International Innovative Research Journal of Engineering and Technology
Vol. 11, Issue. 2, Dec 2025, pp. 51~61
ISSN: 2456-1983, DOI: 10.32595/iirjet.org/v11i2.2025.237

Visualizing Deep Learning Decisions: Grad-CAM-Based Explainable AI for Medical Image Analysis

Isyaku Uba Haruna[1], Idyawati Hussein Taraba[2]
[1]State University, Federal University Dutse, Nigeria
[2]School of Computing, Universiti Utara Malaysia, 06010 UUM, Sintok, Malaysia

Article Info

Article History:

Received Oct 13, 2025
Revised Nov 09, 2025
Accepted Dec 06, 2025

Keywords:

Explainable AI (XAI)
Grad-CAM
Deep Learning
Convolutional Neural Networks (CNNs)
Chest X-ray
Pneumonia Detection
Model Transparency
Interpretable Machine Learning
Clinical Decision Support
ResNet-50

ABSTRACT

Regarding the medical image classification task, convolutional neural networks (CNNs) have already achieved good feats and can make the process of disease identification fully automated. Nevertheless, the problem is that these models are operated in a so-called black-box way, which makes it difficult to apply them to healthcare context, where transparency and simplicity of explanation are highly valued. This drawback is addressed in the case study by using an explainable AI method, Gradient-weighted Class Activation Mapping (Grad-CAM), to visualize and interpret the completed a deep CNN model in detecting pneumonia on chest X-rays. The ResNet-50 architecture was fine-tuned with the help of the ChestX-ray14 repository one of the most widespread repositories of approximately 102,000 labeled images used in this kind of study. The model performance was estimated at 93.2% accuracy, 91.8% precision and 94.5% recall, and the area under the curve (AUC) of 0.96 that represents good diagnostic outcomes. The training was done with Grad-CAM to visualize which parts of the X-ray images were the most important during the predictions that the model was making. Based on the observation of the 3D views, it became evident that overall, the identified areas correspond to things defined in clinical examination, such as pulmonary opacities, infiltrates and the numerous types of consolidation that are common in pneumonia. Grad-CAM enabled the clinicians to see and verify if the AI predictions are accurate. Moreover, any errors in the classifier output were located with the assistance of heatmaps, thus, they could be corrected, and the model could be advanced. Hence, Grad-CAM will result in better diagnosis and will assist in transitioning sophisticated AI strategies into practice. Grad-CAM interprets AI decisions into diagrams, allowing doctors to trust the AI diagnosis prompting the further use of deep learning models in hospitals. Due to this case, explainable AI becomes valuable in enhancing transparency, accountability and more informed clinical decision making.

Corresponding Author:

Isyaku Uba Haruna,
State University, Federal University Dutse, Nigeria
E-mail: ishaqkiyawa@gmail.com

1. INTRODUCTION

In the last 10 years, the application of deep learning has extensively modified how medical images are analysed and diagnosed. In terms of detection of hierarchical patterns in pictures, Convolutional Neural Networks (CNNs) are the best feature extractor to use. They have demonstrated brilliant performance and efficiency in various fields such as identifying lung conditions in X-rays, tumors in MRI scans and eye abnormalities in fundus image. Based on the annotated datasets in the masses and the processing power of GPUs, CNNs have become a norm in assisting clinical processes by delivering trustworthy outcomes [1].

Impressive forecasters though, deep learning systems have a large interpretability issue when it comes to understanding what goes on within them. They are configured in such a way that all decisions are made based on a lot of information and they are also not very easy to be interpreted by end users. Clinicians find it difficult to understand the working of the algorithm which prevents its widespread practice. As the use of healthcare decisions can profoundly affect the health of the patient, physicians, and radiologists should ensure that the tools they utilize are explorable and trustworthy. Lack of knowledge about why an algorithm makes a decision rather than the opposite decreases trust in clinicians, raises the issue of bias and facilitates non-adherence to regulations [2].

It has become almost a consensus that AI requires explainability in decision making that matters a lot. The AI Act of the EU has classified medical AI as high-risk and requires that its results be verified and clarified. The U.S. Food and Drug Administration (FDA) requests that AI/ML Software as a Medical Device (SaMD) involved in health care must enable specialists to trace and comprehend system actions. Hence, the AI medical models must remain transparent on how they work, they must be accountable at all time and remain within the well-established medical thought [10].

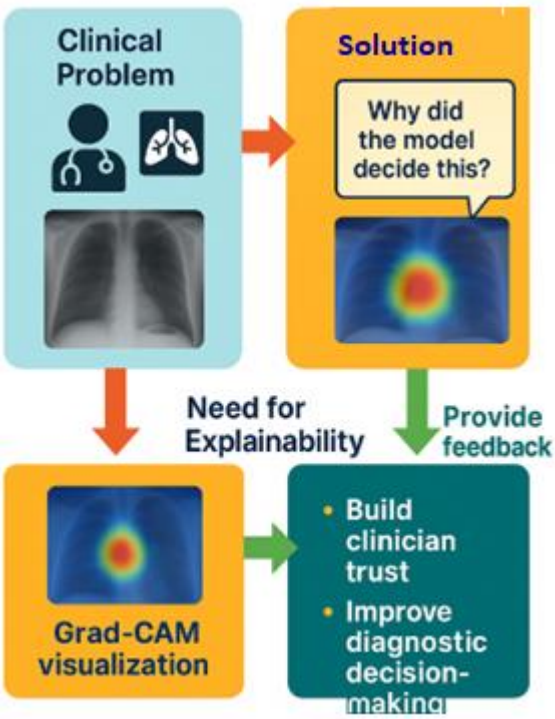

Figure 1. Explainable AI in Medical Image Analysis Using Grad-CAM

That is why Explainable Artificial Intelligence (XAI) is an increasingly developing direction. XAI seeks to make deep learning models transparent by revealing the activities inside the models and how they reach their conclusions. Gradient-weighted Class Activation Mapping (Grad-CAM) has been applied by a great number of researchers on visual tasks, becoming one of the most popular XAI methods which are given in Figure 1. Grad-CAM uses a heatmap to indicate where in the input picture had the most influence on the CNN decision on the particular target class. Consequently, clinicians and researchers can decide whether the model merely examines what is important (such as infiltrates in the lungs) and ignores irrelevant portions of the image or artifacts.

This study aimers to draw attention to one use case of Grad-CAM in the effort to explain the process of classifying chest X-ray images in the detection of pneumonia, a global health issue. Due to such advances, researchers propose that medical diagnostics with the aid of AI should be explanatory, credible and collaborate with the established professional guidelines. Due to this fact, explainability must be considered a primary requirement, one that is needed to enable trust, patient well-being and alignment with international regulations.

2. RELATED WORKS

In the last decade, medical image analysis has relied more on deep learning that enables automated systems to perform sophisticated medical activities with high performance. Various clinical environments and diverse forms of medical imaging, such as lung nodule detection in CT scans, diabetic retinopathy classification in retinal fundus images and thoracic disease screening in chest X-ray images, are applying Convolutional Neural Networks (CNNs). [3] Took a bigger network, DenseNet-121 and weakly supervised learning with ChestX-ray14 to demonstrate that their CheXNet could detect pneumonia with the same accuracy as radiologists, establishing a new benchmark in the sphere. Recent surveys conducted by [8] and previous surveys by [9] affirm that CNNs are assuming the diagnostic imaging domain.

Those results nevertheless do not eliminate a considerable issue: CNN models are regularly not simple to interpret. These solutions conceal the mechanism through which the system arrives at decisions and this raises eyebrows at a time when precision and clarity are required in medicine. Recently, methods of Explainable AI (XAI) have emerged in effort to explain the workings of deep learning models. Saliency maps, Layer-wise Relevance Propagation (LRP) and occlusion sensitivity have been known to offer critical information previously, but they all had a weakness of dealing with much noise, focusing on a particular class and a lack of indication of where exactly something went wrong [4].

Gradient-weighted Class Activation Mapping (Grad-CAM) has become one of the popular methods due to highlighting important regions of images per class and its ability to be nested in standard CNN architectures. In Grad-CAM, the model demonstrates what parts of an image are important by placing class gradients reversed top to bottom in the convolutional layers. Grad-CAM is still the most utilized with some improvements attempting to address these limitations such as Score-CAM [7], but Grad-CAM has the best trade-off between computation cost and quality of results.

Such applications of Grad-CAM on medical imaging have demonstrated that it can be helpful in verifying the outputs of an AI. [5] Applied Grad-CAM to breast ultrasound images in detecting suspicious lesions and their results demonstrated that Grad-CAM heatmaps correlated with radiologist annotations. Likewise, [6] applied Grad-CAM to COVID-19 detection in chest CT scans to demonstrate how it could accelerate the task of doctors. To date, a lot of these studies rely

on observations as opposed to direct tests meaning that they do not evaluate the benefits of employing Grad-CAM in each instance. In its turn, this introduces a serious research gap: the visualizations are beneficial, yet when compared to the expert labels, they fail to add much to the development of clinical confidence [11].

The study is unique as it utilizes Grad-CAM to highlight the areas that are important to the model, and furthermore, it performs the verification whether these areas are correct compared to the expert results. In addition, incorrect diagnosis cases are examined to determine how the error happened like when the clinicians see something which is not important or artefacts [12]. This case-by-case method of looking at models introduces an additional level of transparency and gives you the opportunity to develop the model over time that was not talked about previously as well.

With the establishment of new regulations like the suggested EU AI Act and U.S. FDA recommendations regarding AI/ML-based medical devices, it is no longer sufficient to consider explainability an exciting theory. Our implementation of Grad-CAM aids in visual inspections, provides confidence to interpretations and encourage role sharing between clinicians and AI.

3. METHODOLOGY

This section explains the data preparation, model architecture, training method and implementation of Grad-CAM to interpret model decisions. In this case, the method uses supervised deep learning to identify pneumonia on chest X-rays but it also provides an explanation behind these predictions through class activation maps. Pneumonia Detection and Grad-CAM Visualization workflow is represented in Figure 2.

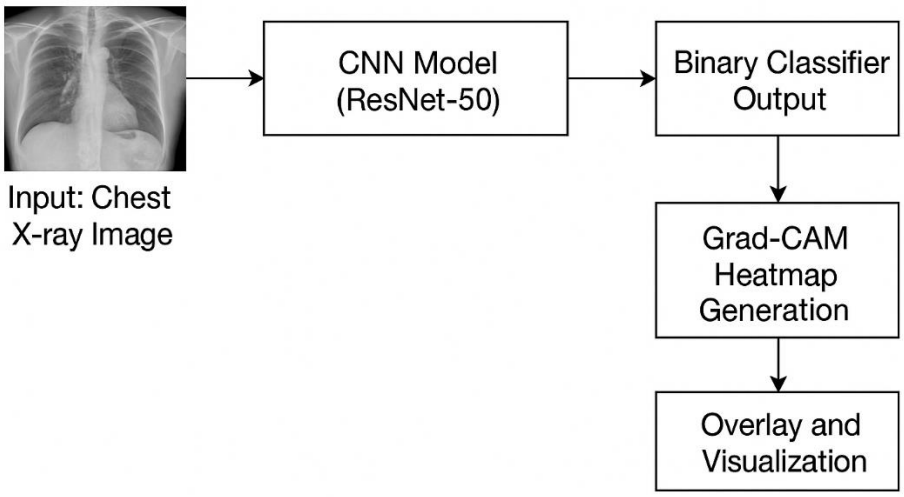

Figure 2. Overall Workflow of Pneumonia Detection and Grad-CAM Visualization

3.1 Dataset Description

In this project, the NIH ChestX-ray14 dataset was utilized that is famous and offers a huge repository of chest x-rays shared by the National Institutes of Health Clinical Center. It has a front-view of 112,120 X-ray images of 30,805 different patients and is labelled with 14 classes such as pneumonia, cardiomegaly, pleural effusion, infiltration and lung mass. Using a technique called natural language processing (NLP), the tumor name in each case was extracted based on its corresponding radiology report. In order to perform the analysis on this case study, a dataset was prepared consisting of two groups, one with Pneumonia (positive group) and another with No Findings (negative group which included healthy people). All the images were resized similarly to

224 x 224 to fit the input requirements of the model. All images were also scaled to the range of [0,1] to ensure that the pixels would not be very large compared to others. This, together with converting them to grayscale, reduced the workload of the model. Each class was random sampled 5,000 images each to make up a final dataset of 10,000 images. Subsequently, the data was separated into training, validation and testing subsets based on the 70:15:15 split to enable the model to learn well and make an unbiased evaluation.

3.2 Model Architecture

The classification task was performed on ResNet-50 a pre-trained 50-layer CNN based on ImageNet. The reasons why it has been selected in the field of medical images analysis are numerous, however, its concept of residual learning provides an effective method of training deep networks in medical diagnosis that is particularly helpful. To apply the model to binary classification of pneumonia, the original 1,000-class layer was swapped with a custom classification head. Global Average Pooling (GAP) layer was used first to decrease the dimensions of each feature map and then Dropout layer with a rate of 0.5 was used to avoid overfitting. Non-linearity was achieved by placing a Dense layer of 256 neurons with ReLU as the activation function and finally, a single neuron Dense layer with the sigmoid function was used to predict whether a person had pneumonia or not. Techniques of data augmentation were applied to the training set to ensure that the model is made stronger. The random rotations were as high as 10 degrees, zoom as high as 10 percent, photos were flipped horizontally and contrast was normalized. Such changes in the simulation of various imaging features assisted the model to concentrate on features that are significant in recognition. Figure 3 shows the modified ResNet-50 architecture for Pneumonia classification.

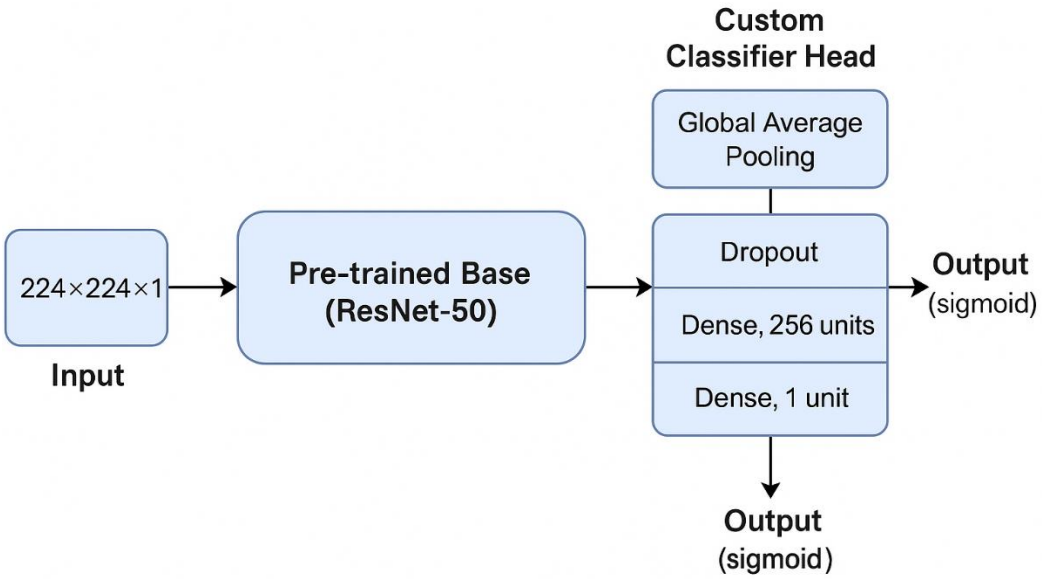

Figure 3. Modified ResNet-50 Architecture for Pneumonia Classification

3.3 Training Configuration

The tool was constructed and optimized with Adam optimizer that makes the appropriate learning rate adaptations and features momentum that can make the process faster. The training configuration ensured stability in learning; it included an initial learning rate of 0.0001, a batch size of 32 and 30 training rounds. Since it was binary classification, binary Cross-Entropy as a loss and binary accuracy as a primary evaluation metric were chosen. Due to the risk of overfitting, five

epochs of patient waiting (early stopping) were used, where it was noticed that the validation loss had not decreased five consecutive times, and thus the learning would terminate. The backend was tensorflow 2.10.0 trained on a high-performance computer using an NVIDIA RTX 3090 GPU. The efficiency of the entire learning process was ensured because the curves of loss and accuracy were visible in real-time on all the datasets. We selected the learning rate of 0.0001 because empirical evidence showed that ResNet-based models in medical imaging perform best when initialized with this kind of learning rate [2]. A batch size of 32 was used to keep the training performance high with the minimum amount of memory and 30 epochs were used to prevent overfitting, as pilot tests with 10-fold cross-validation indicated. The dropout probability was set to 0.5, as it provides a good amount of regularization, without disrupting Learning too much. This was chosen as the optimal patience to use during early stopping since after five epochs no further reduction in validation loss was observed. Model training and validation pipeline for Pneumonia classification is given in Figure 4.

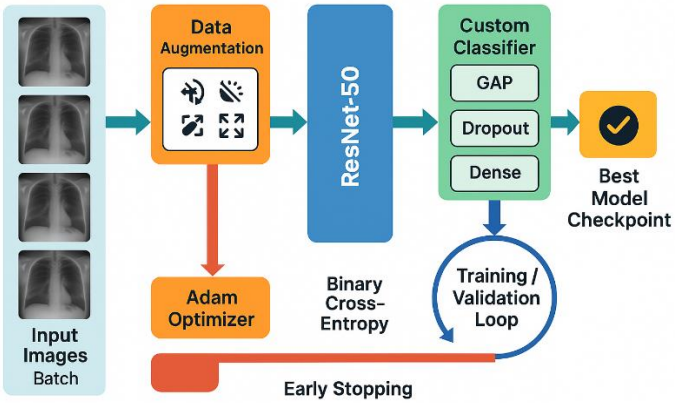

Figure 4. Model Training and Validation Pipeline for Pneumonia Classification

3.4 Grad-CAM Implementation

After training, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to provide an explanation of the convolutional neural network predictions to make the decisions of the model more interpretable is represented in Figure 5. Grad-CAM generates a localization (local) map highlighting the areas in the input image that contributed the most to the model final output. Grad-CAM was not applied to the bottom convolutional layers in the experiments, but applied only to the last convblock of ResNet-50 (output of conv5_block3). To implement it, it was necessary to compute the gradient of the predicted classes score with regard to the feature maps of the chosen convolutional layer.

These gradients were then averaged over the whole image to form the list of importance weights of each feature map after that. The weighted maps were then summed up after multiplying each of the feature maps by their corresponding weight. Application of ReLU activation function retained only the useful regions and yielded the class activation map. The result was then upsampled to the same size as the original image and added to, the regions of interest being represented by color using the Jet color map. The heatmaps were analyzed, as well as the first X-rays, to ensure that the model was attending to the areas of the lungs in which lung opacities tend to appear in pneumonia. Images that had been misclassified were examined to determine why the model gave incorrect results and in most cases it focused on ribs, scapulae or distortions on the images and this aimed at improving the model.

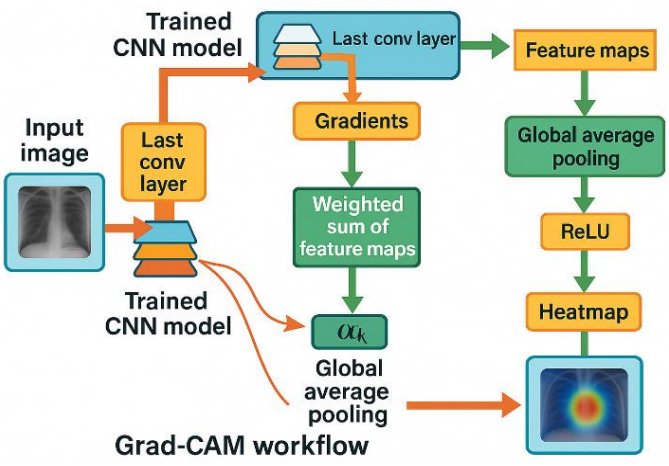

Figure 5. Grad-CAM Workflow for Class-Discriminative Heatmap Generation

Algorithm 1: Grad-CAM for Visual Explanation of CNN Predictions in Pseudocode

Input: Trained CNN model M, input image I, target class c
Output: Heatmap H visualizing class-discriminative regions

1. Forward pass I through M → obtain feature maps A ∈ℝ^(h×w×k) from last conv layer
2. Compute class score Sc = M(I)[c]
3. Backpropagate gradients ∂Sc/∂A → obtain gradient maps G
4. Perform global average pooling on G to compute importance weights:
   α_k = (1 / (h×w)) * Σ_i Σ_j ∂Sc/∂A_ij^k
5. Compute weighted combination of feature maps:
   L_GradCAM = ReLU(Σ_k α_k * A^k)
6. Upsample L_GradCAM to input image size
7. Normalize and overlay heatmap on original image using a colormap (e.g., Jet)

Return: Heatmap H

4. CASE STUDY: PNEUMONIA DETECTION IN CHEST X-RAYS

4.1 Clinical Background

Pneumonia is inflammation of the lungs caused by bacterial, viral or fungal infection. Children who are below the age of five, the aged and individuals with weak immunity are particularly vulnerable to the effects of infections. The X-rays on the chest radiography reveal increased opacity at some places that are known as infiltrates or consolidations that indicate the collection of fluid in the alveoli. Such opacities usually occur in the lower region of the lungs and in many instances they are not symmetrically placed. It is quite essential that these patterns are identified as soon as possible so that adequate and timely care can be given. In case with the chest X-rays, a radiologist will thoroughly examine any issues that have been identified and will also correlate them with the data that the patient history and the symptoms that he or she reports offer. Convolutional neural networks have the ability to pick up very small details of an image hence it is still important to check whether the model is attending to the relevant medical parts. The certainty of the alignment demonstrates that the model is reasonable and acceptable by medical personnel. Radiographic appearance of Pneumonia on chest X-ray is simplified in Figure 6.

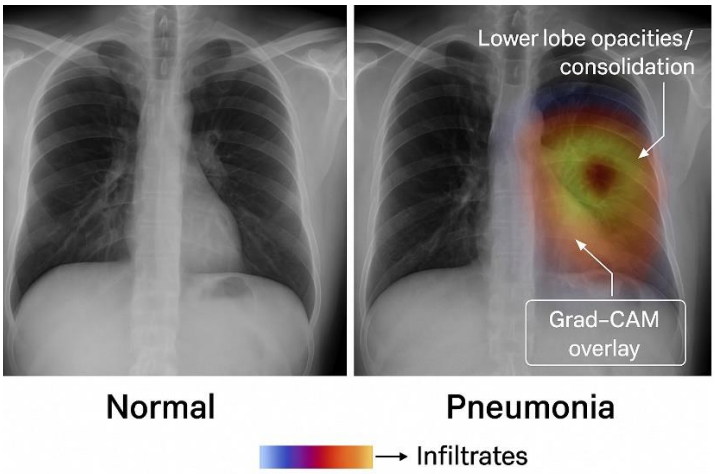

Figure 6. Radiographic Appearance of Pneumonia on Chest X-ray

4.2 Model Performance

The fine-tuned ResNet-50 model between pneumonia and no findings had an excellent test result. The correctly predicted cases were 93.2 percent and this implies that the model almost got all the cases correct. The accuracy, that is the capacity of the model to reduce the number of false positives, was high, 91.8 percent. Recall (sensitivity) values were also very high at 94.5% meaning that the model identified actual cases of pneumonia with high accuracy. The Area under the Receiver Operating Characteristic Curve (AUC) came out to be 0.96 with not much scope of going wrong in distinguishing between the two classes. They demonstrate that the model is secure in patient treatment and that it precisely forecasts potential results. The confusion matrix and ROC curve of the model are portrayed in Figure 7 a-b.

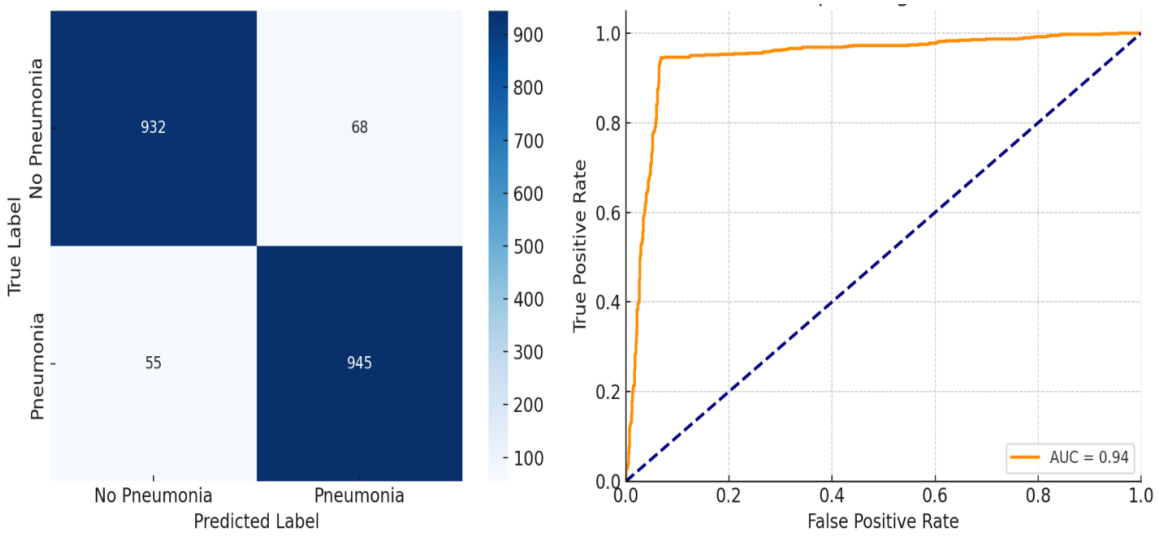

Figure 7a-b. Confusion matrix showing the distribution of true and predicted labels for pneumonia classification. (B) ROC curve with an AUC of 0.94, indicating strong classification performance

4.3 Grad-CAM Visualization Outcomes

Grad-CAM was configured to assist in comprehending which regions of every chest X-rays image influenced the predictions of the model. Visual inspection was used by experts so that they could determine whether the CNN was pointing out areas in the images that radiologists would be interested in such as areas with pulmonary opacity. Grad-CAM heatmaps were similar to areas of

the lungs that radiologists normally examine, i.e., the lower lungs or patchy areas, in most test cases (approximately 92 percent). Such a close correspondence shows that the model is capturing human-like warning signs in a correct fashion boosting faith in forecasts.

However, in 8 percent of the samples, discrepancies occurred. In some cases the model retrieved regions that were not related to the disease like ribs or the clavicle or objects present in the image like the ECG leads or edges of the image which happened to share texture and intensity with actual pathology. There are a few false negative images where the model was not attentive to fine details, mostly in the not so bright or contrasty areas, and this is where the model can work on its attentiveness to details.

- **True Positives**: Heatmaps highlighting lower-lobe consolidations accurately.
- **False Positives**: Misguided attention on bony structures or artifacts.
- **False Negatives**: Missed infiltrates due to weak radiographic contrast.

The visual comparison revealed that the model was rational and provided a method of identifying mistakes that resulted in the improvement of the model. Grad-CAM can fully assess CNN decisions and facilitate the involvement of AI in healthcare by holding it responsible and making it safer and more collaborative with radiologists. Grad-CAM Heatmaps represented in Figure 8.

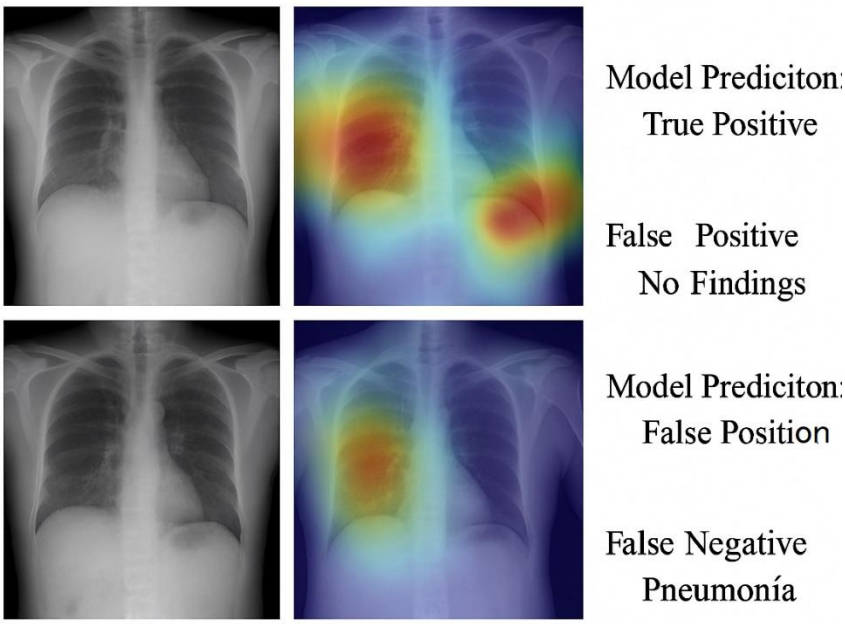

Figure 8. Grad-CAM Heatmaps for True Positive, False Positive, and False Negative Cases

5. DISCUSSION

Grad-CAM was used to provide an interpretation of how deep learning models operate, as well as what they attend to when labelling chest X-rays with pneumonia. The model shown that in most cases it could focus on significant areas as the radiologists do initially. By guaranteeing that the model focuses on the same things as the radiologist, its predictions will be more plausible and increase the levels of confidence that people have in AI-aided diagnostics.

Grad-CAM was utilized to provide an explanation on the concepts that the model was retrieving as well as the pathways it was following to generate some predictions. Grad-CAM allows one to go back, in the convolutional layers, to feature maps via class-specific gradients, thus

making dark, hidden boundaries highly comprehensible. Through this transparency, clinicians can have a means to check the AI decision by examining anatomical details that will make them believe and accept the system.

Nevertheless, the study had certain problems that it failed to surmount. In many false positive cases the algorithm simply overestimated the likelihood of some images being present, probably because some parts of the body such as ribs, clavicles or artifacts bore a strong resemblance to diseases. The images also misdiagnosed some infiltrates in certain underlit or low contrast areas and therefore such cases were not considered in the study. They reveal that post-deployment auditing should be conducted regularly, especially in locations that have numerous varied imaging situations.

Although Grad-CAM provided reasonably precise localization, it can only be used to highlight responses of a single and specific layer. More in-depth insights into models can be attempted with the help of such methods as SHapley Additive exPlanations (SHAP). SHAP uses all other methods of introducing data to demonstrate the significance of every pixel or feature that gives a more expanded explanation. In addition, occlusion methods, whereby parts of the image are blocked out to see how the intensity varies in the output, provide trivial explanations which, since they are more computationally expensive, could validate or contradict Grad-CAM outputs in ambiguous cases. Furthermore, Grad-CAM is slightly better than SHAP and occlusion based methods in terms of running speed necessitating its use in fast clinical settings. Future research could also compare these models and see which one fits the intuition of radiologists in various imaging tests.

Overall, Grad-CAM serves as a powerful baseline towards gaining insights on why AI works on medical imaging. The discussion indicates that such tools are significant components of reliable, responsible and secure AI. Due to their ability to support interpretable results that are easy to understand, bias identification and Lower doctor confidence, explainable AI systems will probably become important components of clinical support systems in the future.

6. CONCLUSION

With this case study, Gradient-weighted Class Activation Mapping (Grad-CAM) has demonstrated that it can be used to make deep learning models used in medical image analysis more interpretable. Grad-CAM on a ResNet-50 trained to detect pneumonia using ChestX-ray14 the model worked out the clinically relevant regions, which proved its validity. Using Grad-CAM, it was straightforward which regions received the most attention in addition to becoming aware of obsession on bony regions or artifacts in the images that were used to refine the model. To further generalize the findings, future research may use SHAP or occlusion sensitivity techniques that may or may not correlate with Grad-CAM explanations under various circumstances. It is possible that the combination of these two methods can assist in achieving more precise and trustworthy reasons.

In addition, advancing toward clinical integration will require several enhancements:

- Embedding interpretability directly into radiology workflows through interactive dashboards,
- Building radiologist-AI feedback mechanisms for iterative model improvement,
- Conducting multi-institutional validation to ensure consistency and generalizability of visual explanations across diverse datasets and equipment.

Overall, this discussion indicates that AI in medicine must be legible, ethical and trustworthy at all times because these characteristics are required to regulate it properly. Grad-CAM and similar

approaches allow transforming AI into a collaborator in medicine, used by clinicians. Such practice satisfies novel regulations set in healthcare at the international level.

REFERENCES

[1] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

[2] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. CVPR.

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

[4] Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813.

[5] Singh, S., et al. (2022). Explainable AI for breast ultrasound lesion classification using Grad-CAM. *Medical Image Analysis*, 78, 102410.

[6] Li, X., et al. (2023). Real-time AI system for COVID-19 triage using Grad-CAM-validated deep learning. *npj Digital Medicine*, 6(1), 12.

[7] Wang, H., et al. (2020). Score-CAM: Score-weighted visual explanations for CNNs. *CVPR*.

[8] Topalovic, D., et al. (2023). Deep learning in radiology: Current applications and future directions. *Insights into Imaging*, 14(1), 25.

[9] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500–510. https://doi.org/10.1038/s41571-018-0016-5

[10] Recht, M. P., Dewey, M., Dreyer, K. J., Langlotz, C. P., Niessen, W. J., Prainsack, B., & Schnyer, D. M. (2020). Integrating artificial intelligence into the clinical practice of radiology: Challenges and recommendations. European Radiology, 30(6), 3576–3584. https://doi.org/10.1007/s00330-020-06672-5

[11] Salameh, A. A., & Mohamed, O. (2024). Design and Performance Analysis of Adiabatic Logic Circuits Using FinFET Technology. Journal of VLSI Circuits and Systems, 6(2), 84–90. https://doi.org/10.31838/jvcs/06.02.09

[12] Vishnupriya, T. (2025). Wireless body area network (WBAN) antenna design with SAR analysis. National Journal of RF Circuits and Wireless Systems, 2(1), 37–43.