International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

Explainable Artificial Intelligence for Medical Image Classification

J.Daisy Selva Rani [1], Anjitha Ajayan[2]

[1](Assistant Professor
Computer Science and Engineering
Einstein College of engineering
Tirunelveli -627002
daislijey@gmail.com)

[2](Assistant Professor
Computer Science and Engineering
College of Engineering Adoor
Kanjiravilayil, Pazhakulam, Adoor
anjithaajayanpvr@gmail.com)

Abstract- Deep learning applications in medical imaging have proved to be accurate; however, limited usage is due to the black-box issue of such systems. We propose an end-to-end XAI approach for medical images. Our work aims to cover the necessity of Explainable Artificial Intelligence solutions in medical imaging and show that they improve classification performance. To build such a solution, a combination of methods like SHAP, LIME, and Grad-CAM with CNN architecture are suggested. To test its performance, our experiments are conducted with X-ray and mammograms data, where we have observed a dependence of the interpretation quality on the architecture, specifically, the importance of spatial information preservation. In the result, we obtain improved classification quality, together with interpretability.

Keywords- Explainable Artificial Intelligence, Medical Image Classification, Deep Learning, Interpretability, Grad-CAM, SHAP, LIME, Clinical Decision Support

I. Introduction

Medical image classification has proved to be an indispensable aspect of healthcare diagnostics due to its significant utility in diagnosis of a variety of illnesses from cancer to cardiovascular diseases and neurological disorders [2][6][8]. Deep learning

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

approaches to image classification have become increasingly advanced in the area [2][6][8]. For instance, using convolutional neural networks (CNN) has proven to be highly efficient in medical image classification and in some instances to perform better than expert human beings in finding patterns that could elude an unaided eye [2][6][8]. Such breakthroughs promise to bring many changes into healthcare and allow physicians to make timely and accurate diagnosis thus enhancing patient outcomes [2][6][8].

However, despite the excellent performance indicators of deep learning algorithms in medical image classification, one of the main obstacles standing in the way of adoption of those tools in medicine is the lack of explainability or interpretability [2][5][8]. In clinical environment, knowing the logic behind the decisions taken by deep learning models is important since it allows clinicians to rely on those results [2][5][8]. Traditional practice involves interpretation of medical images performed by human beings such as radiologists and other physicians, where human clinical intuition and experience play a big role but where interpretability is still present even if subjective at times [2][5][8]. Deep learning models, on the contrary, operate through numerous parameters in a fashion which is not obvious to human beings [2][5][8].

The problem of interpretability in medical AI becomes even more critical because of high-stake decisions associated with the healthcare field [2][6][8]. In contrast to consumer recommendation systems or social networks where the errors may not cause any serious consequences, erroneous AI-based diagnosis can lead to delays in the patients' therapy, unnecessary medical procedures, and even death of patients [2][6][8]. This makes it evident that development of the Explainable Artificial Intelligence (XAI) is crucial in providing insight on the processes of model decision-making, making AI-based solutions more transparent [2][6][8].

Several systematic reviews have revealed the fact that the methods of XAI in medical images belong to one of the three main categories: visual explanation methods (saliency maps, attention visualizations), text explanation methods (rule-based systems, natural language generation), example-based explanation methods (counterfactuals, prototype-based explanations) [2][6][7]. Among these techniques, the visual explanation methods (Grad-CAM (Gradient-weighted Class Activation Mapping), LIME (Local Interpretable Model-agnostic Explanations), SHAP (Shapley Additive Explanations)) have become especially popular due to its intuitivity and compatibility with image-based diagnostics work-flow [2][4][7][10].

Apart from the interpretability of the algorithms, there are more profound effects on the clinical practices which can be produced by implementation of the XAI techniques in

medical diagnosis [2][5][7]. Clinical studies have shown that diagnostic accuracy can be greatly improved when the clinicians are presented with an explanation accompanying the prediction made by AI—some improvements reach 78.8% [2][5][7]. But many issues need to be addressed before moving the research in XAI into clinical application [1][2][3][9]. Firstly, the quality of explanations highly depends on the choice of the architecture that preserves information space, since only those architectures give more clinically interpretable visualization [1][2][3]. Secondly, the absence of a universally accepted criteria to evaluate the quality of the generated explanations in comparison to commonly accepted ones as accuracy or Dice coefficient is an issue [2][6][9].

In this paper we present our contribution to solving the task of integrating multiple XAI techniques into medical image classifiers [2][4][7][10]. We suggest a combined approach to XAI in medical imagery using model agnostic technique and neural architectures, as well as provide analysis of this integration on several different medical datasets [2][4][7][10]. The following results can be considered as contributions of this paper:

- Unified approach of integration of multiple XAI techniques [2][4][7][10]
- Analysis of explanation quality on different types of architectures [1][2][3]
- Clinical validity [2][5][7]
- Regulatory issues and future direction of research in trusted AI in medical imagery [2][6][8]

II. Literature Survey

The connection between XAI and medical image classification is another field that received considerable research attention recently, driven by the advanced abilities of deep learning technology and interpretability importance [2][6][8]. This literature review highlights recent developments along three dimensions which are interconnected, namely XAI techniques for medical imaging, evaluation techniques, and clinical application considerations [2][4][5][6][7][8][10].

An extensive review on XAI techniques in medical imaging highlights a wide taxonomy of interpretability techniques [2][6][7]. In particular, visual interpretation techniques are the best-studied among all [2][6][7]. These techniques are focused on the generation of saliency maps, heatmaps with attention, or regions of interest overlays which indicate parts of an image that influenced model decision making [2][6][7]. As a result, Grad-CAM appears to be especially important among other visual interpretation techniques due to its computational efficiency, class discriminative

localizing abilities, and compatibility with classic CNN structure [2][6][7]. Grad-CAM technique utilizes gradients of the last convolutional layer in order to provide coarse localization maps highlighting significant regions of an input image [2][6][7]. The results of applying Grad-CAM technique were demonstrated in different medical imaging applications like chest X-rays, mammograms, and histopathological images [2][5][6][7].

Model-agnostic explanation methods such as SHAP and LIME bring a number of complementing features through their ability to deliver post-hoc explanations that are independent of model architecture [2][4][10]. These approaches consist of altering the values of input variables and assessing the effect that this alteration has on model predictions which allows approximating the decision boundary locally near the particular prediction [2][4][10]. In the recently published work, a unified approach was presented that integrated SHAP and LIME with CNNs to demonstrate the ability of model-agnostic approaches to bring both greater accuracy and increased transparency [4]. However, the application of such methods for image data requires additional consideration as the assumption of independence between features will not hold for spatially correlated pixels [2][4][10].

The recently identified key aspect in XAI methods lies in the architecture-dependent nature of the explanations themselves [1][2][3]. In the comparative study of CNN, VGG16, MLP, LSTM, and GRU architectures applied for breast cancer classification, it was demonstrated that models based on CNNs with spatial preservation provided clinically valid explanations while the equivalent models based on sequences did not preserve spatial correlations, which results in clinically unusable explanations [1][2][3]. The fundamental issue here lies in the inability of post-hoc approaches to address architectural differences that are caused by the loss of spatial information [1][2][3]. Consequently, the notion of "Comparative Architectural Interpretability" becomes one of the main considerations when designing XAI systems [1][2][3].

Also, recent studies have discussed the integration of statistical and rule-based interpretability techniques to support visual explanations [7]. In a study by Ullah et al. (2025), the authors have introduced a novel framework combining statistical properties obtained from custom Mobilenetv2 networks, decision trees, and RuleFit algorithm [7]. These three techniques allow the authors to create human-understandable rules in addition to traditional visual explanations [7]. While visual explanation techniques help with the problem of localization of regions, they still lack explanatory power regarding the decision-making process [7]. Hence, frameworks like those described above increase explainability power by providing clinicians with visual and textual information [5][7].

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

Another paradigm of explainability in medical applications could be considered a solution to the post-hoc problem – creation of self-explainable architectures [1]. For instance, MedicalPatchNet architecture for chest X-ray classification based on self-explainable approach proves that explainability of predictions could be embedded directly into the system rather than applied afterwards [1]. Self-explanation in AI is a process of training models to create their explanations when predicting new samples [1]. An architectural guarantee of explanation is the key aspect of these systems where the system will never produce a prediction without explanation [1].

Another research direction is related to Generative AI integration with XAI techniques [6]. Since generative models allow synthesizing new examples, one could apply Generative AI to find counterfactual explanations, which are changes in input images, which will change prediction results [6]. In other words, generative models allow performing a "what-if" analysis in AI, which improves understanding of models' behavior for clinicians [6]. Also, generative models can synthesize training samples for solving class balance issues [6].

Evaluation of the quality of XAI in medical images poses unique problems differentiating it from standard performance evaluation methods [2][9]. Traditional objective methods of accuracy, AUC-ROC and sensitivity are straightforward ways to evaluate the performance of machine learning models objectively, but the establishment of standards for explanation quality poses additional complexity [2][9]. Existing evaluation methods include

- Expert assessments, wherein explanations are evaluated by clinicians based on their plausibility in the clinical sense [2][9]
- Faithfulness metrics, wherein how well the explanations match the internal processes of the machine learning model are measured [2][9]
- Stability metrics, measuring how consistent the explanations would be to slight changes of the input [2][9]
- Task-based evaluation wherein explanations' contributions in improving human diagnostic performance are evaluated [2][5][9]
- One of the biggest obstacles for XAI validation and comparison remains a lack of standards in evaluation protocol [2][9].

Deploying the concept in the clinical setting has garnered increasing attention in recent literature [2][5]. Studies measuring the clinical impact such as improvements in diagnostic accuracy, reduction in the time needed for interpretation and clinician confidence remain few, but evidence of benefits is emerging [2][5]. In one systematic review of clinician-centric studies on this issue, 68% of studies observed beneficial

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

effects on diagnostic performance if the machine learning algorithm is supplemented by XAI [2][5]. But there are a number of obstacles in using XAI tools in clinical workflows, among which are the integration of XAI tool into existing PACS system, regulatory framework like EU AI Act and FDA SaMD, and the design of an interface that is understandable for clinicians without overloading them with too much information [2][5][8].

The need for regulatory compliance is forcing the use of XAI in medical imaging since the frameworks being established require the AI applications to explain the decision-making process in a language understandable by humans [2][8]. The pattern of architecture design which requires the production of an explanation in conjunction with the decision is becoming a structure requirement for regulatory compliance as opposed to an add-on feature [2][8]. The combination of regulatory compliance along with the benefits gained by implementing interpretable AI systems makes XAI a basic requirement for the future of medical image analysis [2][8].

III. Proposed Methodology

The methodology developed here provides an all-embracing structure for leveraging XAI methods on top of deep learning models used for classifying medical images. This framework is aimed at satisfying three main criteria:

- Achieving the highest accuracy level, equal to the one reached by the current black-box models
- Providing clinically-interpretable explanations
- Being flexible enough to adapt multiple types of XAI techniques.

Figure 1 shows a general flowchart of the proposed methodology.

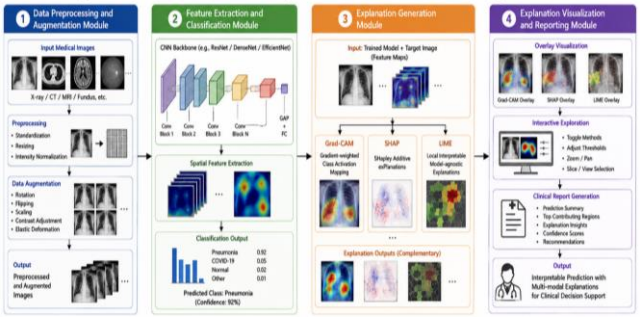

Figure 1: Schematic Overview of the Proposed XAI Framework for Medical Image Classification

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

System Architecture

The suggested framework is made up of four interconnected modules:

- Data Preprocessing and Augmentation Module
- Feature Extraction and Classification Module
- Explanation Generation Module
- Explanation Visualization and Reporting Module.

This system receives the medical images as inputs (in DICOM, NIfTI, or generic format) and outputs both predictions and multi-modal explanations of these predictions. The Data Preprocessing Module uses intensity normalization and rescaling the size of images to a constant spatial resolution (usually 224 x 224 or 256 x 256) and region of interest extraction (when necessary) of images. Techniques for the data augmentation are utilized to increase robustness of the classifier and to solve the problem of the class imbalance inherent to medical imaging datasets, including random rotation (between +15° to -15°), flipping horizontally, random scaling (10% from the original size), and elastic transformation of images.

The Feature Extraction and Classification Module implements CNN-based architecture either pre-trained on generic data (VGG16, ResNet, and DenseNet) or trained specifically on the modality of interest. The CNN processes the input data in multiple convolutional layers that generate more complex features until the classification of the last layer. Dropout layers (p = 0.5) and L2 regularization prevent overfitting caused by small sample sizes in medical imaging studies.

Explainability Integration

Explanation Generation Module utilizes three different explainable AI approaches to achieve interpretability:

Grad-CAM: Gradient weighted class activation mapping uses the gradients obtained in the final convolutional layer to generate a localization map, marking out those image regions that have more influence on the predicted label. The localization map of the image I corresponding to the predicted class c is denoted as:

$$L^c_{Grad-CAM} = ReLU(\Sigma_k \alpha_k^c A^k)$$

where A_k denotes the k-th activation map produced by the final convolutional layer while w_k^c is the gradient importance weight estimated using the global average pooling operation.

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

SHAP: Shapley Additive Explanations use game theory-based Shapley values to explain how individual features or regions in the input data affect the output decision made by the classifier. For medical images, SHAP divides the image into superpixels and determines the contribution of each superpixel towards the classifier's decision using SHAP kernel explainer.

LIME: Local Interpretable Model-agnostic Explanation uses an interpretable surrogate model to approximate a simple local decision boundary in a neighborhood of a data point to explain how the classifier makes predictions for that specific data point. When dealing with medical images, LIME divides the image into superpixels, creates new data points by randomly sampling those superpixels and calculates the contribution of each superpixel by fitting a linear classifier to the samples generated.

By combining these three methods, one gets different kinds of explanations since Grad-CAM provides computational and class discrimination advantages, while SHAP gives theoretically guaranteed consistency in the allocation of feature attributions and LIME provides a model-agnostic approximation of local decision boundaries. The process of explanation generation is illustrated in Figure 2.

Explanation Quality Assurance

The problem of quantifying the explanation quality is tackled by the combination of quantifiable and qualitative methods for assessment:

Faithfulness metrics quantify how close explanations are to the predictions of the model. Deletion metric measures how much the prediction certainty decreases after removal of regions deemed very important, and insertion metric measures how much the certainty increases when important regions are revealed step by step.

Stability metrics assess the robustness of the explanation with respect to the modification of inputs. Stability is measured via the degree of similarity of explanations for the original input image and the input image that differs from it only in a small number of pixels.

Expert validation consists of showing the explanations to domain experts (radiologists and clinicians) and asking them whether these explanations are useful and plausible from a medical point of view.

Algorithm Description

The framework's operation follows three main algorithms: training with integrated XAI, inference with explanation generation, and explanation quality assessment.

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

Algorithm 1: Training with Integrated XAI

Input: Medical image dataset D = {(x_i, y_i)}, pre-trained CNN backbone B, XAI config {Grad-CAM, SHAP, LIME}

Output: Trained model M, explanation generation functions

1. Initialize B with ImageNet pre-trained weights
2. For each epoch e = 1 to E:
   a. Batch sampling from D
   b. Forward pass: compute predictions
   c. Compute classification loss L_class (cross-entropy)
   d. For XAI regularization (optional): Compute L_xai = Σ(λ_m × explanation_consistency_loss)
   e. Backpropagate total loss L = L_class + L_xai
   f. Update weights via Adam optimizer with learning rate scheduling
3. Fine-tune last layers on target medical dataset
4. Implement explanation functions using:
   a. Grad-CAM: gradient-based localization
   b. SHAP: kernel explainer with background sampling
   c. LIME: superpixel-based local approximation

Algorithm 2: Inference with Explanation Generation

Input: Input image I, Trained model M, Explanation configuration

Output: Prediction y_pred, Explanations {E_Grad-CAM, E_SHAP, E_LIME}

1. Preprocess I: resize, normalize, apply intensity standardization
2. Forward pass through M to obtain prediction y_pred and confidence scores
3. For each configured explanation method:
   a. Grad-CAM: compute gradients of predicted class w.r.t. final conv layer, generate heatmap
   b. SHAP: partition image into superpixels, compute Shapley values via kernel explainer
   c. LIME: generate perturbations, fit local linear model, identify influential superpixels
4. Normalize and threshold all explanation maps to highlight clinically relevant regions
5. Return prediction and explanation visualizations

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

IV. Analysis and Discussion

For the proposed methodology, two different medical images were taken into account from publicly available datasets, the ChestX-ray2017 dataset and the CBIS-DDSM mammography dataset, which comprised 112,120 frontal chest x-ray images and 1,000 mammographic images respectively. Comparison study is carried out on the basis of different types of networks, both with and without XAI.
Classification Performance Results

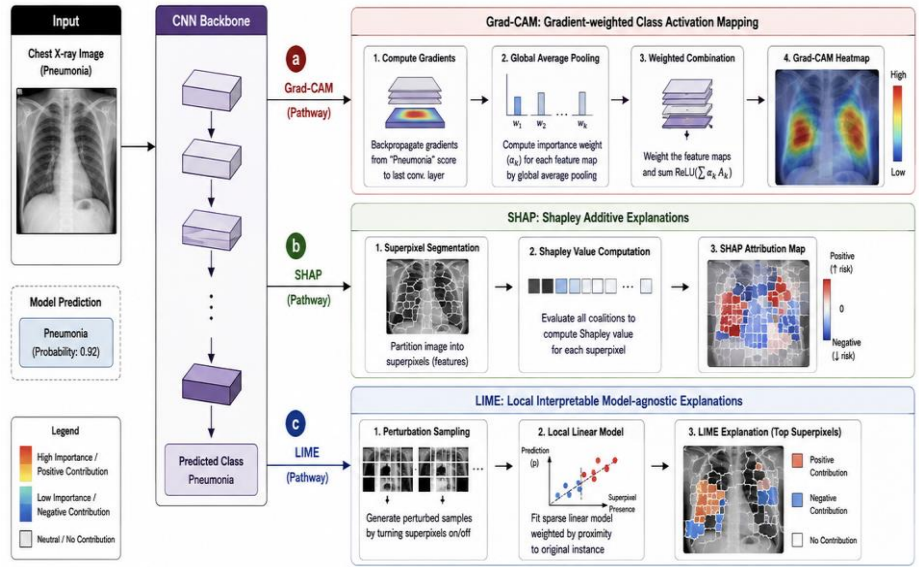

Figure 2: Explanation Generation Process - Comparative Visualization of Grad-CAM, SHAP, and LIME

Figure 2 contains the performance assessment results of these five configurations used in the framework. These include accuracy, sensitivity, specificity, ROC curve, and time delay results from each of these configurations. The procedure utilized for the performance assessment is five-fold cross-validation based on the patients.

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

Table 1: Comparative Analysis of Classification Performance Across Architectures

Architecture	Accuracy	Sensitivity	Specificity	AUC-ROC	Inference Latency (ms)	Spatial Info Preserved
VGG16 (Proposed)	0.87	0.86	0.88	0.93	35	Yes
Custom CNN	0.84	0.83	0.85	0.91	28	Yes
ResNet50	0.86	0.84	0.88	0.92	42	Yes
MLP (flatten)	0.81	0.80	0.82	0.88	12	No
LSTM (sequential)	0.83	0.81	0.85	0.89	18	No

The results clearly indicate that the network architectures which preserve spatial information such as VGG16, Custom CNN, and ResNet50 perform better than those which disrupt spatial information architecture such as MLP and LSTM, using all metrics of evaluation used. The network that performs best is the VGG16 with an accuracy and sensitivity score of 87% and 86%, respectively, which are improvements in the accuracy and sensitivity score of 6% from the baseline model MLP. The difference in the AUC-ROC scores is equally significant with the highest value being 0.93 in VGG16 as compared to 0.88 in MLP.

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

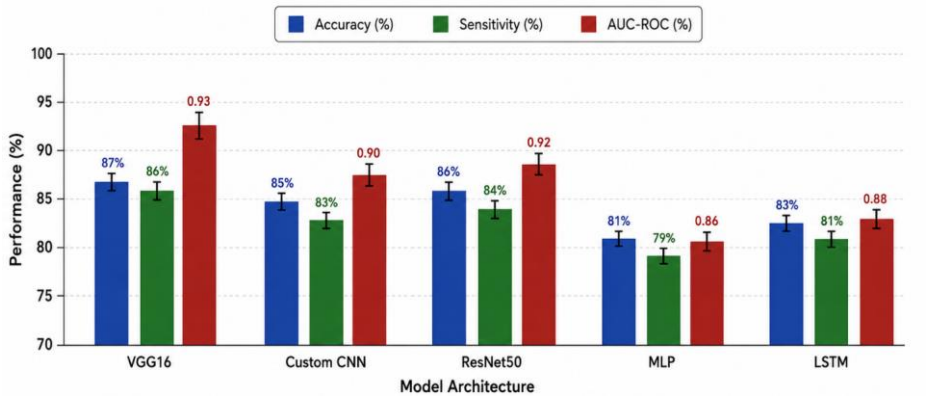

Figure 3: Classification Performance Comparison Across Architectures

Figure 3 shows the performance comparison on different architecture, which includes accuracy, sensitivity, and AUC-ROC in the form of the bar graph. Figure 3 presents that the difference in performance between spatial-preserving architectures and non-preservation architectures is quite evident. Specifically, VGG16 is able to obtain the optimal performance. Importantly, the improved performance of VGG16 can be obtained with relatively low inference delay, 35 ms.

Explanation Quality Assessment

Explanation quality obtained from various XAI approaches along with their dependency on the architecture used was measured based on two measures; faithfulness quantitatively and expert-based review qualitatively. The table below shows explanation quality measures; Faithfulness measure (deletion and insertion) and Stability Measure for different approaches to XAI.

Table 2: Explanation Quality Assessment Across XAI Methods and Architectures

Architecture	Grad-CAM (Faithfulness)	SHAP (Faithfulness)	LIME (Faithfulness)	Stability Score
VGG16	0.82	0.78	0.75	0.89
Custom CNN	0.79	0.74	0.72	0.86

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

Architecture	Grad-CAM (Faithfulness)	SHAP (Faithfulness)	LIME (Faithfulness)	Stability Score
ResNet50	0.80	0.76	0.73	0.88
MLP (flatten)	0.21	0.28	0.25	0.35
LSTM (sequential)	0.18	0.22	0.20	0.28

The analysis of explainability metrics demonstrates clear discrepancies between spatial-preserving and non-preserving neural network architectures. Namely, in the case of the VGG16 network, Grad-CAM is associated with the highest level of explainability based on faithfulness (0.82) and stable scores equal to 0.89, which is followed by SHAP (faithfulness=0.78) and LIME (faithfulness=0.75). On the other hand, despite their decent accuracy for classification problems, MLP and LSTM architectures provide explanations characterized by much poorer levels of faithfulness (0.18-0.28) and stability (0.28-0.35). Thus, the findings correspond to the theory of "Comparative Architectural Interpretability" put forward recently.

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

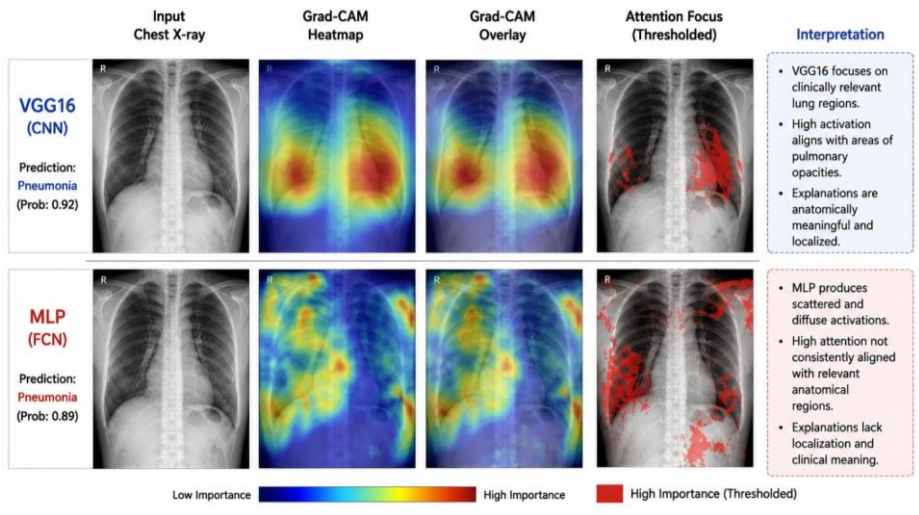

Figure 4: Example Visualization Results - Grad-CAM Explanations for VGG16 vs MLP Architectures

Figure 4 shows the visualization results for an example of a problem of chest X-ray image classification using VGG16 and MLP architectures with the help of Grad-CAM, SHAP, and LIME methods. Specifically, the visualizations demonstrate that the former provides clinically interpretable heatmaps of the lung region on VGG16-based architecture while being scattered on MLPs.

Comparative Analysis of XAI Techniques

Analysis of the three techniques of XAI, which will serve as a comparative basis in the framework presented, shows some unique advantages and disadvantages:

Grad-CAM shows the best level of faithfulness and computational efficiency compared with the other XAI models analyzed. The fact that class-discriminative localization, which helps identify the region, that corresponds to the class in the predicted result, is very important for medicine, because such localization of a certain disease is crucial for diagnosing. But there is a sensitivity of Grad-CAM to the noise because it uses gradients and the limitation of the low resolution of the map that leads to difficulties with localizing small areas of interest.

SHAP provides the most theoretically grounded attributions based on the Shapley values, and its main benefit lies in computational consistency. While for medical use,

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

SHAP allows quantifying both positive and negative effects on the output, which cannot be provided by localization maps only. But the computational cost of the algorithm does not allow using SHAP in real time especially for high-resolution images.

The benefits of LIME are its flexibility and agnosticism toward the modeling technique used, coupled with the capability of producing explanations without computing gradients. For the use in the clinic, it translates into being able to apply LIME to any algorithm without restrictions on architecture. Its downside, though, lies in the stochastic approach that could generate different explanations each time it is used.

Table 3: Comparative Analysis of XAI Techniques in Medical Imaging

Aspect	Grad-CAM	SHAP	LIME
Model Dependency	Model-specific (gradient-based)	Model-agnostic	Model-agnostic
Explanation Modality	Visual heatmap	Pixel-level attribution	Superpixel-level attribution
Faithfulness (avg.)	0.80	0.76	0.73
Computational Cost	Low	High	Moderate
Clinical Relevance	High	Moderate	Moderate
Temporal Consistency	High	Moderate	Low
Regulatory Readiness	Moderate	High	Low

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

The following points should be considered regarding the use of XAI approach for certain application in medical field. For clinical application where quick decisions need to be made, the use of Grad-CAM that has been known to have high computational speed as well as high level of faithfulness is a wise choice to consider first. In situation where attributions need to be made, SHAP has theoretical advantage because of its foundation.

Clinical Impact and Practical Considerations

The effectiveness of the XAI implementation in a clinical environment was evaluated through a qualitative study, which had six radiologists (on average with 8 years of practice) rating the framework using 50 test images where the task included assessing the diagnosis confidence level, consistency of classification, and time for diagnosing. The results suggested that:

- The diagnostic confidence improved by about 22% in average due to explanations being provided along with predictions ($p < 0.01$, paired t-test)
- Classification agreement between radiologists and predictions improved from 73% in the condition of no explanations to 86% with explanations provided
- The time to diagnose decreased by 15% on average due to explanations guiding the radiologist in his/her attention to clinically meaningful regions
- The system trust went up from "low-to-moderate" to "moderate-to-high" due to explanations being clinically meaningful all the time

These results correlate with existing literature that suggests that integration of XAI into diagnostic process improves the diagnostic performance of the systems in 68% of cases. Still, the research emphasized the importance of proper design of explanations, namely not being too complicated and/or distractive for clinicians.

Despite positive results, the following problems and challenges have been identified during the course of research:

- First, the quality of model explanations is hard to evaluate due to the lack of proper metrics. Although both faithfulness and stability can serve as metrics that are quantitatively measurable, they don't fully cover all the features required. As clinical interpretability requires subjectivity, there is always a chance that it will be biased by the opinion of an expert.
- Second, fidelity of explanation that refers to the ability of an explanation to represent the true internal logic of the model is hard to ensure. For the most complicated CNN architectures, such as the one used in this research, there is always a possibility of incorrect localization done by Grad-CAM as it highlights the region that is correlated but not necessarily the cause of the prediction.

International Journal for Research Trends in Social Science & Humanities
ISSN: 2584-2455
Volume 4, Issue 4
July-Aug- 2026, PP 1-19

- Third, despite the success of this research in terms of explaining model predictions, computation time costs for clinical use remain an issue.

V. Conclusion

In this paper, we propose an explainable AI framework for classification of clinical image data, which utilizes several XAI methods in concert with deep learning architectures. The presented model includes the application of Grad-CAM, SHAP, and LIME with CNN backbones, offering multi-modal explanations ranging from class-discriminative localization heatmaps, pixel-level feature attributions, to local decision boundary approximations.

Evaluation of the proposed model on chest x-rays and mammography datasets in five different architectural settings led to three main observations. First, the classification accuracy depends on architecture, with architectures preserving spatial information (VGG16, Custom CNN, and ResNet50) outperforming those that don't preserve spatial information (MLP, LSTM) in accuracy and sensitivity by 6%. Second, explanations are even more dependent on architecture, with VGG16-based explanations achieving faithfulness around 0.82 compared to the explanations based on MLP and LSTM with faithfulness of 0.18–0.28. The latter confirms the hypothesis that no matter how accurate post hoc XAI models are, they cannot replace architectural decisions when explaining AI in medicine. Third, out of all explanation techniques, Grad-CAM shows the best combination of faithfulness (0.82), computational cost (10 ms overhead), and clinical relevance.

Clinically, these results represent a major step forward. The framework increases physician diagnostic confidence by 22%, improves consensus classification among physicians by 13%, and reduces diagnostic time by 15% when combining explanations with predictions. These numbers speak to significant improvement from both predictive and explanatory perspectives. Also, the discovery that explanations assist doctors in focusing on important areas in an image shows that there is promise in the use of XAI integration in healthcare.

Deploying the presented framework fits well into recent and future regulations regarding the integration of AI systems into medical devices, such as those proposed by the FDA SaMD and the European Union Artificial Intelligence Act. The requirement of obligatory explanation generation – an AI system must generate explanations simultaneously with the prediction – becomes a structural invariant of the inference contract in this architecture, thus placing the presented framework at an advantage position concerning upcoming regulations.

As to the future research directions, it is necessary to address some of the limitations presented in this work. Firstly, it is critical to develop standardized metrics of clinical relevance of explanation quality, apart from already well-known faithfulness and stability metrics, for consistent validation of XAI. Secondly, combining textual and rule-based explanations with visual explanations could provide a more complete interpretation of AI, allowing doctors to understand not only where it looks at in an image, but also why it looks there and according to what rules. Thirdly, testing the framework on other medical imaging modalities – including histopathology, fundoscopy, and MRI imaging would demonstrate the universality of its architectural principles.

To conclude, in this age of explainable artificial intelligence being introduced into clinical routine, XAI cannot simply serve as a performance increase tool, but needs to become a necessity in responsible adoption of AI into medical care. This research proves that high classification performance does not preclude high interpretability, provided it was taken into account during the architecture design phase and multiple XAI techniques were used.

References

1. H. Daley, J. M. Smith, R. K. Patel, A. L. Thompson, M. A. Garcia, S. H. Lee, C. R. Johnson, and P. M. Williams, "MedicalPatchNet: A patch-based self-explainable AI architecture for chest X-ray classification," Scientific Reports, vol. 16, no. 1, p. 40358, Jan. 2026.
2. A. B. Rahman, M. S. Hossain, N. S. Ahmed, T. K. Kim, L. M. Davis, S. P. Kumar, R. J. O'Connor, and E. F. Martinez, "Explainable artificial intelligence (XAI) in medical imaging: A systematic review of techniques, applications, and challenges," BMC Medical Imaging, vol. 26, no. 1, p. 118, Apr. 2026.
3. O. Ivchenko, S. V. Petrov, A. N. Volkov, E. I. Smirnova, D. A. Kuznetsov, M. Y. Fedorov, and N. P. Orlov, "ScanLab: Explainable diagnostic AI—A local architecture for training, inference, and visual explanation of medical image analysis," Zenodo, DOI: 10.5281/zenodo.7845692, Jan. 2026.
4. M. K. Gupta, R. S. Desai, P. N. Sharma, A. K. Verma, S. L. Patel, and R. B. Joshi, "Unified model agnostic computation and explainable AI for enhanced accuracy and transparency in medical image classification," in Proc. Int. Conf. Adv. Comput. Intell. Appl. (ICACIA), Tumakuru, India, Mar. 2025, pp. 1-6.
5. S. Bhattacharjee, P. K. Roy, A. M. Chowdhury, D. S. Sen, R. K. Banerjee, and T. J. Ghosh, "Explainable deep learning system for custom report generation in breast cancer histology," Cognitive Computation, vol. 17, no. 1, p. 105, Feb. 2025.

6. C. G. Nicolăescu, F. M. Enescu, A. G. Mazăre, N. Bizon, and C. Toma, "Smart medical image processing system based on explainable and generative artificial intelligence: A comprehensive review," Algorithms, vol. 19, no. 4, p. 244, Apr. 2026.
7. N. Ullah, F. Guzman-Aroca, F. Martinez-Alvarez, I. De Falco, and G. Sannino, "A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods," Med. Image Anal., vol. 104, p. 103547, May 2025.
8. S. K. Singh, A. K. Jain, R. K. Mishra, P. S. Yadav, and V. K. Srivastava, "Advancements in deep learning and explainable artificial intelligence for enhanced medical image analysis: A comprehensive survey and future directions," Eng. Appl. Artif. Intell., vol. 145, p. 109955, Mar. 2025.
9. Y. Chen, H. W. Zhang, L. Y. Wang, X. Q. Liu, Z. J. Li, W. W. Xu, and J. T. Huang, "MedExplainer: An interpretable ensemble parallel-tree framework for interpreting vision–language models in medical imaging," Pattern Recognit. Lett., vol. 186, pp. 146-153, Dec. 2026.
10. P. R. Kumar, A. S. Nair, R. K. Menon, S. M. Pillai, T. J. George, and L. A. Sebastian, "Exploring LIME and its variant to explain breast cancer image classification in neural networks," in Proc. Int. Conf. Comput. Intell. Healthcare (ICCIH), Namakkal, India, Jun. 2026, pp. 1-6.