Explainable Deep Learning for Breast Ultrasound Imaging: Analysis and Comparison of XAI Techniques

Spoorti Udgirkar[1*], Sneha Reddy[2], Jacinta[3], Mohan.K[4]

[1,2,3,4]*CSE (Data Science), Guru Nanak Dev Engineering College , Karnataka , India.*

*spoortiudgirkar2004@gmail.com

Abstract

Breast cancer remains one of the most frequently diagnosed malignancies in women globally, and timely detection serves an important function in improving patient survival. When diagnosed at an early stage, the 5-year survival rate exceeds 99%. Ultrasound imaging is extensively utilized for breast tumor detection due to its non-invasive nature and effectiveness in analyzing dense breast tissue. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated high accuracy in breast ultrasound image analysis. However, their black-box nature limits interpretability in clinical applications. To address this, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM, LIME, and SHAP have been introduced to provide visual explanations of model predictions. This study presents a comparative analysis of XAI methods for breast tumor localization in ultrasound images. The evaluation is based on key factors including localization capability, boundary clarity, noise sensitivity, and computational cost. The analysis highlights a lack of modality-specific comparisons in existing literature. Results indicate that gradient-based approaches, particularly Grad-CAM variants, offer an effective balance between interpretability and performance, supporting their use in clinical decision-making.

Keywords— *Breast Cancer, Breast Ultrasound Imaging, Convolutional Neural Networks (CNNs), Deep Learning (DL), Explainable Artificial Intelligence (XAI), Medical Image Analysis, Tumor Detection, Tumor Localization.*

1. Introduction

Breast cancer continues to be one of the most frequently diagnosed cancers among women worldwide and represents a serious public health issue. Each year, nearly 2.3 million new cases are reported, making it the most prevalent cancer among women globally [1]. It is also a leading cause of cancer-related mortality, accounting for approximately 685,000 deaths annually [2]. In India, breast cancer has emerged as the most common cancer among women, contributing to around 28% of total female cancer cases, with a steadily rising incidence, especially in urban regions [3]. Early-stage detection is critical for improving survival outcomes; however, challenges such as lack of awareness, insufficient screening infrastructure, and delayed diagnosis continue to limit effective disease control in many developing countries [1], [3].

Medical imaging plays a central role in the detection and diagnosis of breast abnormalities. Techniques such as mammography and ultrasound are routinely used for screening, while magnetic resonance imaging (MRI) offers enhanced sensitivity in complex diagnostic scenarios [4]. Despite advancements in imaging technologies, interpretation of medical images largely depends on expert analysis, which can be both time-intensive and subject to variability between radiologists. To enhance diagnostic consistency and efficiency, computer-aided diagnosis (CAD) systems have been developed to support clinicians in identifying suspicious patterns in medical images [5]. In recent years, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown remarkable success in automating breast tumor classification by learning intricate feature representations directly from imaging data [6].

Among available imaging modalities, ultrasound is widely preferred due to its non-invasive nature and ease of access. Unlike mammography and computed tomography (CT), ultrasound does not use ionizing radiation, making it suitable for frequent examinations and for younger women with dense breast tissue [7]. Furthermore, studies have demonstrated that combining ultrasound with mammography can improve cancer detection rates in high-risk patients [8]. The ability of ultrasound to provide real-time imaging also supports precise tumor localization and assists in guiding biopsy procedures, reinforcing its importance as a complementary diagnostic tool [9].

Deep learning techniques have substantially enhanced medical image interpretation by enabling automated extraction of relevant features and improving predictive performance. CNN-based approaches have been successfully applied to a range of tasks, including tumor classification, detection, and segmentation across multiple imaging modalities such as ultrasound, MRI, mammography, and CT scans [10], [11]. Although these models achieve high accuracy, their decision-making processes are often not transparent, which limits their acceptance in clinical practice. To address this issue, Explainable Artificial Intelligence (XAI) methods have been introduced to provide insights into model predictions by highlighting the most relevant regions within input images. Techniques such as Grad-CAM, LIME, and SHAP are commonly used to generate visual explanations and improve model interpretability [13], [21], [22], [23].

While existing research has explored explainable AI across various medical imaging domains, relatively fewer studies have focused specifically on breast ultrasound imaging [14], [15]. Ultrasound images present unique challenges, including speckle noise, low contrast, and irregular tumor shapes, which can influence the effectiveness of different explainability methods. Therefore, this work presents a qualitative, literature-based comparison of widely used XAI techniques, including Grad-CAM, Grad-CAM++, LIME, and SHAP, for breast ultrasound tumor analysis. The comparison evaluates these methods based on tumor localization capability, interpretability, and computational complexity.

2. Background

A. Breast Ultrasound Imaging

Breast ultrasound is commonly utilized for tumor assessment due to its safe operation, real-time imaging capability, and wide availability. In contrast to imaging modalities such as mammography and computed tomography (CT), ultrasound does not rely on ionizing radiation, making it appropriate for repeated use and for examining younger patients with dense breast tissue [7]. Its comparatively lower cost and portable design further support its use in rural and resource-constrained healthcare environments [8]. The ability to capture real-time images also assists clinicians in accurately identifying lesion locations and guiding biopsy procedures [9].

Despite these advantages, ultrasound imaging is associated with several inherent limitations. Image quality is often affected by speckle noise, reduced contrast between tissues, and poorly defined tumor boundaries, which can complicate visual interpretation [9]. These factors increase dependence on radiologist expertise and highlight the need for automated approaches to support consistent and accurate tumor detection and classification.

B. Deep Learning in Medical Imaging

Advances in deep learning have significantly reshaped medical image analysis by enabling models to learn informative patterns directly from raw data. CNNs are particularly effective in capturing spatial relationships within images and have been widely applied to classification, detection, and segmentation tasks across modalities such as ultrasound, MRI, mammography, and CT [10], [11]. By learning hierarchical representations, these models reduce reliance on manually engineered features and often achieve improved diagnostic performance.

In breast ultrasound applications, CNN-based models have been successfully employed for identifying tumor regions, segmenting lesions, and classifying lesions as benign or malignant cases [6]. Architectures such as ResNet and U-Net have further contributed to improved accuracy and robustness in automated analysis [12], [17]. However, despite their strong predictive capability, these models provide limited insight into how decisions are made, which restricts their acceptance in clinical environments where interpretability is essential.

C. Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence techniques have been introduced to address the interpretability limitations associated with deep learning models. These approaches aim to reveal which regions of an input image contribute most significantly to a model's prediction, thereby improving transparency. Gradient-based methods such as Grad-CAM and Grad-CAM++ utilize internal network information to highlight influential regions in images [13], [23]. In contrast, model-agnostic approaches like LIME and SHAP analyze the impact of input variations on model output, providing more generalized explanations across different models [21], [22].

By offering visual or feature-based insights, these techniques enable clinicians to verify whether predictions are aligned with relevant tumor regions, thereby improving trust in AI-assisted systems. Although XAI methods have been explored in various medical imaging domains, their application in breast ultrasound imaging remains relatively limited [14], [15]. The presence of noise, low contrast, and irregular lesion structures can influence the reliability of generated explanations, making domain-specific evaluation essential.

D. Research Gap

Existing studies on explainable AI in medical imaging primarily focus on individual methods or on imaging modalities other than ultrasound [14], [15]. Comparative analyses that specifically examine multiple XAI techniques in the context of breast ultrasound imaging are limited. Furthermore, challenges such as speckle noise, low contrast, and irregular tumor boundaries are not comprehensively addressed in current literature.

Therefore, there is a need for a focused, literature-based comparison of commonly used XAI approaches. Evaluating these methods in terms of tumor localization capability, interpretability, and computational complexity can provide valuable insights for selecting appropriate techniques and guiding future research in breast ultrasound analysis.

3. Deep Learning Models for Breast Ultrasound Analysis

Deep learning has reshaped medical image interpretation by enabling automated and scalable analysis of breast ultrasound-based imaging for tumor detection and evaluation. Earlier approaches relied heavily on manually designed features such as texture descriptors, intensity patterns, and geometric properties, which often struggle to represent the high variability observed in tumor appearance. In contrast, modern deep learning techniques learn feature representations directly from raw data, allowing models to capture complex visual patterns more effectively.

Among these methods, Convolutional Neural Networks (CNNs) have emerged as effective models, offering enhanced diagnostic accuracy through hierarchical feature learning that supports improved sensitivity and specificity in breast cancer detection [6], [11].

A. CNN-based Methods

CNNs form the backbone of most deep learning frameworks applied in medical image analysis. These networks employ stacked convolutional layers, pooling operations, and nonlinear activations to progressively extract features ranging from basic edges to high-level semantic representations. In the context of breast ultrasound imaging, CNNs have shown strong capability in distinguishing benign from malignant lesions, while also identifying subtle tumor boundaries under challenging imaging conditions [6], [11].

Well-known architectures such as AlexNet and VGGNet have been adapted for medical datasets to improve classification performance. Evidence from clinical studies suggests that CNN-based systems achieve strong diagnostic performance in medical imaging, including breast ultrasound applications [6], [16].

B. Transfer Learning Models

Transfer learning has emerged as a practical solution when large annotated medical datasets are not available. In this approach, models pretrained on large-scale datasets (e.g., ImageNet) are fine-tuned for specific medical imaging tasks. Architectures such as ResNet and DenseNet are commonly used due to their ability to retain and adapt previously learned visual features [11], [12].

For breast ultrasound analysis, transfer learning helps improve classification accuracy while reducing overfitting and training time. Studies indicate that deeper pretrained models often outperform conventional CNNs, particularly in identifying malignant lesions with higher sensitivity while maintaining reliable specificity [11].

C. Segmentation Models (U-Net)

Segmentation is a critical step in breast ultrasound analysis, as it enables precise localization and delineation of tumor regions. The U-Net architecture and its extensions are widely adopted for this purpose due to their effectiveness in pixel-level prediction tasks [17], [18].

The encoder–decoder structure of U-Net allows simultaneous capture of global contextual information and fine spatial details. This is particularly useful in ultrasound imaging, where noise, low contrast, and irregular tumor shapes are common challenges. Enhanced variants such as U-Net++ introduce dense skip connections to improve feature propagation and boundary refinement, resulting in more accurate segmentation outputs [18].

In many systems, segmentation is combined with classification models to provide both localization and malignancy prediction, thereby improving overall diagnostic performance.

D. Hybrid and Ensemble Models

Hybrid approaches integrate multiple techniques to leverage their complementary strengths. For instance, CNNs can be used for feature extraction, followed by traditional classifiers including classifiers like Support Vector Machines (SVMs) and Random Forests for final prediction. Similarly, attention mechanisms are often incorporated into segmentation networks to focus on clinically relevant regions [7], [14].

Ensemble learning further enhances robustness by combining predictions from multiple models, reducing variability and minimizing false positives across different datasets. These strategies improve generalization and make deep learning models more suitable for real-world clinical applications.

TABLE I presents an overview and comparison of deep learning architectures used in breast ultrasound imaging.

Model / Architecture	Key Features	Application	Advantages	Limitations	References
CNN (VGG, AlexNet)	Multiple convolution + pooling layers; learns hierarchical features	Tumor classification (benign vs malignant)	High accuracy; automated feature extraction; robust for varied tumor shapes	Requires large datasets; limited interpretability	[6], [7]
ResNet / DenseNet (Transfer Learning)	Residual connections (ResNet) or dense connections (DenseNet); pretrained on ImageNet	Tumor classification; cross-modality adaptation	Handles small datasets; faster convergence; strong generalization	May require fine-tuning; still somewhat opaque	[11], [12]
U-Net / U-Net++	Encoder-decoder architecture; skip connections;	Tumor segmentation and localization	Accurate tumor boundary delineation; handles low	Computationally expensive; requires annotated	[17], [18]

	nested connections (U-Net++)		contrast; works on noisy ultrasound	masks	
Hybrid Models	Combination of CNN + classical ML classifiers or segmentation + attention modules	Classification + localization; robustness enhancement	Leverages strengths of multiple methods; improves robustness	Complex architecture; higher computational cost	[7], [14]
Ensemble Models	Aggregates predictions from multiple models	Classification / Detection	Reduces false positives; stabilizes performance	Requires multiple models; longer inference time	[7], [10]

TABLE I: Deep Learning Models for Breast Ultrasound Analysis

4. Explainable AI (XAI) for Breast Ultrasound

Reliable and interpretable tumor localization is a critical requirement for integrating Deep Learning (DL) systems into clinical breast ultrasound workflows. While Convolutional Neural Networks (CNNs) demonstrate strong classification capability, their decision-making process is often opaque, which restricts their acceptance in safety-critical medical environments. To address this limitation, Explainable Artificial Intelligence (XAI) techniques have been introduced to provide visual interpretations of model outputs by identifying regions that contribute most significantly to predictions. This section reviews prominent XAI approaches applied to breast ultrasound imaging, focusing on their effectiveness in tumor localization, boundary delineation, robustness to noise, computational efficiency, and ease of integration into clinical workflows

4.1 Overview of XAI Methods

XAI methods enhance the interpretability of DL models by revealing the contribution of image regions to prediction outcomes. In breast ultrasound imaging, these approaches typically generate heatmaps that highlight diagnostically relevant areas, assisting clinicians in validating automated decisions [13], [14].

Fig. 1 illustrates a standard pipeline in which ultrasound images are pre-processed and analysed using a deep learning model, followed by the application of XAI techniques such as Grad-CAM, LIME, and SHAP to produce visual explanations for tumor localization.

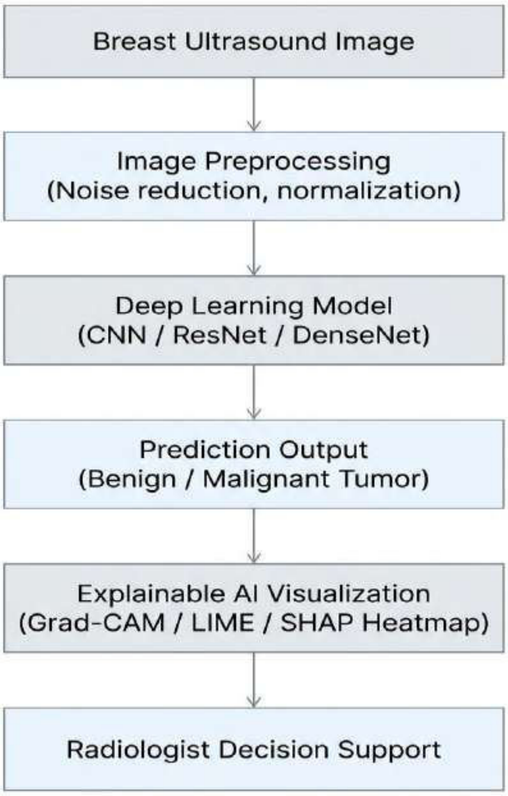

Fig. 1. Deep learning and XAI-based pipeline for breast ultrasound tumor analysis.

4.1.1 Gradient-Based Methods

Gradient-based techniques utilize derivatives of model outputs with respect to input features or intermediate activations to produce saliency maps.

Grad-CAM: Grad-CAM generates class-discriminative localization maps by combining gradient information with convolutional feature maps. It provides reliable tumor localization with relatively clear boundaries, making it suitable for clinical interpretation [13].

Grad-CAM++: An extension of Grad-CAM, this method improves localization performance in images containing multiple regions of interest. It produces more precise boundary delineation and exhibits reduced sensitivity to noise, while maintaining moderate computational requirements. These characteristics make it highly effective for ultrasound-based tumor localization tasks [23].

Guided Grad-CAM: Guided Grad-CAM combines Grad-CAM with guided backpropagation to generate high-resolution and visually sharp explanations. It enhances boundary clarity and highlights fine-grained tumor structures more effectively than standard Grad-CAM. However, this improvement comes with increased computational cost compared to basic gradient-based methods [13].

4.1.2 Perturbation-Based Methods

Perturbation-based approaches systematically modify input images to assess their impact on predictions.

LIME (Local Interpretable Model-Agnostic Explanations): Perturbs image super pixels and measures prediction changes. Medium localization and poor boundary clarity; computationally expensive, not ideal for real-time clinical use [21].

SHAP (Shapley Additive explanations): Provides pixel-level attribution based on Shapley values. Medium localization, medium boundary clarity, and high computational cost; useful for understanding model decisions but slow for real-time deployment [22].

Score-CAM: Combines activation maps with forward-passing model scores to generate more reliable heatmaps. Reduces noise and provides medium to high localization with moderate computational overhead [19].

4.1.3 Backpropagation-Based Methods

Backpropagation-based techniques propagate output signals backward through the network to identify influential pixels.

Saliency Maps: Highlight influential pixels using raw gradients. Low localization, poor boundary clarity, and high noise make it less reliable for clinical applications [20].

Layer-wise Relevance Propagation (LRP): Distributes prediction scores to input pixels using layer-wise decomposition. High localization, good boundary clarity, but medium-high computational cost; clinically useful but complex to implement [14].

Integrated Gradients: Computes average gradients along the path from baseline to input. High localization and boundary clarity with moderate computational cost; provides interpretable results similar to Grad-CAM [24].

4.1.4 Model-Agnostic Methods

Model-agnostic techniques can be applied to a wide range of architectures without requiring internal model modifications.

Occlusion Maps: This method evaluates prediction sensitivity by masking different regions of the input image. While it provides intuitive explanations with moderate localization accuracy, it is computationally expensive and unsuitable for real-time applications [21].

RISE (Randomized Input Sampling for Explanation): RISE generates heatmaps by averaging predictions across randomly masked inputs. It produces high-quality localization with improved boundary clarity but requires significant computational resources, limiting its clinical practicality [19].

Attention-Based Methods: Attention mechanisms embedded within deep learning models highlight regions of interest directly during inference. These approaches are computationally efficient and provide moderate localization performance, although their effectiveness depends on model design [14].

Counterfactual Explanations: Counterfactual methods generate modified inputs to analyze decision boundaries. While useful for understanding model behaviour, they are less effective for precise tumor localization and are computationally demanding [15].

4.2 Research Contribution: Unified XAI Comparison Table

Existing XAI studies lack a breast ultrasound-specific survey that systematically compares methods based on localization, boundary clarity, noise robustness, computational cost, and clinical suitability [7], [15].

To address this gap, we present TABLE II, summarizing 13 XAI techniques applied to breast ultrasound tumor detection. This table consolidates information on localization accuracy, boundary clarity, noise sensitivity, computational decision boundaries. While useful for understanding model behaviour, they are less effective for precise tumor localization and are computationally demanding [10], [15].

Method	What it Shows (Breast Ultrasound)	Tumor Localization	Boundary Clarity	Noise	Computational Cost	Clinical Suitability	References
LIME	Super pixel regions on lesion	Medium	Poor	High	High	Not ideal	[21]
SHAP	Pixel importance map	Medium	Medium	Medium	Very High	Slow	[22]
Score-CAM	Activation-based heatmap	Medium–High	Good	Low	Medium	Good	[19]
Grad-CAM	Tumor region heatmap	High	Good	Low	Low	Best	[13]
Guided Grad-CAM	Sharper tumor map	High	Very Good	Medium	Medium	Good	[13]
Grad-CAM++	Multiple tumor localization	Very High	Very Good	Very Low	Medium	Best	[23]
Saliency Map	Pixel gradient importance	Low	Poor	Very High	Low	Noisy	[20]

Method	What it Shows (Breast Ultrasound)	Tumor Localization	Boundary Clarity	Noise	Computational Cost	Clinical Suitability	References
LRP	Pixel relevance map	High	Good	Medium	High	Complex	[14]
Integrated Gradients	Smooth attribution map	High	Good	Low	Medium	Similar	[24]
Occlusion	Patch removal effect	Medium	Medium	Low	Very High	Too slow	[21]
RISE	Random mask heatmap	High	Good	Low	Very High	Heavy	[19]
Attention-based	Attention heatmap	Medium	Medium	Low	Low	Model-specific	[14]
Counterfactual	Change lesion explanation	Low	Poor	Low	High	Not localization	[15]

TABLE II: Comparative Analysis of Explainable AI (XAI) Methods for Breast Ultrasound Tumor Localization

Key contributions of TABLE II include:

Modality-Specific Focus: Exclusive focus on breast ultrasound, accounting for challenges such as low contrast, speckle noise, and irregular tumor boundaries [7], [9].

Comprehensive Method Comparison: Structured evaluation of multiple XAI techniques, considering both interpretability and computational aspects [14], [15].

Integration of Clinical Relevance: Connects technical performance with clinical usability, supporting radiologists in practical decision-making [7], [14].

Actionable Reference: Serves as a consolidated guide for researchers by bringing together diverse XAI methods into a single comparative framework [14], [15].

4.3 Research Gap and Motivation for Unified Comparison

Despite the increasing use of XAI in medical image analysis, several limitations persist in the current literature. Many existing surveys adopt a modality-agnostic approach. They cover imaging techniques such as MRI, mammography, and CT, without focusing specifically on breast ultrasound [10], [14]. However, ultrasound imaging presents unique challenges. These include speckle noise, low contrast, and irregular lesion boundaries, which are not adequately addressed in generalized studies [7], [9]. As a result, their applicability to ultrasound-based tumor localization remains limited.

Lack of Systematic Comparison**:** Most studies evaluate individual XAI techniques such as Grad-CAM, Grad-CAM++, LIME, and SHAP. However, they do not provide a structured comparison across key metrics like tumor localization, boundary clarity, noise robustness, and computational cost [14], [15].

Clinical Interpretation Gap: Existing work primarily focuses on technical explanations and visualization outputs. Clinical usability is often not considered. This limits the ability of radiologists to rely on these methods in real-world diagnosis [7], [15].

Absence of Unified Reference Table: There is no consolidated framework that compares multiple XAI methods specifically for breast ultrasound. The lack of such a unified table makes it difficult for researchers and practitioners to select appropriate techniques [14], [15].

Motivation for Our Work:
To address these limitations, this study presents a comprehensive comparative framework through a unified summary table (TABLE II), accompanied by a systematic evaluation of 13 XAI techniques specifically applied to breast tumor localization in ultrasound imaging. The key contributions of this work are outlined as follows:

A breast ultrasound-specific survey addressing modality-specific challenges.

Comparative evaluation of multiple XAI methods, considering localization, boundary clarity, noise, computational cost, and clinical suitability.

Integration of clinical relevance with technical performance, aiding radiologists in practical adoption.

A single consolidated reference (TABLE II) for researchers to quickly identify suitable XAI methods for breast tumor detection.

By explicitly highlighting these gaps and linking them to our contribution, we create a logical transition from method review to research novelty, ensuring the reader understands the significance of the unified XAI table in advancing both research and clinical practice [7]–[15].

5. Discussion and Insights

The comparative analysis reveals several important insights into different XAI techniques used in breast ultrasound imaging. It focuses on tumor localization, boundary clarity, noise sensitivity, and computational cost. These factors directly influence how effectively the results can be interpreted in clinical practice.

Since this study is survey-based, the findings rely on previously reported results. No new experiments or model evaluations are performed.

Key Observations:

Gradient-Based Methods: Grad-CAM and Grad-CAM++ show strong tumor localization performance. They also produce clear boundaries with low noise levels [13], [23]. Their computational cost is moderate, which supports practical use. Guided Grad-CAM improves boundary sharpness further. However, it increases computational requirements [13]. Overall, these methods are well-suited for clinical interpretation. They provide visually reliable tumor maps for radiologists.

Perturbation-Based Methods: LIME and SHAP explain model decisions through feature importance. However, their effectiveness for tumor localization is limited. LIME produces coarse boundaries, while SHAP requires high computation [21],[22]. Score-CAM generates reliable heatmaps using activation maps and forward-pass scores. It reduces noise and provides good localization with moderate computational cost [19]. RISE improves localization performance. However, it is computationally expensive [19]. This limits its use in real-time clinical settings.

Backpropagation-Based Methods: Saliency Maps generate simple pixel-level explanations. However, they are highly sensitive to noise [20]. LRP produces more structured and meaningful maps. Yet, it involves complex computation [14]. Integrated Gradients produce smoother attribution maps and improve interpretability, although clinical understanding may still be challenging [24].

Model-Specific CAM and attention-based methods depend on the model architecture. They offer moderate localization performance. These methods are computationally efficient. However, boundary clarity is often limited [13], [14]. This reduces their effectiveness in clinical applications.

Clinical Implications:

The best balance of localization, clarity, and usability is observed with Grad-CAM++ and Grad-CAM, making them prime candidates for interpretability in clinical breast ultrasound analysis [13], [23].

Methods like LIME, SHAP, and RISE, although powerful for feature attribution, may be less practical in routine screening due to high computation or poor boundary delineation.

Integrating these XAI methods into diagnostic pipelines can enhance radiologist trust and reduce reliance on black-box AI systems, even without new experimental testing [15].

Limitations of the Survey:

This work does not perform experiments or new model evaluations; findings are based solely on reported literature.

The effectiveness of XAI methods may vary with dataset characteristics, imaging quality, and tumor types, which should be considered when interpreting TABLE II.

6. Conclusion

This survey addresses an important gap in explainable artificial intelligence (XAI) for breast ultrasound imaging. It provides a unified comparison of commonly used explanation methods. Unlike prior studies, this work focuses specifically on breast ultrasound. It considers modality-specific challenges such as speckle noise, low contrast, and irregular tumor boundaries. The paper systematically reviews gradient-based, perturbation-based, and backpropagation-based approaches. It highlights their strengths and limitations for tumor localization and interpretation.

The primary contribution is a structured comparison of 13 XAI techniques. The analysis is based on clinically relevant factors such as localization accuracy, boundary clarity, noise sensitivity, computational cost, and interpretability. Furthermore, this survey introduces a single consolidated comparison table (TABLE II) that serves as a unified reference for researchers and clinicians when selecting appropriate explainability techniques for breast ultrasound analysis. The structured comparison also highlights that gradient-based localization approaches, particularly Grad-CAM and Grad-CAM++, provide improved tumor localization and better visual interpretability compared to model-agnostic and perturbation-based methods reported in previous literature [13],[23].

This survey brings together fragmented research into a single framework. It improves understanding of how different XAI methods perform in clinical scenarios. The study also supports the integration of interpretable AI into breast tumor detection systems. This can enhance transparency and increase clinician confidence in automated decisions.

Future work should focus on extending XAI analysis to multi-modal breast imaging. It should also focus on improving robustness to ultrasound-specific noise. Further validation through radiologist-centered studies is necessary to assess real-world usability [7], [21], [22].

7. References

[1] World Health Organization, “Breast cancer,” WHO Fact Sheet, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] S. Malvia, S. A. Bagadi, U. S. Dubey, and S. Saxena, “Epidemiology of breast cancer in Indian women,” *Asian Pacific Journal of Cancer Prevention*, vol. 18, no. 12, pp. 3317–3321, 2017.

[4] R. M. Mann, K. Balleyguier, C. Baltzer, *et al.*, “Breast MRI: EUSOBI recommendations for women's information,” *European Radiology*, vol. 25, no. 12, pp. 3669–3678, 2015.

[5] M. L. Giger, “Machine learning in medical imaging,” *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 512–520, 2018.

[6] A. Dhungel, G. Carneiro, and A. P. Bradley, "Fully automated classification of breast tumors in ultrasound images using deep learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 421–433, 2018.

[7] R. M. Mann, C. Balleyguier, C. Baltzer, *et al.*, "Breast ultrasound: Current status and future directions," *European Journal of Radiology*, vol. 116, pp. 84–96, 2019.

[8] T. A. Berg, J. Zhang, and D. Lehrer, "Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk," *JAMA*, vol. 307, no. 13, pp. 1394–1404, 2012.

[9] A. Athanasiou, A. Tardivon, and M. Tanter, "Breast ultrasound: A review of techniques and clinical applications," *Insights into Imaging*, vol. 1, no. 3, pp. 197–206, 2010.

[10] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[11] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[13] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[14] S. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[15] A. Holzinger *et al.*, "What do we need to build explainable AI systems for the medical domain?" arXiv:1712.09923, 2017.

[16] F. Ciompi *et al.*, "Automated classification of pulmonary nodules using 3D CNNs: Evaluation on clinical datasets," *Medical Image Analysis*, vol. 18, pp. 560–574, 2014.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[18] Z. Zhou *et al.*, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis*, 2018.

[19] H. Wang *et al.*, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE CVPR Workshops*, 2020.

[20] D. Smilkov *et al.*, "SmoothGrad: Removing noise by adding noise," arXiv:1706.03825, 2017.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[23] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[24] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 3319–3328.