\circ$, random horizontal flipping, zoom variation $\pm 15\%$, translation $\pm 5\%$ in both axis, and brightness adjustment (factor range: 0.85–1.15) applied which were clinically plausible geometric and photometric transformations that were applied only to the training set. These transforms are chosen so as to cover all the reasonable possibilities of acquisition without inducing anatomically unreasonable distortions.

3.3 Class Imbalance Handling

Instead of artificially balancing the dataset through oversampling or undersampling, this study addresses class imbalance by calculating inverse-frequency class weights, which penalize misclassification of the minority class proportionally during training without altering the underlying data distribution or introducing synthetic samples. Class weights are calculated using scikit-learn's balanced formula:

$$w_c = \frac{N}{K \cdot n_c}, \quad N = 4,709, \quad K = 2, \quad n_{\text{NORMAL}} = 1,262, \quad n_{\text{PNEUMONIA}} = 3,447 \quad (1)$$

These weights are passed to model.fit() through the class_weight_dict argument, scaling the binary cross-entropy contribution to the gradient by roughly $2.73\times$ for normal-class samples relative to pneumonia-class samples.

3.4 Proposed CNN Model

The proposed model is ResNet-50 a 50-layer residual CNN pre-trained on ImageNet and was selected due to its depth, skip-connection and ability to perform well on the CXR dataset as shown in the literature [6, 14, 17]. The architecture consists of four stacked bottleneck blocks, each performing a $1\times 1 \rightarrow 3\times 3 \rightarrow 1\times 1$ convolutional sequence with an identity shortcut with the exception of the last Global Average Pooling (GAP) layer. The original 1,000-class ImageNet head was replaced with a binary pneumonia classifier and applied a sigmoid activation function which output a pneumonia probability score ranging from (0 to 1). The complete flow of the backbone, from the initial convolution through the four bottleneck stages to the pooled feature vector and the final sigmoid output, is illustrated in Fig. 3.

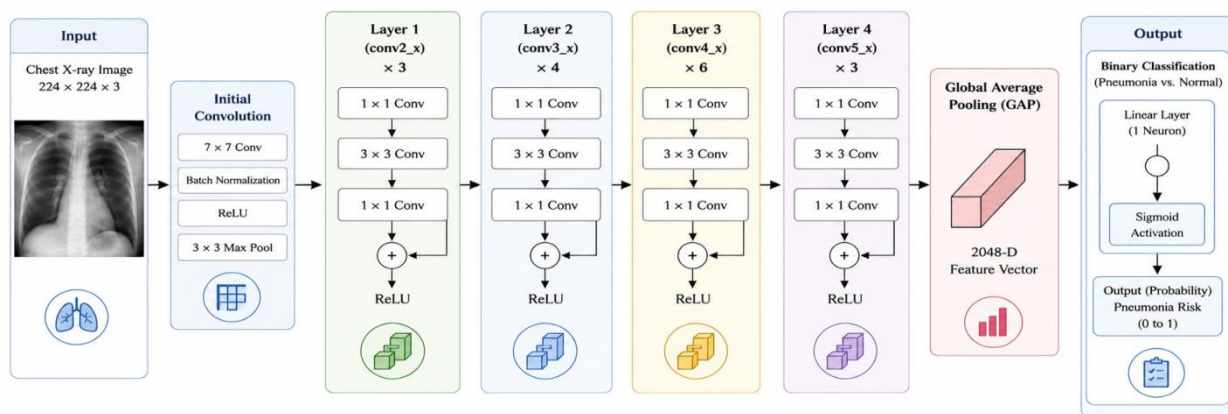


Fig. 3. Model Architecture

3.5 Interpretability Techniques

Interpretability methods break down model predictions into attribution signals that describe a region of the image where the model relied on, and can help clinicians confirm that the classification is based on the pixels used in the diagnostically relevant anatomy and not on noise in the image.

Layer-wise Relevance Propagation was used to interpret the pneumonia classifier by propagating the output score backward through the network and assigning a relevance value to each input pixel, while preserving the total relevance at every layer. Different propagation rules were applied depending on the layer type, including rules suited to the dense layer, the projection and standard convolutions, and the skip connections, with batch normalization statistics fused into the convolutions beforehand and relevance blocked from flowing through inactive units [1, 13].

Grad-CAM was applied to the deepest convolutional layer of the ResNet-50 backbone, since it offers the best balance of semantic detail and spatial resolution. Channel-wise importance weights were obtained by averaging the gradients of the predicted class score with respect to the feature maps, which were then combined with the activations and passed through a ReLU to retain only the regions supporting the prediction [2, 3].

Adversarial training was used to make the model's saliency maps more anatomically coherent by incorporating FGSM-based perturbations into the training process, with the model updated only on these perturbed examples rather than on clean images. This was motivated by evidence that adversarially trained models push their decision boundaries further from the data manifold, producing sharper and more reliable saliency maps, and that models without such defenses can lose substantial accuracy under perturbation [24, 26, 27].

A spatial attention module was built directly into the classification head so that attention weighting is learned jointly with the classification task rather than added afterward. The module applies a convolution and normalization step to the final feature map, followed by a spatial softmax that redistributes attention across locations, and the resulting attention map is multiplied with the feature map before pooling. This pushes the model to focus on diagnostically relevant lung regions while suppressing areas such as bone borders or scanner artifacts [4].

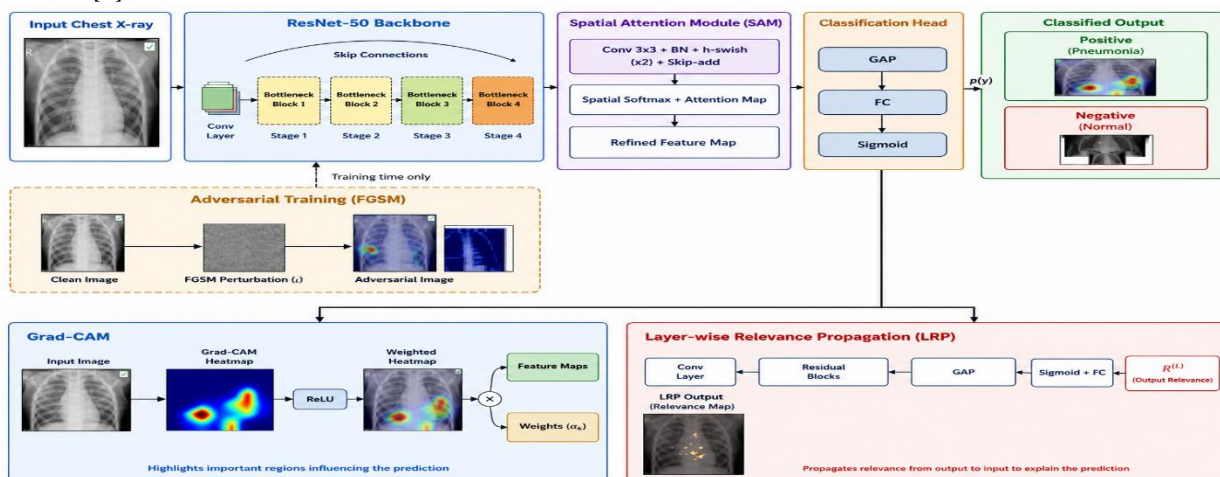


Fig. 4. Unified interpretability framework combining adversarial (FGSM) training, the spatial attention module (SAM), Grad-CAM and layer-wise relevance propagation (LRP) on a shared ResNet-50 backbone

Fig. 4 traces how a single chest X-ray moves through the shared backbone, the spatial attention module and the classification head to reach a sigmoid probability, with the two branches that make the model interpretable shown around this common path. The adversarial training branch on the left operates only during training: a clean image is perturbed with an FGSM gradient step and the perturbed image, not the clean one, is what the backbone is trained on for this variant. Grad-CAM and LRP, shown in the bottom two panels, both start from the classification head's output rather than from the attention module directly; Grad-CAM works forward from the last convolutional feature maps and the class score to build a weighted heatmap, while LRP works backward from the output relevance through the sigmoid, fully connected, pooling, residual and convolutional layers to assign a relevance value to every input pixel. Reading the two bottom panels side by side is the fastest way to see why the two methods disagree on some cases: Grad-CAM answers "which broad region drove this decision", while LRP answers "which individual pixels were used", and the difference in granularity carries through to the Mean Relevance Score comparisons in Section 4.

These methods are based on different principles and provide different perspectives of the decision making process of a CNN. All four methods were used on the same trained ResNet-50 pneumonia classifier based on the controlled comparison framework introduced by Colin et al. [6] so that a direct and fair evaluation of the quality and spatial fidelity of the explanations have been performed.

3.6 Results Evaluation

Classification performance is assessed using some complementary metrics accuracy, sensitivity, specificity, F1 score, AUC-ROC and trade-off score. A high-sensitivity classification threshold applied throughout.

3.6.1 Accuracy

Accuracy, the ratio of the number of instances correctly predicted by a model to the number of instances is calculated:

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

3.6.2 Sensitivity (True Positive Rate)

Sensitivity (Recall) measures the model's performance in correctly identifying true positive cases out of all positive cases:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

3.6.3 Specificity (True Negative Rate)

Specificity is the ability of the model to correctly identify all actually negative cases. In the case of detecting pneumonia, a high specificity means the model is good at excluding false positives, minimizing unnecessary worry for patients who don't have pneumonia:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

3.6.4 F1-Score

The F1-score is a single score combining both precision and recall, and it penalizes model imbalance. In the formula PPV represents the reliability of the positive model output from the clinician's point of view; a high PPV will indicate a higher degree of confidence that when the alert is raised, it is indeed a true alert.

$$\text{F1 Score} = 2 \times \frac{\text{Sensitivity} \times \text{PPV}}{\text{Sensitivity} + \text{PPV}} \quad (5)$$

3.6.5 Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The AUC-ROC depicts the graph of how well the model distinguishes between pneumonia and normal cases at different decision points.

$$\text{AUC - ROC} = \int_0^1 \text{TPR}(t) d(\text{FPR}(t)) \quad (6)$$

3.6.6 Trade-off Score

The Trade-off Score T is a new added metric for Adversarial Training and SAM specifically, because it is a unique feature of these two methods that they modify the training process, while still maintaining classification performance [6].

$$\text{Trade-off Score} = \text{Accuracy} - \lambda \times (1 - \text{MRS}) \quad (7)$$



3.6.7 Mean Relevance Score (MRS)

The MRS is the primary interpretability metric selected because it translates the spatial focus of each explanation method into a single clinically grounded number [6]. Unlike qualitative heatmap inspection, MRS enables direct quantitative comparison across all four techniques on the same scale.

$$\text{MRS} = \frac{1}{N} \sum_{i=1}^N R_i \quad (8)$$

3.7 Statistical Reliability Framework

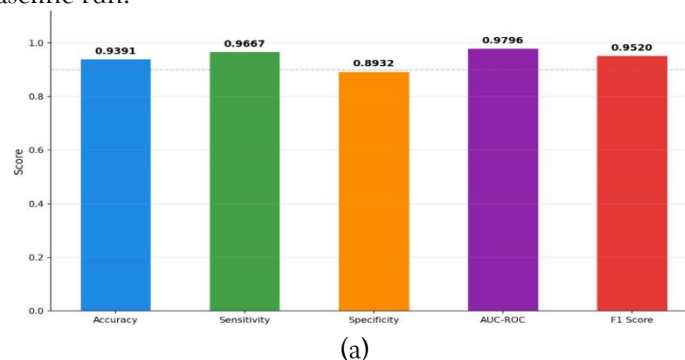
In this research work, 95% confidence intervals for each classification and interpretability metric were estimated via 1,000 bootstrap resamples, percentile method, and McNemar's exact test was applied to each of the six pairwise combinations of the four XAI methods to assess whether per-sample classification accuracy differs significantly between methods. Family-wise error rate across the six comparisons is controlled via Bonferroni correction, yielding an adjusted significance threshold of $\alpha^* = 0.0083$.

4. Results and Discussion

The quantitative and visual results of the baseline ResNet-50 and four variants of interpretability (LRP, Grad-CAM, Adversarial Training, and SAM) trained and tested on a common experimental setup on the CXR Pneumonia dataset are provided here. For each variant classification metrics, a confusion matrix, training history and saliency map are provided as visual output, while the lung region dedicated MRS serves as the anchor for the other approaches, with a comparative summary showing the differences in terms of classification performance and the quality of the spatial explanations provided by all four approaches.

4.1 Pre-Trained ResNet50 without Interpretability Technique Results

The ResNet-50 serves as the performance ceiling against which all interpretability variants are measured. Its results reflect the best possible classification achievable without any interpretability constraint applied during or after training. The classification threshold was set on the validation set so as to ensure the algorithm was able to achieve a minimum sensitivity of 0.90 reflecting a clinical priority of minimizing missed pneumonia cases. The baseline achieved 93.91% accuracy with 96.67% sensitivity a clinically significant result given a specificity of 89.32% indicates that the test is reasonably accurate in correctly identifying normal cases. The F1 score is 95.20 % and AUC-ROC is 97.96 %, confirming a good balance in sensitivity and specificity. These results are summarised in Fig. 5, which reports the five classification metrics alongside the confusion matrix and the training history for this baseline run.



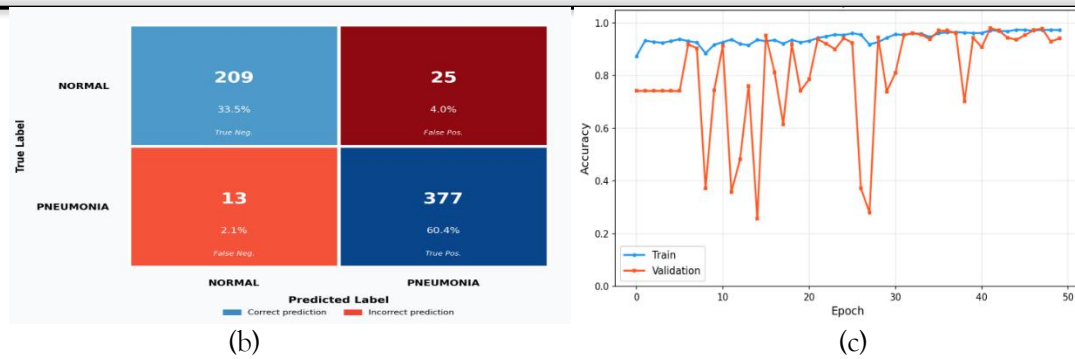


Fig. 5. Baseline ResNet-50 classification performance without an interpretability technique

4.2 Layer-Wise Relevance Propagation Implementation Results

The variant of Composite Hybrid LRP maintained the classification capabilities, with improved anatomical focus. The accuracy was equal to the baseline, which was 93.91%, and sensitivity was 95.90 % and specificity was 90.60 %, proving no diagnostic compromise due to the hybrid rule assignment. F1 (95.17%), and AUC-ROC (98.24%) complete the classification profile. As shown in Figure 6, the overall MRS is (0.9078) which validates the composite rule structure and demonstrates better focus on the lung region.

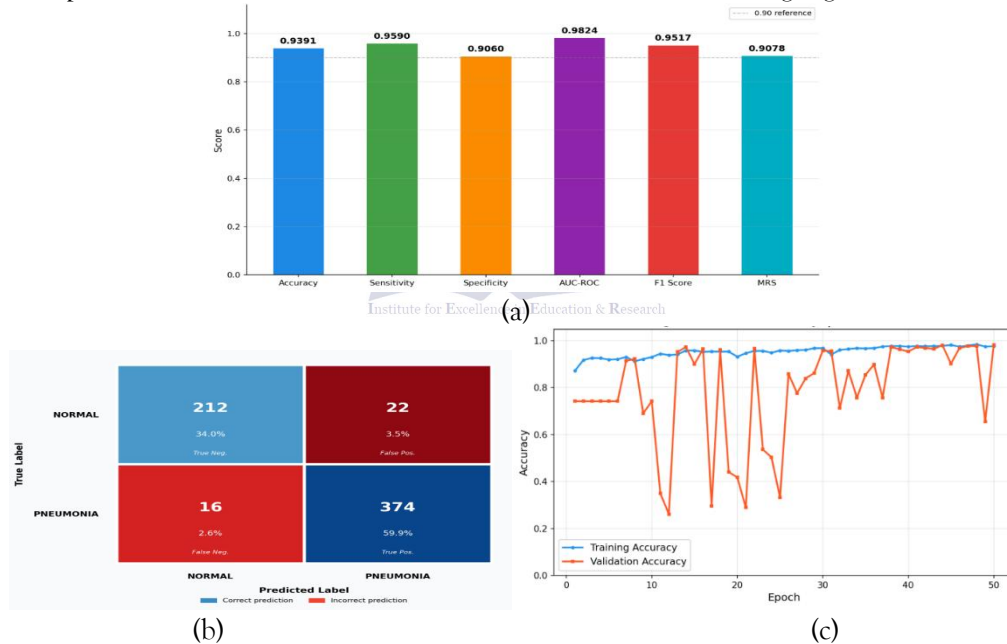


Fig. 6. Composite Hybrid LRP classification performance

LRP maps exhibit a high degree of spatial consistency and are consistent with the expected clinical anatomy, as shown in Fig. 7, presented in three panels: original CXR, raw relevance heatmap, and fused heatmap overlaid on the original CXR. For pneumonia, the high relevance regions are tightly clustered within consolidation/opacity regions of the lung parenchyma.

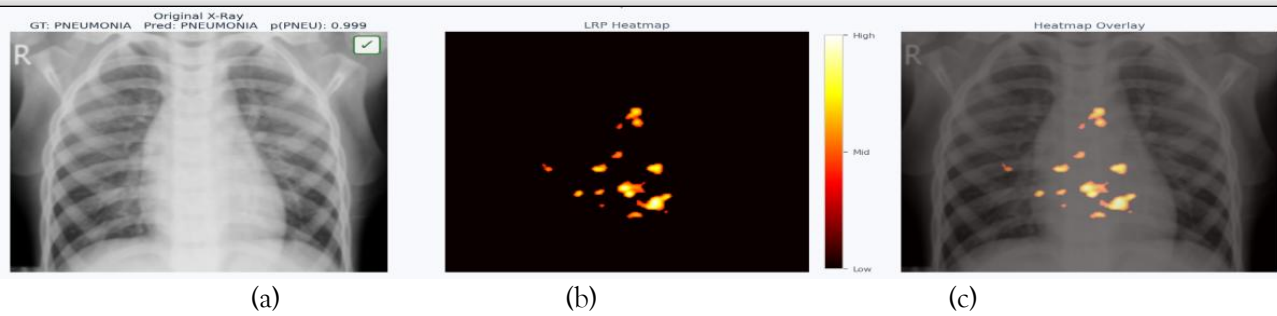


Fig. 7. Composite Hybrid LRP explanation for a representative pneumonia case: (a) original chest X-ray with the model's prediction and probability, (b) raw LRP relevance heatmap, (c) relevance heatmap fused onto the original image

4.3 Class Activation Map Implementation Results

The highest classification performance was achieved by Grad-CAM with all five configurations. It was able to achieve 95.03% accuracy with a sensitivity of 94.87% on the conv5_block3_out layer without changing any training or architecture, and outperforming both the baseline and LRP variant. Specificity achieved 95.30% and F1-score of 95.98% along with the AUC-ROC of 98.72% indicate that this is the most balanced classification profile among all the compared ones. Fig. 8 presents the corresponding metrics, confusion matrix and training curve.

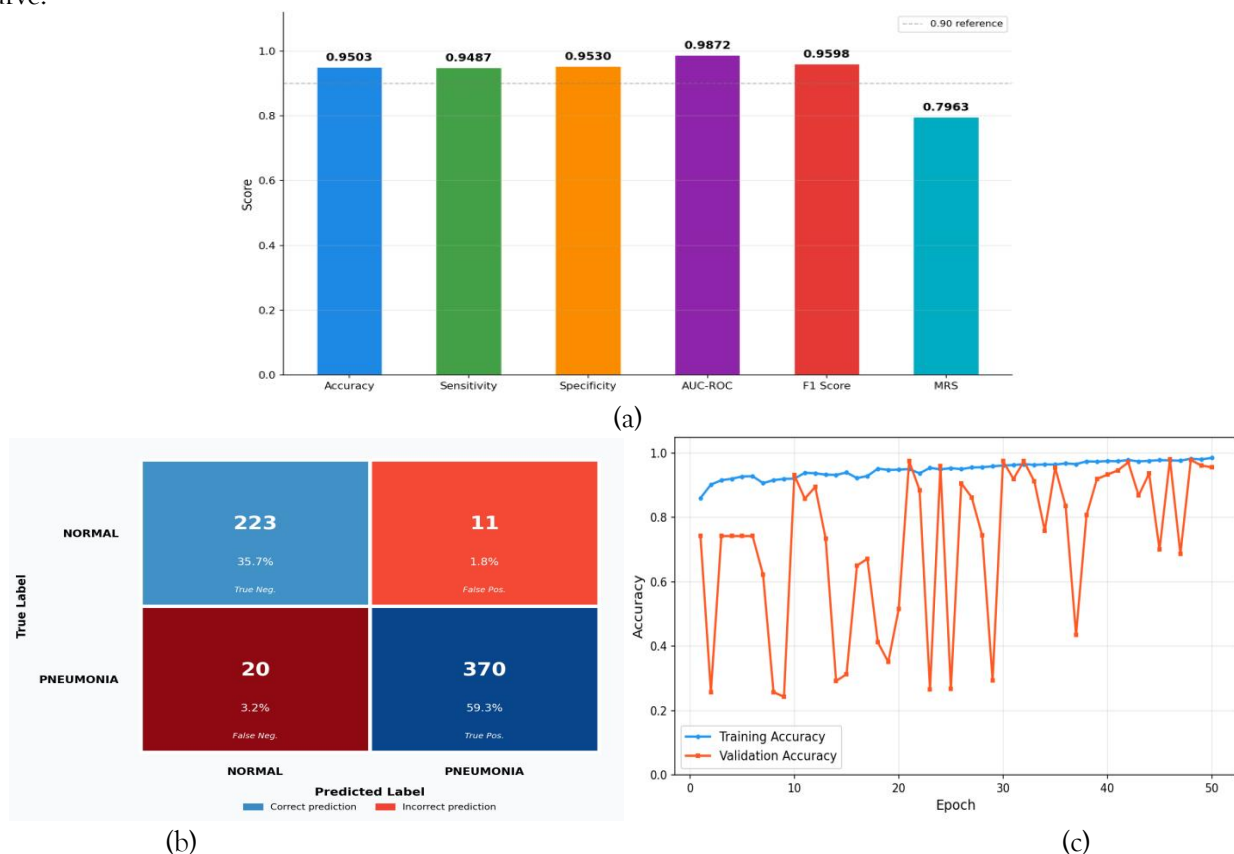


Fig. 8. Grad-CAM classification performance: (a) accuracy, sensitivity, specificity, AUC-ROC, F1-score and MRS; (b) confusion matrix; (c) training and validation accuracy over 50 epochs

Activations of pneumonia are more concentrated in the lower and middle portions of the lung field, which is consistent with the usual consolidation and interstitial pneumonia activation area, as shown for a representative

case in Fig. 9. The size of the activation blobs is larger for Grad-CAM than for LRP, but is still diagnostically meaningful, and not blurred by peripheral artefacts around the bones or soft tissue.

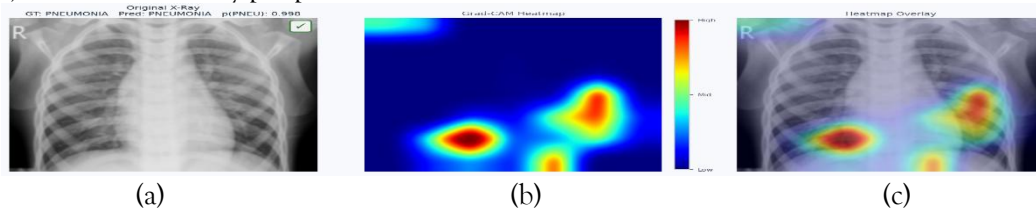


Fig. 9. Grad-CAM explanation for a representative pneumonia case: (a) original chest X-ray with the model's prediction and probability, (b) raw Grad-CAM heatmap, (c) heatmap fused onto the original image

4.4 Attention Mechanism Implementation Results

SAM is the only variant for which interpretability is architecturally built into the model. However, classification performance was not the best; accuracy 93.75%, sensitivity 94.87%, specificity 91.88%, F1-score 94.99% and AUC-ROC 97.94%. The results of the interpretability expose an interesting constraint. The overall MRS was 0.6681, which is the lowest of all four methods. Fig. 10 shows the classification metrics, confusion matrix and training curve for this configuration.

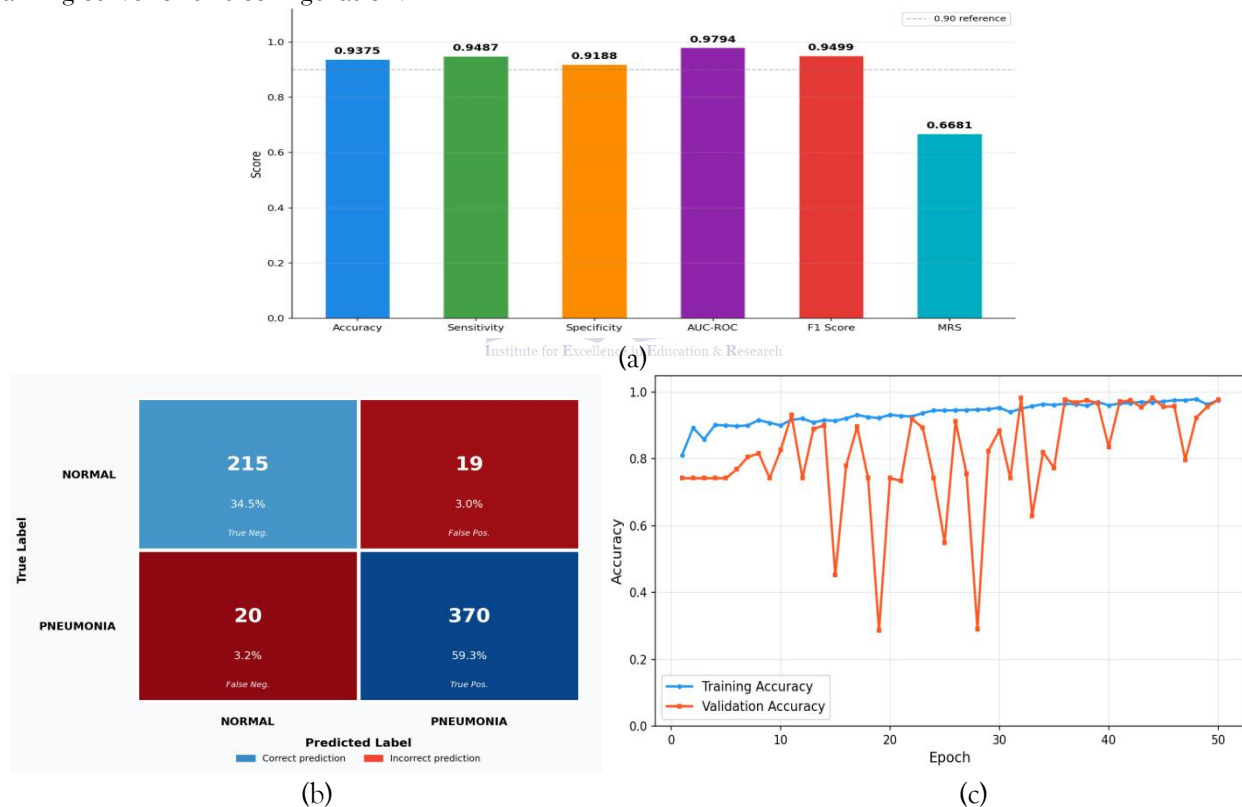


Fig. 10. Spatial Attention Module (SAM) classification performance: (a) accuracy, sensitivity, specificity, AUC-ROC, F1-score and MRS; (b) confusion matrix; (c) training and validation accuracy over 50 epochs

The SAM attention maps are learned softmax distribution over 7×7 feature map, up-sampled to 224×224 , either as an original image, or as a colormap-based attention map, or combined into an overlay, as illustrated in Fig. 11 for a representative pneumonia case. In normal cases, there is a focused attention with primary lung fields. In pneumonia the attention is shifted to neighboring soft tissue and neighboring pleural space which is a core weakness of SAM: the coarse 7×7 feature grid results in a broad probability distribution in the presence of spatially heterogeneous pathology, diminishing the precision of localization when it is most clinically required.

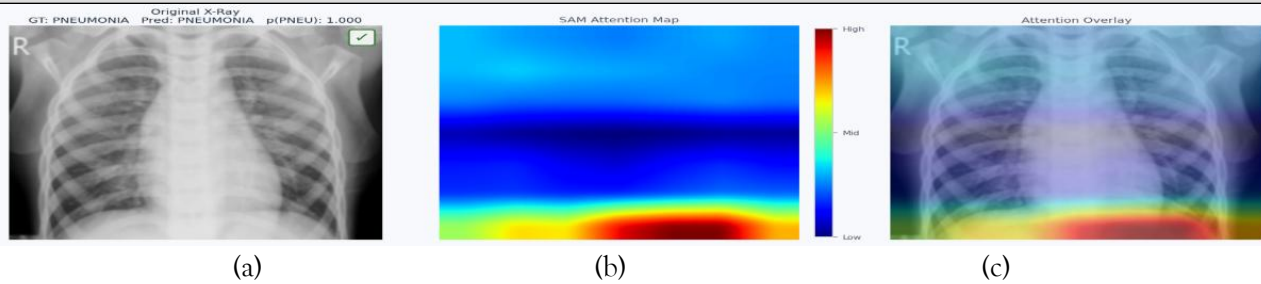


Fig. 11. SAM explanation for a representative pneumonia case

4.5 Adversarial Training Implementation Results

Unlike the other three variants, Adversarial Training only uses FGSM-perturbed images, at a value of $\epsilon=0.01$, and gradient saliency maps $|\partial L/\partial x|$ as the interpretability output from the model. The classification cost of this approach is relatively low: 92.79% accuracy, which is the lowest variant achieved by augmenting the classification with interpretability but just 1.12 points below the baseline without augmentation. Sensitivity of 94.62%, specificity of 89.74%, F1-score of 94.25% and AUC-ROC of 97.45% demonstrate model's ability to remain effective in its diagnostics without ever having been trained on clean images. These results are summarised in Fig. 12.

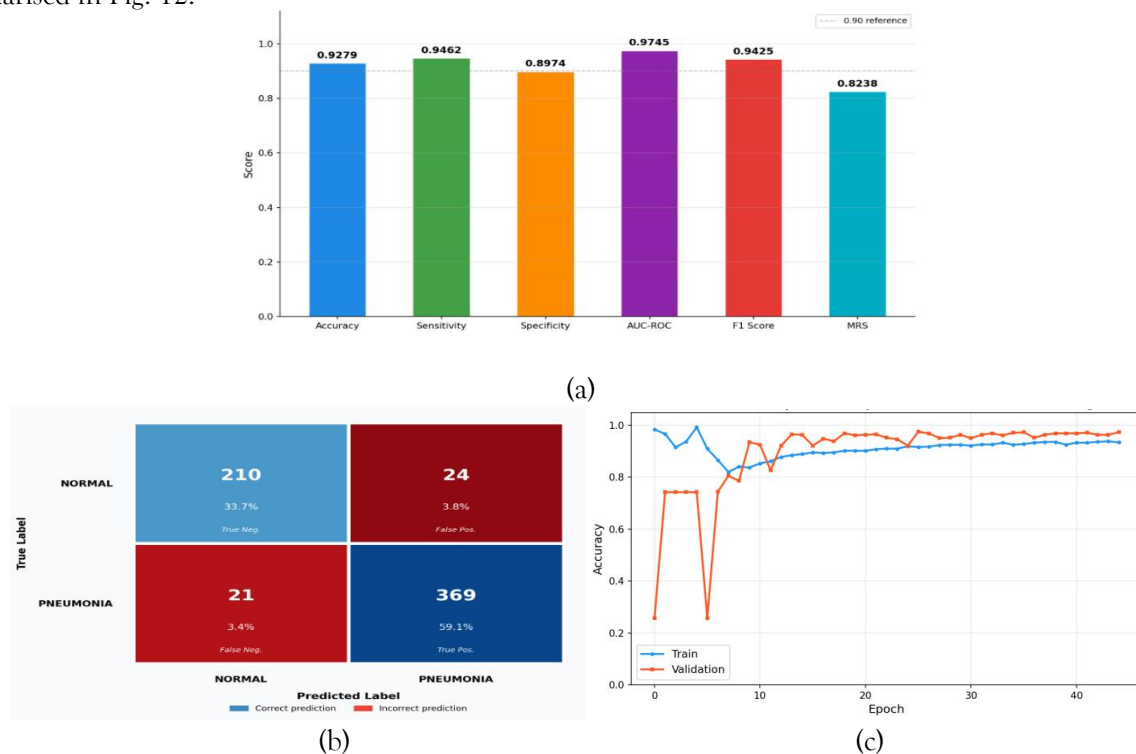


Fig. 12. FGSM Adversarial Training classification performance

Adversarial Training maps are displayed in three different ways: adversarial data, adversarial data + jet-colormap saliency, and adversarial data + raw absolute gradient $|\partial L/\partial x|$, as illustrated for a representative pneumonia case in Fig. 13. Gaussian smoothed and power normalized $|\partial L/\partial x|$. The regions with warmer temperatures represent the most sensitive regions to FGSM perturbation.

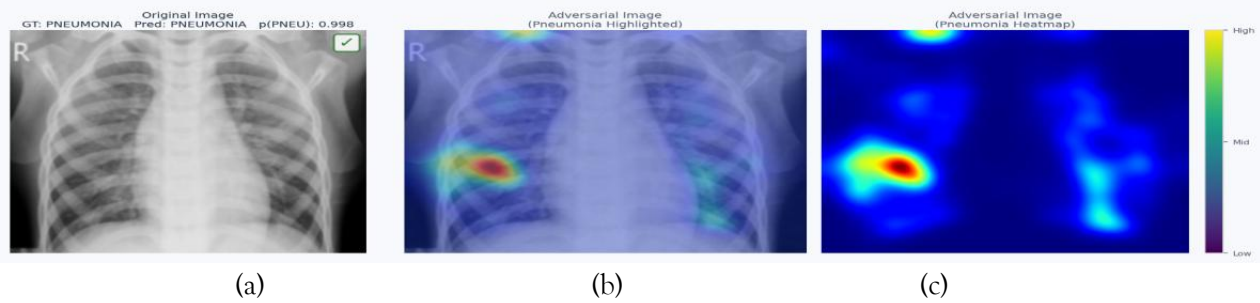


Fig. 13. FGSM Adversarial Training explanation for a representative pneumonia case: (a) original image with the model's prediction and probability, (b) adversarial image with the predicted region highlighted, (c) raw absolute-gradient saliency map

The table summarizes classification and interpretability metrics for each of the 5 configurations along with 95% confidence intervals for each metric, allowing direct comparison of intervals as well as point estimates of the classification and interpretability metrics. The non-overlapping MRS confidence intervals confirm that the interpretability ranking (LRP $\rightarrow$ Adversarial $\rightarrow$ Grad-CAM $\rightarrow$ SAM) is stable across uncertainty bounds. By contrast, the rather wide accuracy confidence intervals suggest that differences in accuracy between variants are very small in absolute terms and might not represent true model-level differences, which we formally tested.

Table 1: Classification and Interpretability Metrics with 95% Confidence Intervals.

Model	Accuracy	Sensitivity	Specificity	AUC-ROC	MRS
ResNet-50 (Baseline)	0.9391 (0.915–0.963)	0.9667 (0.949–0.985)	0.8932 (0.854–0.932)	0.9796 (0.9560–1.000)	~
ResNet-50 + LRP	0.9391 (0.915–0.963)	0.9590 (0.940–0.978)	0.9060 (0.869–0.943)	0.9824 (0.9726–0.9906)	0.9078 (0.903–0.913)
ResNet-50 + Grad-CAM	0.9503 (0.929–0.972)	0.9487 (0.927–0.971)	0.9530 (0.926–0.980)	0.9872 (0.9787–0.9938)	0.7963 (0.790–0.803)
ResNet-50 + SAM	0.9375 (0.914–0.961)	0.9487 (0.927–0.971)	0.9188 (0.884–0.954)	0.9794 (0.9691–0.9883)	0.6681 (0.659–0.677)
ResNet-50 + Adversarial Training	0.9279 (0.903–0.953)	0.9462 (0.924–0.969)	0.8974 (0.858–0.937)	0.9745 (0.9620–0.9846)	0.8238 (0.818–0.830)

A non-parametric hypothesis test was used to evaluate whether the observed classification differences are real or due to sampling variation, on the $n = 624$ test partition. For paired comparison of binary classification results, we use McNemar's exact test, which is specially designed for the comparison of two classifiers tested on the same test samples. Statistical significance was determined at $\alpha = 0.05$ for six pairwise comparisons using Bonferroni correction ($\alpha^* = 0.008$).

Table 2: McNemar's Test Results for Pairwise Classification Accuracy Comparison.

Comparison Pair	χ^2	p-value	Conclusion
Grad-CAM vs. LRP	7.875	0.005	Significant *
Grad-CAM vs. SAM	1.750	0.186	Not significant
Grad-CAM vs. Adversarial Training	4.225	0.040	Not significant
LRP vs. Adversarial Training	0.980	0.322	Not significant
SAM vs. Adversarial Training	0.735	0.391	Not significant

Comparison Pair	χ^2	p-value	Conclusion
LRP vs. SAM	3.380	0.066	Not significant

As presented in Table 2, the differences in classification accuracy among most of the model pairs were not statistically significant at the Bonferroni-corrected threshold except for the LRP vs Grad-CAM pair ($\chi^2 = 7.875$, $p = 0.005$). For the remaining five pairs classification differences are non-significant, indicating that the other variants achieve broadly equivalent classification performance on this test partition. This result confirms an interesting finding, namely that these four XAI variants are statistically equal on this test partition. The choice of the XAI method should be based on the interpretability quality and on clinical coherence, rather than on minor differences in accuracy, that are the result of sampling variability.

4.6 Comparative Analysis

The most important improvements that have been added to this study are compared against Colin et al. [6] which was the foundation study for this research work. The overall comparison of all five variants and six for Adversarial training and SAM of the proposed model, as well as the baseline ResNet-50 model of the aforementioned study is shown in table.

Table 3. *Comparison of Results Colin et al. [6].*

Model Variant	Metric	Colin & Surantha	This Study	Δ Improvement
ResNet-50 (Baseline)	Accuracy	0.90	0.9391	+0.0391
	Sensitivity	0.92	0.9667	+0.0467
	Specificity	0.88	0.8932	+0.0132
	AUC-ROC	0.93	0.9796	+0.0496
ResNet-50 + LRP	Accuracy	0.91	0.9391	+0.0291
	Sensitivity	0.90	0.9590	+0.0590
	Specificity	0.92	0.9060	-0.0140
	AUC-ROC	0.93	0.9824	+0.0524
ResNet-50 + Grad-CAM	MRS	0.85	0.9078	+0.0578
	Accuracy	0.9050	0.9503	+0.0453
	Sensitivity	0.86	0.9487	+0.0887
	Specificity	0.83	0.9530	+0.1230
	AUC-ROC	0.89	0.9872	+0.0972
	MRS	0.70	0.7963	+0.0963
ResNet-50 + SAM	Accuracy	0.8800	0.9375	+0.0575
	Sensitivity	0.90	0.9487	+0.0487
	Specificity	0.87	0.9188	+0.0488
	AUC-ROC	0.91	0.9794	+0.0694
ResNet-50 + Adversarial Training	MRS	0.12	0.6681	+0.5481
	Trade-off Score	0.42	0.6056	+0.1856
	Accuracy	0.85	0.9279	+0.0779

Model Variant	Metric	Colin & Surantha	This Study	Δ Improvement
	Sensitivity	0.87	0.9462	+0.0762
	Specificity	0.79	0.8974	+0.1074
	AUC-ROC	0.87	0.9745	+0.1045
	MRS	0.76	0.8238	+0.0638
	Trade-off Score	0.79	0.7517	-0.0383

All 5 configurations have sustained improvements as illustrated in Table 3. Altogether, these results validate the added value of the enhanced pipeline in terms of classification accuracy and interpretability quality when compared with the original baseline. The Trade-off Score (TS) was introduced by Colin and Surantha [6] as a composite metric for adversarial training and SAM only capturing joint diagnostic accuracy and spatial explanation quality, reflecting the dual requirement of clinical AI systems. High TS indicates a model that both classifies correctly and produces anatomically coherent explanations.

5. Conclusion and Future Work

This work meets the challenge by assessing four XAI approaches within a standardized framework, on a common ResNet-50 backbone and a novel anatomical attribution quality metric. CXR imaging is the most common method used in the diagnosis of pneumonia, which is one of the most common causes of death in children under five years of age worldwide. This study aimed to compare LRP, Grad-CAM, Adversarial Training with FGSM and SAM in a controlled four-way scenario with the same experimental setup. All, four methods were used with the same ResNet-50 backbone trained on the Kaggle CXR Pneumonia dataset. LRP obtained perfect interpretability quality (MRS = 0.9078, 100% of the relevance is conserved) without any sacrifice of classification (93.91% accuracy). Grad-CAM obtained the highest classification accuracy (95.03%) and AUC-ROC (98.72%) with the low focused on relevant region in comparison of LRP (MRS = 0.7963). SAM performed competitively well (93.75%) and had the lowest MRS (0.6681). While the adversarial training have high MRS value after LRP which is (0.82). The non-overlapping 95% confidence intervals for MRS across the four methods indicate consistent, meaningful differences in interpretability quality, while McNemar's test found the difference in classification accuracy to be statistically significant for only one pair (LRP vs Grad-CAM, $\chi^2 = 7.875$, $p = 0.005$), with the remaining five pairs statistically equivalent. The choice of methods for XAI should, therefore, be based on the quality of interpretability and clinical consistency and not on small differences of accuracy.

This work contributes to the state of the art in interpretable pneumonia detection and directly improves the baseline for all the techniques analyzed however some limitations and open directions still encourage researchers to continue.

1. To replicate the full pipeline on other datasets, like NIH ChestX-ray14 or CheXpert.
2. Combinations of the methods such as merging LRP relevance maps with SAM attention or spatially constraining the LRP backward pass with Grad-CAM activation may result in better explanations.
3. The existing framework only allows binary pneumonia vs normal classification, multiclass is required.
4. More capable attention architectures like convolutional block attention, deformable attention, confined to a lung segmentation mask, could overcome low MRS issues.

References

1. Bassi, P.R., S.S. Dertkigil, and A. Cavalli, *Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization*. Nature Communications, 2024. 15(1): p. 291.

2. Manju, A., et al. *High-Resolution Pneumonia Detection and Multi-Class Classification Using VGGNet-19 with Transfer Learning and Explainable AI for Medical Imaging*. in 2025 2nd International Conference on Computing and Data Science (ICCDs). 2025. IEEE.
3. Ennab, M. and H. Mcheick, *Enhancing Pneumonia Diagnosis through AI Interpretability: Comparative Analysis of PixelLevel Interpretability and Grad-CAM on X-ray Imaging with VGG19*. IEEE Open Journal of the Computer Society, 2025.
4. Potharaju, S., et al., *Enhanced X-ray image Classification for Pneumonia Detection Using Deep Learning Based CBAM and SE Mechanisms*. Intelligence-Based Medicine, 2025: p. 100299.
5. Hadavale, G., *Beyond Accuracy: An Interpretability-Driven Audit of Deep Learning Models for Pneumonia Detection from Chest X-Rays*. Authorea Preprints, 2026.
6. Colin, J. and N. Surantha, *Interpretable deep learning for pneumonia detection using chest x-ray images*. Information, 2025. **16**(1): p. 53.
7. Priyatharsini, G.S., et al. *Revolutionizing lung disease diagnosis: a unique hybrid deep learning framework for explainable chest X-ray analysis*. in *Proceedings of the 3rd International Conference on Optimization Techniques in the Field of Engineering (ICOFE-2024)*. 2024.
8. Patil, A., A. Singh, and S. Amaran. *X-ray image classification for pneumonia detection using neural networks*. in *AIP Conference Proceedings*. 2025. AIP Publishing LLC.
9. Yazdani, E., et al., *Explainable artificial intelligence for pneumonia classification: clinical insights into deformable prototypical part network in pediatric chest x-ray images*. Journal of medical imaging and radiation sciences, 2025. **56**(5): p. 102023.
10. Mahfuz, S., et al. *Explainable AI Based Multi-Class Pneumonia Classification and Anomaly Detection from Chest X-ray*. in 2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS). 2025. IEEE.
11. Reddy, K.T., *Explainable deep learning for automated pneumonia detection from chest X-ray images using MobileNetV2 and Grad-CAM CNN*. Authorea Preprints, 2025.
12. Prodhan, M.M.A. and M.N. Hossain. *Interpretable Pneumonia Detection in Chest X-Rays via VICReg Self-Supervised Pretraining*. in 2025 IEEE 8th International Conference on Multimedia Information Processing and Retrieval (MIPR). 2025. IEEE.
13. Otsuki, S., et al. *Layer-wise relevance propagation with conservation property for resnet*. in *European Conference on Computer Vision*. 2024. Springer.
14. Neenu, S. and B. Ankayarkanni, *Enhanced ResNet-50 with multi-feature fusion for robust detection of pneumonia in chest X-ray images*. Diagnostics, 2025. **15**(16): p. 2041.
15. Akhila, P., et al. *Interpretable Deep Learning Model for Pneumonia Diagnosis from Chest X-Rays*. in 2025 Third International Conference on Emerging Applications of Material Science and Technology (ICEAMST). 2025. IEEE.
16. Ali, M., et al., *Pneumonia detection using chest radiographs with novel efficientnetv2l model*. IEEE Access, 2024. **12**: p. 34691-34707.
17. Chakraborty, S., et al., *Deep Learning and X-Ray Imaging Innovations for Pneumonia Infection Diagnosis: Introducing DeepPneuNet*. Computational Intelligence, 2025. **41**(1): p. e70029.
18. Dalvi, T., et al., *Pneumonia Detection Using Chest X-Rays: A Comparative Study of CNN Architectures and Explainable AI Techniques*, in *Advancements in IoT Sensors and Security: Harnessing AI/ML for Secured IoT Sensor Data*. 2026, Springer. p. 219-245.
19. Hasnat, M.A., et al., *Interpretability of Machine Learning Models in Medical Image: Analysis A Study Using the RSNA Pneumonia Dataset*. Authorea Preprints, 2025.
20. Li, Y., et al., *A comprehensive segmentation of chest X-ray improves deep learning-based WHO radiologically confirmed pneumonia diagnosis in children*. European Radiology, 2024. **34**(5): p. 3471-3482.

21. Erukude, S.T., V.C. Marella, and S.R. Veluru. *Explainable Deep Learning in Medical Imaging: Brain Tumor and Pneumonia Detection*. in 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). 2025. IEEE.
22. Jihad, M., et al. *Explainable Artificial Intelligence for Deep Learning-Based Detection of Pneumonia in Chest X-ray Images*. in International Conference on Innovations in Data Analytics. 2024. Springer.
23. Mahamud, E., et al., *An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning*. Decision analytics journal, 2024. 12: p. 100499.
24. Ashraf, S.N., R. Siddiqi, and H. Farooq, *Auto encoder-based defense mechanism against popular adversarial attacks in deep learning*. PloS one, 2024. 19(10): p. e0307363.
25. Mozhegova, E., et al., *Assessing the adversarial robustness of multimodal medical AI systems: insights into vulnerabilities and modality interactions*. Frontiers in Medicine, 2025. 12: p. 1606238.
26. Rafferty, A., R. Ramaesh, and A. Rajan. *CoRPA: Adversarial Image Generation for Chest X-rays Using Concept Vector Perturbations and Generative Models*. in 2025 IEEE 13th International Conference on Healthcare Informatics (ICHI). 2025. IEEE.
27. Wong Zi Ling, N. and N. Surantha, *Synthetic-augmented adversarial training for robust chest x-ray classification*. Discover Artificial Intelligence, 2026. 6(1): p. 222.
28. Vyas, R. and R.K. Pandey, *Deep learning for pneumonia detection from X-ray: A systematic review of models, datasets, and clinical translation challenges*. Biomedical Signal Processing and Control, 2026. 115: p. 109272.

