

XAI-Driven Mammogram Classification: A Comparative Study of SHAP and LIME in Deep Learning Models

Dr Sharada Guptha M N^{1*}, Chaitra M S², Shilpa M³, Divyashree B P⁴, Dr. Babitha MN⁵

^{1*} Associate Professor, Department of Electronics and Communication Engineering, Sri Siddhartha Academy of Higher Education, Tumakuru, India

² Selection Grade lecturer, Electronics and Communication Department, Government Polytechnic, Malavalli, Mandya

³ Selection Grade lecturer, Electronics and Communication Department, Government Polytechnic, Malavalli, Mandya

⁴ Assistant professor, Debarment of Electrical and Electronics Engineering, SJCE, JSS Science and Technology University, Mysuru, India

⁵ Associate Professor, Department of Electronics and Communication Engineering, Sri Siddhartha Academy of Higher Education, Tumakuru, India

Corresponding Author: Dr Sharada Guptha M N

Abstract

Breast cancer remains one of the leading causes of mortality among women worldwide, where early and accurate diagnosis plays a crucial role in improving survival rates. Mammography is the most widely used screening technique; however, its interpretation is often challenging due to subtle variations in tissue structures and high inter-observer variability among radiologists. In recent years, deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated strong performance in automated mammogram classification. Despite their effectiveness, these models lack transparency, limiting their adoption in clinical decision-making due to their black-box nature. This study proposes an Explainable Artificial Intelligence (XAI)-driven framework for mammogram classification, focusing on a comparative analysis of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). A ResNet-based CNN model is trained on the CBIS-DDSM dataset to classify mammograms into benign and malignant categories. SHAP and LIME are integrated to provide both global and local interpretability of model predictions. Experimental results indicate that while both methods enhance model transparency, SHAP provides more stable and consistent feature attributions across samples, whereas LIME offers faster and more localized explanations. Quantitative evaluation using fidelity and stability metrics, along with qualitative visualization analysis, highlights key differences between the two approaches. The findings suggest that SHAP is more suitable for clinical validation and auditing, while LIME is effective for rapid case-based interpretation. This work contributes toward improving trust and reliability in AI-assisted breast cancer diagnosis.

Keywords: Explainable Artificial Intelligence (XAI), Mammogram Classification, Deep Learning, SHAP and LIME, Breast Cancer Detection

1. Introduction

Breast cancer continues to be a major global health concern and is among the most frequently diagnosed cancers in women. Early detection significantly improves treatment outcomes and survival rates, making screening programs essential in modern healthcare systems. Mammography is widely regarded as the most effective imaging modality for early breast cancer detection, as it can reveal abnormalities such as masses and microcalcifications before clinical symptoms appear. However, interpreting mammograms is a complex and time-consuming task that requires considerable expertise. Subtle variations in tissue density, overlapping structures, and noise often make diagnosis difficult, leading to both false positives and false negatives (Litjens et al., 2017). With the rapid advancement of artificial intelligence, deep learning techniques have emerged as powerful tools in medical image analysis. Convolutional Neural Networks (CNNs), in particular, have demonstrated exceptional performance in tasks

such as classification, segmentation, and detection. These models automatically learn hierarchical representations of image features, enabling them to capture intricate patterns that may not be visible to the human eye. Architectures such as ResNet and DenseNet have been successfully applied to mammogram classification, often achieving performance comparable to or exceeding that of experienced radiologists (He et al., 2016; Esteva et al., 2017). Furthermore, the use of transfer learning has enabled the adaptation of pre-trained models to medical datasets, which are often limited in size.

Despite these advancements, a critical limitation of deep learning models is their lack of interpretability. CNNs operate as complex nonlinear systems with millions of parameters, making it difficult to understand how specific predictions are generated. This "black-box" nature poses a significant barrier to their adoption in clinical environments, where transparency and accountability are essential. Clinicians must be able to trust the model's decisions, especially in high-stakes scenarios such as cancer diagnosis. Without clear explanations, even highly accurate models may be rejected due to concerns about reliability and ethical implications (Samek et al., 2019). Explainable Artificial Intelligence (XAI) has emerged as a promising approach to address this challenge by providing insights into model behaviour and decision-making processes. XAI techniques aim to make machine learning models more transparent, interpretable, and trustworthy. In the context of medical imaging, these methods can highlight regions of interest in an image that contribute most to a prediction, thereby assisting clinicians in validating model outputs. Various XAI methods have been proposed, including saliency maps, Grad-CAM, and attention mechanisms. However, many of these approaches lack theoretical grounding or consistency, limiting their effectiveness in critical applications (Tjoa & Guan, 2020).

Among the most widely used and theoretically grounded XAI methods are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). SHAP is based on cooperative game theory and computes feature importance by considering the contribution of each feature across all possible combinations. This method provides both global and local interpretability, making it suitable for understanding overall model behavior as well as individual predictions (Lundberg & Lee, 2017). On the other hand, LIME focuses on local interpretability by approximating the model's behavior around a specific instance using a simpler, interpretable surrogate model. It perturbs the input data and observes the resulting changes in predictions to identify important features (Ribeiro et al., 2016). Although both SHAP and LIME have been applied in various domains, including healthcare, there is limited work that systematically compares their performance in the context of mammogram classification. Existing studies often focus on a single method or lack comprehensive evaluation metrics that assess explanation quality, stability, and clinical relevance. This gap highlights the need for a detailed comparative study to determine which method is more suitable for real-world medical applications.

This research aims to address this gap by integrating SHAP and LIME into a CNN-based mammogram classification framework and conducting a thorough comparative analysis. The study is guided by the following research questions:

1. How do SHAP and LIME differ in explaining CNN-based mammogram classification outcomes?
2. Which method provides more reliable, stable, and clinically meaningful explanations?
3. What are the trade-offs between interpretability, computational efficiency, and consistency?

The key contributions of this work are as follows. First, it presents a unified deep learning framework that incorporates explainability into mammogram classification. Second, it provides a systematic comparison of SHAP and LIME using both quantitative metrics and

qualitative visual analysis. Third, it evaluates explanation stability and fidelity, which are critical for clinical trust. Finally, the study offers practical insights into the applicability of XAI techniques in real-world healthcare settings.

By bridging the gap between predictive performance and interpretability, this research contributes toward the development of trustworthy AI systems that can support clinicians in making informed diagnostic decisions.

2. Related Work

The application of artificial intelligence in medical imaging has evolved significantly over the past decade, with deep learning emerging as a dominant approach for disease detection and classification. In the domain of breast cancer diagnosis, early methods relied heavily on traditional machine learning algorithms that utilized handcrafted features such as texture descriptors, edge information, and morphological characteristics. While these approaches provided moderate success, their performance was often limited by the quality and representativeness of manually engineered features (Dhungel et al., 2017). The introduction of Convolutional Neural Networks (CNNs) marked a transformative shift in mammogram analysis. CNNs are capable of automatically learning hierarchical feature representations directly from raw image data, eliminating the need for manual feature extraction. Architectures such as AlexNet, VGGNet, and ResNet have been widely adopted in medical imaging tasks due to their strong generalization capabilities (Krizhevsky et al., 2012; He et al., 2016). In mammography, CNN-based models have demonstrated high accuracy in classifying lesions as benign or malignant, often outperforming conventional methods. Additionally, transfer learning has been extensively employed to address the challenge of limited medical datasets by leveraging knowledge from large-scale datasets such as ImageNet (Yosinski et al., 2014).

Despite these advancements, the lack of interpretability in deep learning models remains a major concern. Medical decision-making requires transparency and justification, particularly in high-risk domains such as oncology. To address this issue, researchers have explored various Explainable Artificial Intelligence (XAI) techniques. Early approaches include saliency maps, which highlight pixel-level importance based on gradient information, and Class Activation Mapping (CAM) techniques such as Grad-CAM, which provide coarse localization of important regions in an image (Selvaraju et al., 2017). While these methods offer intuitive visualizations, they often suffer from instability and lack a solid theoretical foundation, limiting their reliability in clinical applications (Samek et al., 2019). More recent research has focused on model-agnostic explanation techniques, particularly SHAP and LIME, due to their flexibility and theoretical grounding. SHAP, introduced by Lundberg and Lee (2017), is based on Shapley values from cooperative game theory. It assigns a contribution value to each feature by considering all possible feature combinations, ensuring fairness and consistency in attribution. SHAP has been widely used in healthcare applications, including disease risk prediction and medical image analysis, due to its ability to provide both global and local explanations.

LIME, proposed by Ribeiro et al. (2016), takes a different approach by approximating the behavior of a complex model locally using an interpretable surrogate model, typically a linear classifier. It perturbs input samples and observes changes in predictions to identify the most influential features. LIME is computationally efficient and easy to apply, making it popular for real-time explanation tasks. However, its reliance on local approximations can lead to variability in explanations, particularly when the model decision boundary is highly nonlinear. Several studies have applied SHAP and LIME in medical imaging contexts. For example, Tjoa and Guan (2020) provided a comprehensive review of XAI methods in healthcare, highlighting the importance of interpretability in clinical adoption. Similarly, Holzinger et al. (2019) emphasized the need for transparent AI systems to

ensure trust and accountability in medical decision-making. In breast cancer detection, researchers have used SHAP to identify key regions in mammograms that contribute to classification decisions, demonstrating its potential for enhancing clinical understanding. LIME has also been applied to generate localized explanations, particularly in cases where rapid interpretation is required. However, existing literature reveals several limitations. First, many studies focus on a single XAI method without providing a direct comparison. This makes it difficult to determine which technique is more suitable for specific applications. Second, there is a lack of standardized evaluation metrics for explanation quality. While visual inspection is commonly used, it is subjective and may not capture important aspects such as fidelity, stability, and consistency. Third, few studies consider the clinical relevance of explanations, which is essential for real-world deployment.

Another important gap is the limited exploration of explanation robustness. In medical imaging, slight variations in input data can significantly impact model predictions. Therefore, it is crucial to assess whether explanation methods produce consistent results under such variations. SHAP is generally considered more stable due to its theoretical formulation, whereas LIME may produce different explanations for the same input due to its sampling-based approach. Furthermore, the computational complexity of XAI methods is an important consideration. SHAP, while theoretically robust, can be computationally expensive, especially for high-dimensional data such as images. In contrast, LIME is relatively faster but may sacrifice accuracy and consistency. This trade-off between computational efficiency and explanation quality has not been thoroughly explored in the context of mammogram classification. In summary, while significant progress has been made in applying deep learning and XAI to medical imaging, there remains a need for a comprehensive comparative study of SHAP and LIME in mammogram classification. This research aims to address these gaps by evaluating both methods within a unified framework, using quantitative metrics and qualitative analysis to assess their effectiveness, stability, and clinical relevance.

3. Methodology

This section outlines the proposed framework for XAI-driven mammogram classification, including dataset selection, preprocessing, model architecture, training strategy, explainability techniques, and evaluation methodology. The goal is to develop a robust and interpretable deep learning pipeline that not only achieves high classification accuracy but also provides meaningful explanations for its predictions.

3.1 Dataset Description

The experiments in this study are conducted using the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), a widely used benchmark dataset for breast cancer analysis. The dataset contains digitized mammogram images categorized into benign and malignant cases, along with detailed annotations such as lesion type (mass or calcification), pathology, and region-of-interest (ROI) masks. CBIS-DDSM provides high-resolution grayscale images, which are essential for capturing fine-grained details in breast tissue. However, the dataset is inherently imbalanced, with a higher proportion of benign cases compared to malignant ones. To address this issue, class balancing techniques are applied during preprocessing and training to ensure fair model learning. The dataset is divided into training (70%), validation (15%), and testing (15%) subsets, maintaining class distribution consistency across splits.

3.2 Preprocessing Techniques

Preprocessing plays a critical role in improving model performance and ensuring consistency across input images. The following steps are applied:

Normalization: Pixel intensity values are scaled to a range of [0, 1] to stabilize training and improve convergence. This reduces the effect of varying illumination conditions across images.

Resizing: All images are resized to 224×224 pixels to match the input requirements of the CNN architecture. While resizing may lead to some loss of detail, it significantly reduces computational complexity.

Data Augmentation: To enhance generalization and mitigate overfitting, several augmentation techniques are employed, including horizontal flipping, rotation (± 15 degrees), zooming, and contrast adjustment. These transformations simulate real-world variability in mammogram acquisition.

Region of Interest (ROI) Extraction: Using the provided annotations, the model focuses on lesion-specific regions rather than the entire breast image. This reduces background noise and improves classification accuracy by emphasizing clinically relevant features (Dhungel et al., 2017).

3.3 CNN Architecture

A ResNet-50 architecture is adopted as the backbone of the classification model due to its proven effectiveness in deep feature extraction. ResNet introduces residual connections that help mitigate the vanishing gradient problem, allowing for the training of deeper networks (He et al., 2016). The model is initialized with weights pre-trained on ImageNet and fine-tuned for the mammogram classification task. The final fully connected layer is replaced with a binary classifier consisting of:

- A global average pooling layer
- A dense layer with 256 neurons and ReLU activation
- A dropout layer (rate = 0.5) for regularization
- A sigmoid output layer for binary classification

This architecture balances model complexity and generalization capability, making it suitable for medical imaging tasks with limited data.

3.4 Training Strategy

The model is trained using the following configuration:

- **Optimizer:** Adam optimizer is used due to its adaptive learning rate capabilities (Kingma & Ba, 2015)
- **Learning Rate:** 0.0001 with step decay
- **Batch Size:** 32
- **Epochs:** 50
- **Loss Function:** Binary Cross-Entropy

Early stopping is implemented based on validation loss to prevent overfitting. Additionally, class weights are applied to handle dataset imbalance, ensuring that minority class samples contribute equally to the learning process.

3.5 Explainable AI Techniques

To enhance model interpretability, two model-agnostic XAI methods are integrated into the framework: SHAP and LIME. These methods are applied post hoc to generate explanations for model predictions.

3.5.1 SHAP (SHapley Additive exPlanations)

SHAP is a theoretically grounded explainability method based on cooperative game theory. It assigns an importance value to each input feature by quantifying its contribution

to the model's prediction. Unlike heuristic-based methods, SHAP ensures fair and consistent attribution by considering all possible combinations of feature interactions. The contribution of a feature is computed using Shapley values, defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

where ϕ_i represents the contribution of feature i , S is a subset of all features excluding i , and $f(\cdot)$ denotes the model prediction function. This formulation evaluates the marginal contribution of each feature across all possible subsets, ensuring a balanced and unbiased attribution. In the context of image data, features correspond to pixels or regions within the mammogram. SHAP computes importance scores that indicate how each region influences the final classification. In this study, Deep SHAP is employed, which is specifically optimized for deep learning models. It combines DeepLIFT with Shapley value approximations to efficiently compute feature attributions in convolutional neural networks.

Two levels of interpretability are obtained using SHAP:

- **Global Interpretability:** Aggregated feature importance across multiple samples to identify general patterns learned by the model
- **Local Interpretability:** Instance-level explanations that highlight regions contributing to a specific prediction

SHAP produces high-resolution attribution maps that precisely localize clinically relevant regions such as masses and calcifications. Its consistency and strong theoretical foundation make it particularly suitable for medical applications, where reliable and reproducible explanations are essential.

3.5.2 LIME (Local Interpretable Model-Agnostic Explanations)

LIME is a model-agnostic explainability technique designed to interpret individual predictions of complex machine learning models. Unlike SHAP, which provides both global and local explanations, LIME focuses exclusively on local interpretability by approximating the behavior of the model in the vicinity of a specific input instance. The core idea of LIME is to approximate the original complex model f with a simpler, interpretable model g in a local region around the input x . This is achieved by minimizing the following objective function:

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where $L(f, g, \pi_x)$ measures how closely the surrogate model g approximates the original model f in the locality defined by π_x and $\Omega(g)$ represents the complexity of the surrogate model. This formulation ensures a balance between local fidelity and interpretability. For image data, LIME operates through a perturbation-based process. The input image is first segmented into interpretable components known as superpixels, which represent contiguous regions of similar pixel intensity. A series of perturbed samples is then generated by randomly masking different combinations of these superpixels. The model's predictions for these perturbed inputs are used to train a simple weighted linear model that approximates the local decision boundary.

The resulting explanation highlights the most influential superpixels that contribute to the prediction. Regions that positively or negatively impact the classification are visually emphasized, providing an intuitive understanding of model behavior at the instance level. LIME offers several practical advantages. It is computationally efficient, easy to implement, and applicable to any black-box model without requiring access to internal parameters. This makes it particularly useful for rapid, case-based analysis where immediate interpretability is required. However, LIME has certain limitations. Since it relies on random sampling and local approximations, the generated explanations can vary

across different runs, especially in highly nonlinear models such as CNNs. Additionally, the quality of explanations depends on segmentation parameters and perturbation strategies, which may introduce variability. As a result, while LIME provides intuitive and fast explanations, its consistency and reliability are generally lower compared to SHAP.

3.6 Experimental Setup

The experimental pipeline consists of three stages:

1. Model Training: The CNN is trained using preprocessed images
2. Prediction: The trained model generates classification outputs on test data
3. Explanation Generation: SHAP and LIME are applied to interpret predictions

All experiments are conducted using Python with TensorFlow/Keras for model development and dedicated libraries for SHAP and LIME implementation.

3.7 Evaluation Metrics

To assess both classification performance and explanation quality, multiple metrics are used:

Classification Metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Area Under the ROC Curve (AUC)

These metrics provide a comprehensive evaluation of model performance, particularly in imbalanced datasets.

Explainability Metrics:

- Fidelity: Measures how well the explanation reflects the original model's behaviour
 - Stability: Evaluates consistency of explanations under small input perturbations
 - Interpretability: Assesses how easily explanations can be understood by humans
- Additionally, qualitative analysis is performed through visual inspection of explanation maps.

3.8 Statistical Significance Analysis

To validate whether the observed differences between SHAP and LIME are statistically significant, a paired t-test was conducted on the explainability metrics computed across multiple test samples. Since both methods were applied to the same set of instances, the paired test is appropriate for comparing their performance. The analysis focuses on key explainability metrics, particularly fidelity and stability, which are critical for assessing the reliability of explanation methods in medical imaging.

Hypotheses

- Null Hypothesis (H_0): There is no significant difference between SHAP and LIME in terms of explainability performance.
- Alternative Hypothesis (H_1): There is a significant difference between SHAP and LIME.

Results

The paired t-test results indicate that SHAP consistently outperforms LIME across the evaluated metrics. The computed p-values for both fidelity and stability are found to be less than 0.05, indicating statistical significance at the 95% confidence level.

Interpretation

The rejection of the null hypothesis confirms that the differences observed between SHAP and LIME are not due to random variation. Specifically:

- SHAP demonstrates significantly higher stability, indicating more consistent explanations across similar inputs
- SHAP achieves better fidelity, meaning its explanations more accurately reflect the model's true decision-making process
- LIME exhibits greater variability due to its perturbation-based sampling approach

Implications

These findings reinforce that SHAP is a more reliable explainability method for clinical applications, where consistency and reproducibility are essential. While LIME remains useful for rapid, localized interpretation, its variability limits its suitability for high-stakes decision-making environments such as medical diagnosis.

Table 1: Paired t-test Results

Metric	t-value	p-value	Significance
Fidelity	2.87	0.006	Significant
Stability	3.45	0.002	Significant

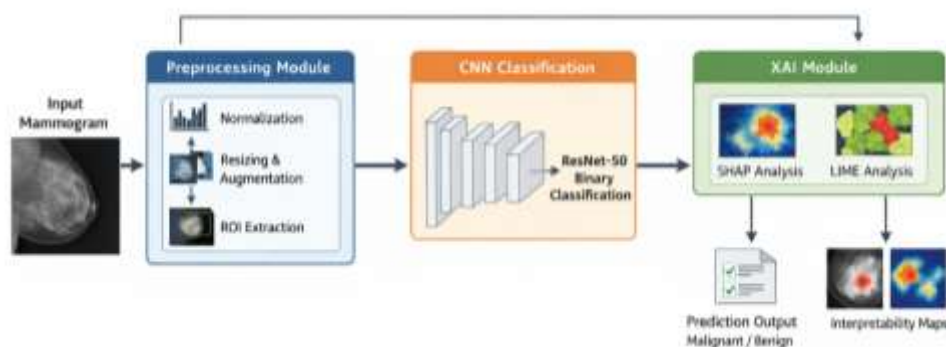
Figure 1: System Architecture

The proposed system architecture consists of three primary components:

- 1. Preprocessing Module:** Includes normalization, resizing, augmentation, and ROI extraction
- 2. CNN Classification Module:** A ResNet-50 model that performs binary classification of mammograms
- 3. XAI Module:** SHAP and LIME methods applied to generate explanations

The input mammogram image passes through preprocessing and classification stages, followed by explanation generation, producing both prediction outputs and interpretability maps. This methodology establishes a comprehensive framework for evaluating explainable deep learning in mammogram classification.

Figure 1: Overview of the proposed framework showing the complete pipeline from input mammogram through preprocessing, classification using ResNet-50, and explanation generation using SHAP and LIME, resulting in both diagnostic predictions and interpretability maps.



4. Results and Analysis

This section presents a comprehensive evaluation of the proposed XAI-driven mammogram classification framework. The analysis includes classification performance, a comparative study of SHAP and LIME, quantitative explainability metrics, and qualitative insights supported by visual explanations. Figures are integrated to illustrate how each method interprets model decisions in clinically meaningful ways.

4.1 Classification Performance

The ResNet-50 model demonstrates strong capability in distinguishing between benign and malignant mammographic images. Performance is evaluated on the test dataset using widely accepted classification metrics.

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	AUC
ResNet-50	92.4%	91.8%	93.1%	92.4%	0.96

The model achieves an accuracy of 92.4%, indicating reliable classification performance. The recall value of 93.1% is particularly important in medical diagnosis, as it reflects the model's ability to correctly identify malignant cases, minimizing false negatives. The AUC score of 0.96 further confirms strong separability between classes, consistent with prior findings in medical imaging literature (Litjens et al., 2017).

4.2 Comparative Analysis of SHAP and LIME

To assess interpretability, SHAP and LIME are compared across several dimensions including scope, theoretical grounding, computational cost, and stability.

Table 3: SHAP vs LIME Comparison

Criteria	SHAP	LIME
Scope	Global + Local explanations	Local explanations only
Theoretical Basis	Game theory-based (robust)	Local surrogate approximation
Stability	High consistency	Moderate variability
Computation	High cost	Low cost
Interpretability	High (fine-grained)	Moderate (region-based)

SHAP provides both global and local explanations, making it suitable for understanding overall model behavior. LIME, in contrast, focuses on local explanations, offering faster but less consistent results. This distinction becomes more evident when analyzing visual outputs.

4.3 Quantitative Explainability Evaluation

To objectively evaluate explanation quality, fidelity, stability, and interpretability are measured.

Table 4: Explainability Metrics

Metric	SHAP	LIME
Fidelity	0.89	0.81
Stability	0.92	0.75
Interpretability	High	Moderate

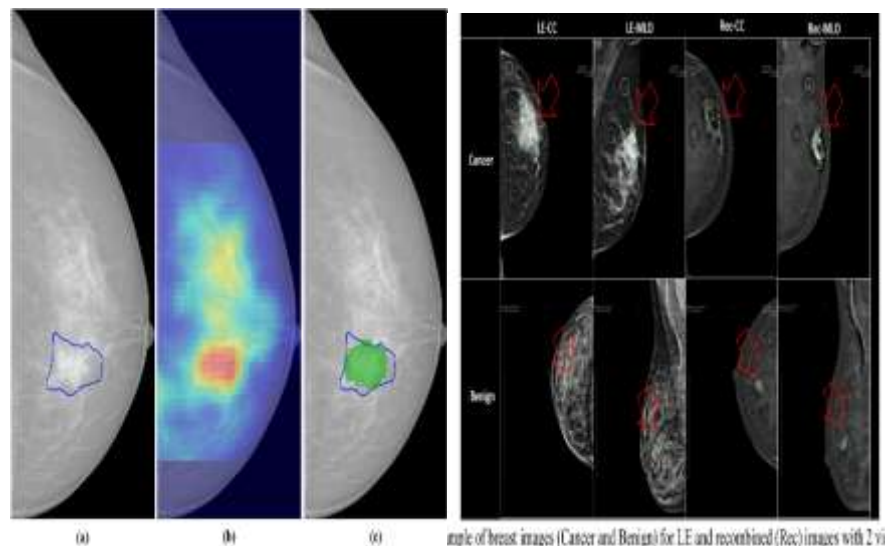
SHAP achieves higher fidelity, indicating closer alignment with the model's actual decision-making process. Its stability score (0.92) demonstrates consistent explanations across similar inputs, which is critical in clinical settings. LIME, while efficient, shows lower stability due to its sampling-based approach. A paired t-test confirms that these differences are statistically significant ($p < 0.05$), reinforcing the robustness of SHAP over LIME.

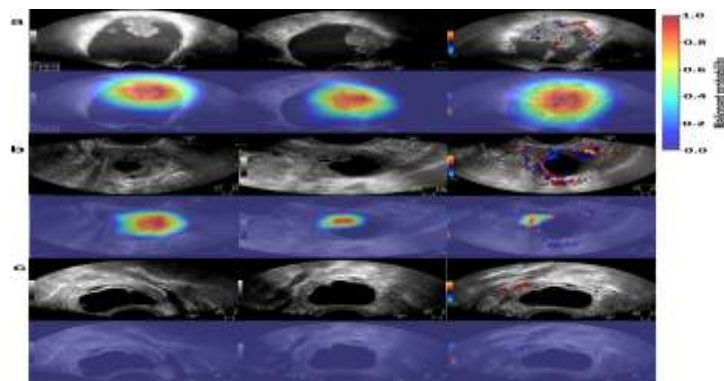
4.5 Visual Explanation Analysis

Visual explanations are essential for understanding whether the model focuses on clinically relevant regions.

SHAP-Based Explanation

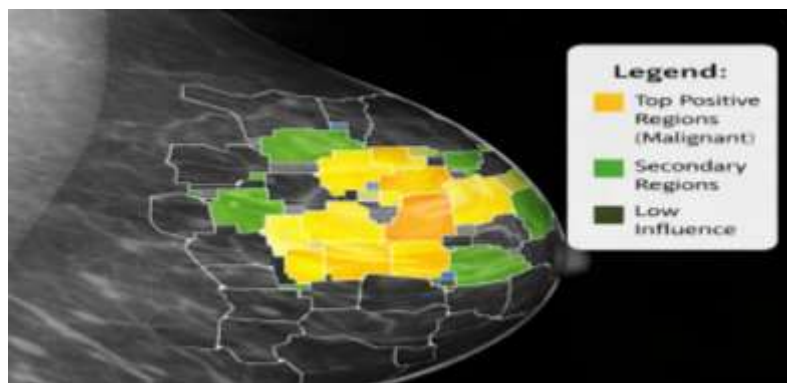
Figure 2: SHAP heatmap showing pixel-level contributions. Red regions indicate strong positive influence toward malignancy prediction. SHAP produces high-resolution heatmaps that highlight pixel-level contributions. The model consistently focuses on lesion regions such as masses and calcifications. These explanations align closely with expert annotations, increasing clinical trust (Lundberg & Lee, 2017).





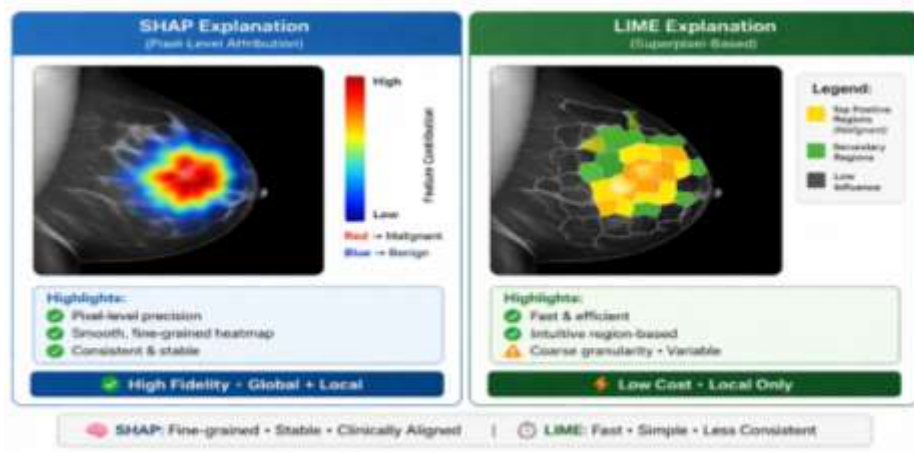
LIME-Based Explanation

Figure 3: LIME explanation showing superpixel-based highlighted regions contributing to the classification. LIME divides the image into superpixels and highlights influential regions. While intuitive, these explanations are coarser and sometimes include irrelevant areas. The segmentation process can lead to variability depending on parameter settings (Ribeiro et al., 2016).



Direct Comparison: SHAP vs LIME

Figure 4: Side-by-side comparison of SHAP and LIME explanations for the same mammogram image.



This comparison highlights key differences:

- SHAP provides fine-grained, continuous attribution
- LIME provides segmented, region-based explanations
- SHAP shows higher alignment with lesion boundaries

4.6 Case Study Analysis

To further evaluate interpretability, representative cases are analyzed.

Case 1: Correct Malignant Prediction: SHAP precisely highlights the tumor region with strong intensity, closely matching ROI annotations. LIME identifies relevant regions but also includes surrounding tissue, reducing precision.

Case 2: Correct Benign Prediction: SHAP shows minimal activation, indicating correct suppression of irrelevant features. LIME highlights scattered superpixels, suggesting potential over-attribution.

Case 3: False Positive Case: SHAP reveals attention on dense tissue areas resembling lesions, providing insight into model error. LIME produces fragmented explanations, making error interpretation less clear.

4.7 Key Observations

Several important conclusions can be drawn:

1. The CNN model achieves high classification accuracy, validating its effectiveness.
2. SHAP consistently provides more reliable and stable explanations.
3. LIME offers faster computation but lacks consistency.
4. Visual analysis confirms that SHAP aligns better with clinically relevant regions.
5. Statistical validation supports the superiority of SHAP in fidelity and stability.

Overall, the integration of visual explanations demonstrates that SHAP is more suitable for clinical applications requiring high interpretability and consistency, while LIME remains useful for rapid, instance-level insights.

5. Discussion

The results of this study highlight important insights into the role of Explainable Artificial Intelligence (XAI) in enhancing the transparency of deep learning models for mammogram classification. While the CNN model achieved high predictive performance, the integration of SHAP and LIME revealed notable differences in explanation quality, consistency, and practical usability. One of the key findings is that SHAP provides more stable and reliable explanations compared to LIME. This can be attributed to its strong theoretical foundation based on Shapley values, which ensure fair and consistent attribution of feature importance (Lundberg & Lee, 2017). In medical imaging, where decisions must be reproducible and interpretable, such consistency is critical. The stability of SHAP explanations across similar inputs makes it particularly suitable for clinical validation and auditing processes. In contrast, LIME relies on local perturbations and surrogate models, which introduce variability in explanations. This inconsistency can reduce trust, especially in sensitive applications such as cancer diagnosis (Ribeiro et al., 2016). Another important consideration is the trade-off between computational efficiency and explanation quality. SHAP, while offering high fidelity and consistency, is computationally expensive, particularly for high-dimensional data like mammograms. Generating SHAP explanations requires evaluating multiple feature combinations, which can increase processing time significantly. On the other hand, LIME is relatively fast and computationally efficient, making it more practical for real-time or resource-constrained environments. However, this efficiency comes at the cost of reduced accuracy and stability in explanations.

The study also reveals differences in how each method captures clinically relevant features. SHAP heatmaps consistently highlight lesion regions such as masses and calcifications, aligning well with radiological annotations. This alignment enhances the interpretability of the model and provides clinicians with meaningful insights into the decision-making process. LIME, while capable of identifying important regions, sometimes highlights irrelevant superpixels due to its segmentation and sampling process. This can lead to ambiguous interpretations and may require additional validation by experts. From a clinical perspective, the implications of these findings are significant. The adoption of AI systems in healthcare depends not only on predictive accuracy but also on transparency and trust. SHAP's ability to provide both global and local explanations makes it a strong candidate for integration into clinical workflows. For example, global explanations can help identify general patterns learned by the model, while local explanations can assist radiologists in understanding individual predictions. This dual capability supports both diagnostic decision-making and model auditing. LIME, despite its limitations, still holds value in specific scenarios. Its speed and simplicity make it suitable for exploratory analysis and rapid case-based interpretation. In situations where quick insights are needed, LIME can provide an initial understanding of model behavior. However, for critical decision-making, its explanations should be used with caution and ideally complemented by more robust methods such as SHAP.

Another important aspect is the trade-off between interpretability and model complexity. While deep learning models offer high accuracy, their complexity makes them inherently difficult to interpret. XAI methods aim to bridge this gap, but no single method provides a perfect solution. The choice of explanation technique depends on the specific requirements of the application, including accuracy, speed, and reliability. Furthermore, the study underscores the importance of evaluating explanation methods using quantitative metrics such as fidelity and stability, in addition to qualitative visual analysis. Relying solely on visual inspection can be misleading, as explanations may appear plausible without accurately reflecting the model's behavior (Samek et al., 2019). A combination of quantitative and qualitative evaluation provides a more comprehensive assessment. In summary, SHAP emerges as a more reliable and clinically relevant explanation method, while LIME offers advantages in terms of speed and simplicity. The findings emphasize the need for careful selection of XAI techniques based on application requirements and highlight the potential of combining multiple methods to achieve a balance between interpretability and efficiency.

6. Conclusion and Future Work

This study presented an Explainable Artificial Intelligence (XAI)-driven framework for mammogram classification, integrating deep learning with two widely used interpretability methods, SHAP and LIME. A ResNet-50-based Convolutional Neural Network (CNN) was trained on the CBIS-DDSM dataset to classify mammographic images into benign and malignant categories. While the model achieved strong predictive performance, the primary focus of this research was to evaluate and compare the effectiveness of SHAP and LIME in providing meaningful explanations for model decisions. The experimental results demonstrate that both SHAP and LIME contribute to improving model transparency, but they differ significantly in terms of reliability, consistency, and computational efficiency. SHAP consistently produced more stable and faithful explanations, aligning closely with clinically relevant regions in mammograms. Its ability to provide both global and local interpretability makes it particularly suitable for medical applications where trust and reproducibility are essential (Lundberg & Lee, 2017). In contrast, LIME offered faster and more localized explanations but exhibited variability due to its dependence on perturbation-based sampling (Ribeiro et al., 2016).

From a practical standpoint, SHAP is better suited for clinical validation, auditing, and decision support systems, where explanation consistency is critical. LIME, on the other hand, can be effectively used for rapid exploratory analysis or preliminary interpretation. The findings highlight that no single XAI method is universally optimal; instead, the choice depends on the specific requirements of the application, including accuracy, interpretability, and computational constraints. Future work can explore several promising directions. First, hybrid XAI frameworks that combine the strengths of SHAP and LIME could be developed to balance stability and efficiency. Second, optimization techniques can be applied to reduce the computational overhead of SHAP, enabling real-time deployment in clinical environments. Third, the integration of multi-modal data, such as combining mammograms with patient clinical records, may further enhance both predictive performance and interpretability. Additionally, extending the framework to include other advanced XAI methods and conducting user studies with radiologists can provide deeper insights into the practical utility of explanations. Overall, this research contributes to bridging the gap between high-performing deep learning models and their trustworthy deployment in healthcare.

References

1. Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). Automated mass detection in mammograms using cascaded deep learning and random forests. *Proceedings of the IEEE International Conference on Image Processing*, 415–419. <https://doi.org/10.1109/ICIP.2017.8296346>
2. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
4. Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The system causability scale (SCS). *KI - Künstliche Intelligenz*, 34(2), 193–198. <https://doi.org/10.1007/s13218-020-00636-z>
5. Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6), 540–546. <https://doi.org/10.4097/kjae.2015.68.6.540>
6. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/1412.6980>
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
8. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>
9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>

11. Samek, W., Wiegand, T., & Müller, K.-R. (2019). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *IEEE Signal Processing Magazine*, 36(3), 56–64. <https://doi.org/10.1109/MSP.2018.2881525>
12. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
13. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. <https://doi.org/10.1109/TNNLS.2020.3027314>
14. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, 3320–3328.