Explainable Artificial Intelligence for Breast Cancer Detection: A Review of Deep Learning, Multimodal Imaging, and Interpretable Diagnostic Frameworks

Monika Lamba[1], Deepak Gupta[2], Shashi Gupta[3]

[1,3] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia; [2] Maharaja Agrasen Institute of Technology, India

Email ID : missmonikalamba@gmail.com, drdeepakgupta.cse@gmail.com, shashigupta@lincoln.edu.my,

Abstract: Breast cancer is a significant global health problem and early detection with accurate diagnosis is required. Computer-aided detection (CAD) systems based on Machine learning (ML) and Deep Learning (DL) have been developed. However, CAD systems suffer from poor interpretability which limits clinical translation. This manuscript provides an overview of Explainable Artificial Intelligence methods for breast cancer detection with medical imaging datasets and molecular marker datasets. Topics include multimodal deep learning, hybrid CNN-Transformer models, genetic algorithm design of interpretable ML models, and transfer learning coupled with explainability algorithms (Grad-CAM, LIME, Integrated Gradients, Occlusion Analysis). Classification accuracies of over 99% can be reached when using XAI-enabled deep learning models, without compromising transparency. Explainability techniques also allow clinicians to identify which regions of medical images are diagnostically relevant, improving trust and interpretability of AI models. Clinical decision support and precision oncology are potential application areas where XAI-integrated CAD systems can be utilized to their strengths. Integration of multimodal learning, federated learning, and image genomics represents a future direction for scalable, interpretable, and clinically deployable breast cancer detection solutions.

Keywords: machine learning; deep learning; decision making; breast cancer; explainable; artificial intelligence.

Introduction

Breast cancer is one of the most frequent cancers worldwide. It is a leading cause of morbidity among women worldwide. According to statistics released by the International Agency for Research on Cancer, almost 700,000 deaths resulted from breast cancer worldwide in 2020. Breast cancer has become one of the most fatal cancers among women. Timely and accurate diagnosis can help initiate treatment measures earlier, which can lead to better survival rates. The conventional techniques include radiological analysis of scans followed by biopsy analysis and clinical examination. However, it is time-consuming and is subject to human error.

The advancements in Artificial Intelligence have revolutionized every industry, including healthcare. Machine learning and deep learning models help identify intricate patterns present in medical data, assisting doctors to recognize anomalies in patients' health. Approaches such as Convolutional Neural Networks outperform traditional techniques when processing mammography, ultrasound, MRI scans, and histopathology images. Transformer-based models have also been very successful in this domain.

AI-based diagnostic models often exhibit high predictive performance. However, most of these models are “black boxes” and non-transparent, making it difficult for clinicians to understand their decision-making processes. The “black box” nature of these systems poses challenges to reliability, accountability, and clinical trust. XAI can solve the black-box problem by making predictions of the AI understandable by humans. Gradient-weighted Class Activation Mapping (Grad-CAM), lime, and Integrated Gradient are all approaches that could be implemented to generate explanations for predictions made by models.

Grad-CAM and similar techniques can produce visual heatmaps that highlight important regions in the image for decision-making.

Researchers have developed CAD frameworks that incorporate both high-performance deep learning architectures and explainability techniques. This paper aims to provide a review on explainable AI-based breast cancer detection frameworks using multimodal learning techniques and hybrid deep learning models.

Related work

Singh and Patnaik (2026) suggested multimodal medical imaging as a possible solution. Their approach involved the use of MammXAI, an explainable multimodal deep learning framework for breast cancer classification tasks. This framework consists of EfficientNetV2-Small as the backbone network for feature extraction, followed by a custom-designed multi-head transformer block for encoding long-range spatial information, as well as a capsule network for retaining the spatial hierarchy of multimodal medical images. The suggested framework, called ETCapsNet, achieved state-of-the-art accuracy of 0.996 compared to baseline models such as VGG16, VGG19, ResNet50, and ConvNeXtTiny. The multimodal MammXAI dashboard employs various XAI methods such as Grad-CAM, ScoreCAM, SmoothGrad, Integrated Gradients, and LIME for intuitive visual explanations [1].

Convolutional neural networks combined with transformer models have also been explored to learn both local and global features in medical images. Abdelsabour et al. proposed a hybrid deep learning framework combining MobileNetV1 convolutional neural network and transformer networks for multi-class BI-RADS breast cancer image classification task [3]. This work proposed dual-stream architecture that utilizes MobileNetV1 as a fine-grained spatial feature extractor and transformer module as global context learner. Following feature-level fusion strategy with bagging-based logistic regression classifier further improved stability of prediction. Evaluation on King Abdulaziz University Breast Cancer Mammogram Dataset reported classification accuracy, sensitivity, and specificity higher than 99%. Visualization was performed with Grad-CAM and Grad-CAM++ methods [2].

Feature optimization approach has also been applied to develop models that are interpretable. Ansari et al. (2026) developed a hybrid model combining Genetic Algorithms (GA) optimization technique with machine learning classifiers such as Random Forest, Decision Tree, Naive Bayes Classifier (NB), Support Vector Machine (SVM), XGBoost Classifier, and K-Nearest Neighbors. The GA determines optimal feature selection for classification accuracy while minimizing the number of features used. Further, to address explainability of prediction models, LIME was used to report explanations for individual predictions. The authors reported that upon using GA for feature selection, accuracy scores for Random Forest and XGBoost classifiers increased to 99.12% [3].

Transfer learning approach can be effectively applied to medical imaging problems when there is insufficient data to train a model from scratch. Islam et al. (2026) demonstrated an explainable deep learning approach for breast tissue classification based on expression of Biglycan (BGN) biomarker on immunohistochemical images. The authors compared several transfers learning-based CNN architectures including EfficientNet-B0, DenseNet-161, and ResNet-50, finding that EfficientNet-B0 outperformed other models with precision of 0.97, accuracy of 99, recall of 1.00, and F1-score of 0.99. Class Activation Mapping (CAM) visualization with Grad-CAM helped highlight diagnostically relevant regions of breast tissue associated with prominent biomarker staining patterns [4].

AI-based algorithms have been integrated into most breast imaging modalities to improve lesion detection accuracy. According to Mondal et al. (2026), AI enabled computer aided detection systems assist radiologists in achieving better cancer detection rates with fewer reads. Current challenges that AI developers face include inadequate annotated datasets for algorithm training and validation, poor performance in dense breasts and uninterpretable AI decision-making processes. Techniques such as federated learning, multimodal AI and privacy preserving AI have the potential to overcome current limitations and help to develop clinical decision-support tools [5-6].

Table 1. Comparative Analysis of Existing Approaches

Study	Methodology	Dataset Type	XAI Techniques	Accuracy
Singh & Patnaik (2026) [1]	EfficientNetV2 + Transformer + Capsule Network (ETCapsNet)	Multimodal imaging	Grad-CAM, Grad-CAM++, ScoreCAM, SmoothGrad, Integrated Gradients, Occlusion, LIME	99.60%
Abdelsabour et al. (2026) [2]	MobileNetV1 + Vision Transformer + Bagging LR	Mammogram dataset (KAUBC)	Grad-CAM, Grad-CAM++	>99%
Ansari et al. (2026) [3]	GA-optimized ML (RF, SVM, XGBoost, etc.)	Tabular clinical data	LIME	99.12%
Mondal et al. (2026) [4]	AI/DL models (CNN, imaging-based systems)	Multimodal (mammography, MRI, ultrasound, genomics)	Not specified (focus on transparent AI)	
Islam et al. (2026) [5]	Transfer Learning (EfficientNet-B0, DenseNet, ResNet)	Biomarker pathology images (Biglycan dataset)	Grad-CAM	99%
Joshi & Patel (2026) [6]	ML, DL, IoT, RNN, CNN-based models	Multiple medical datasets	Not specified	
Abdullakutty et al. (2024) [7]	Multimodal DL (encoder-decoder, GNN, attention models)	Histopathology + genomic + clinical	Grad-CAM, SHAP, LIME, attention-based XAI	
Hassan et al. (2026) [8]	Multimodal fusion (transformers, GNN, ensembles)	Multimodal datasets	Grad-CAM, SHAP, attention weights	

Razzaq et al. (2026) [9]	ML/DL + XAI (ensemble & hybrid models)	Imaging + genomic + clinical	SHAP, LIME	
Dash et al. (2026) [10]	Multimodal DL with attention-based fusion	Mammography + MRI + clinical datasets	Grad-CAM++, SHAP	
Khurana et al. (2026) [11]	Domain-wise XAI review in cancer care	Imaging + genomics + multimodal	SHAP, LIME, Grad-CAM, counterfactuals	
Sultana & Ishaq (2025) [12]	Deep learning (CNN-based CAD systems)	Multi-image modalities (MRI, ultrasound, mammography)	Not specified	
Mubasshira et al. (2026) [13]	AI models (SVM, CNN, RF, XGBoost) for treatment	Imaging + clinical data	Limited mention of XAI	
Sharma & Sharma (2026) [14]	Multimodal DL (FNN + GRU)	Breast Cancer Wisconsin dataset (tabular + text)	SHAP	96.49%

Discussion

This review shows that research on XAI for breast cancer detection has rapidly evolved from traditional machine learning models to hybrid DL and multimodal approaches. The summary of reviewed works reveals that hybrid models combining CNN and transformer architectures, as well as capsule networks, show promising results in extracting both local and global features for breast cancer identification. Furthermore, multimodal models that integrate imaging, clinical, and genomic data outperform single-modality models in terms of robustness and generalizability.

All reviewed works make some contribution towards making DL models more interpretable by adopting XAI techniques such as Grad-CAM, LIME, SHAP, and Integrated Gradients. Furthermore, explainable breast cancer detection models help reveal diagnostic-relevant regions and features, which help radiologists trust and adopt these AI-based tools. Contrary to common belief, there is no evidence that making these models explainable reduces their accuracy.

While significant progress has been made in developing explainable breast cancer detection models, there are limitations in the field. First, there is still a lack of public large-scale multimodal datasets needed to train and test generalizable models. Second, hybrid deep learning and multimodal models are often more computationally expensive and less interpretable than traditional machine learning models. Furthermore, using multiple XAI techniques can create a trade-off between model interpretability and usability. Third, there are still limitations in deploying these models into real-world clinical settings due to regulatory restrictions, ethical considerations, and lack of usability.

For future work, researchers can explore federated learning approaches to overcome data privacy and scarcity and develop better multimodal fusion techniques that can integrate genomic data and biomarkers

for precision cancer medicine. Moreover, research can be done on developing intrinsically interpretable models and clinician-in-the-loop approaches.

Conclusion

Breast cancer is one of the most lethal diseases if it is detected in late stages. Therefore, there is an increasing demand for interpretable, accurate, and swift breast cancer detection systems. Even though many artificial intelligence-based solutions have been proposed which outperform traditional methods, most of these deep learning approaches are deployed as black-box models which prevent them from being used in real-time clinical settings. Thus, this paper aims to summarize the recent literature concerning Explainable Artificial Intelligence in breast cancer detection systems that leveraged hybrid deep learning models, multimodal learning frameworks, transfer learning and data optimization methods along with explainability integrated AI models like Grad-CAM, LIME, and SHAP. According to our findings hybrid and multimodal deep learning techniques can yield state-of-the-art accuracy often above 99%. Explainable AI methods can help vastly improve transparency and interpretability of black-box models for clinical usage. Multimodal systems outperform their unimodal counterparts by making predictions based on information obtained from multiple sources. Lastly, we found that adding explainability to AI models did not decrease model performance while allowing clinicians to trust the predictions made by the model. Some research gaps include lack of public multimodal datasets, computational complexity of proposed models which limit real-time deployment, and lack of model integration into clinical practice and regulatory bodies. Researchers investigate distributed learning techniques like federated learning to remove privacy concerns, improving multimodal fusion approaches, looking into other forms of data such as biomarkers and gene expression, and designing models that are intrinsically interpretable.

References

1. S. K. Singh and K. S. Patnaik, "MammXAI: An XAI integrated adaptive multi-model deep learning approach for breast cancer detection using multi-modality images", Biomedical Signal Processing and Control, vol. 113, p. 109173, 2026.
2. I. Abdelsabour, A. Elgarayhi, M. Sallah and M. Elmogy, "Different BI-RADS breast cancer diagnosis using MobileNetV1 and vision transformer based on explainable artificial intelligence (XAI)", Scientific Reports, 2026.
3. Z. A. Ansari, M. S. H. Ansari, A. Khan, H. Pant, S. Fahad and P. V. H. Prasad, "Explainable breast cancer diagnosis: integrating genetic algorithms with LIME-based machine learning", Evolutionary Intelligence, vol. 19, no. 1, p. 11, 2026.
4. J. Mondal, M. M. Rahman, J. Ferdush, M. M. H. Parvez, M. N. Uddin and L. Akter, "Artificial Intelligence (AI) for Breast Cancer Detection: Trends, Challenges, and Future Directions", in Nano Theragnostics in Breast Cancer: Advances, Challenges, and Future Prospects, Singapore: Springer Nature, pp. 707–750, 2026.
5. M. M. Islam, N. Akter, M. Assaduzzaman, M. M. H. Shimul and R. K. R. Sarker, "Explainable AI for breast cancer detection: Biglycan biomarker classification with transfer learning", Intelligence-Based Medicine, p. 100340, 2026.

6. S. Joshi and U. Patel, "Breast cancer diagnosis with AI technologies", Journal of Multiscale Modelling, 2026.
7. F. Abdullakutty, Y. Akbari, S. Al-Maadeed, A. Bouridane, I. M. Talaat and R. Hamoudi, "Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis", Frontiers in Medicine, vol. 11, p. 1450103, 2024.
8. M. M. Hassan, A. Tahsin, M. G. R. Alam, D. Alzamil, S. Garg, M. Z. Uddin and G. Fortino, "Explainable multimodal fusion for breast carcinoma diagnosis: A systematic review, open problems, and future directions", 2026.
9. N. Razzaq, A. Rasool and P. Fatima, "A systematic review of machine learning and explainable AI in breast cancer detection and diagnosis: From black-box models to interpretable clinical decision support systems", in Journal of Conferences Proceedings Publication, Jan. 2026.
10. S. Dash, L. Bewoor, Y. Dongre, A. Bhosle, K. Patil, S. Jadhav and B. Walia, "Explainable multi-modal deep learning for transparent cancer diagnosis: integrating radiology, clinical features, and decision visualization", Frontiers in Artificial Intelligence, vol. 9, p. 1767612, 2026.
11. D. Khurana, R. Kaur, R. K. Roul and S. Batra, "Explainable artificial intelligence in cancer care: A domain-wise review of adoption, challenges and opportunities", Expert Systems, vol. 43, no. 4, p. e70238, 2026.
12. A. Sultana and M. Ishaq, "Breast cancer diagnostics with deep learning schemes using multi-image modalities", in International Conference on Advances in Smart Computing and Applications, Cham: Springer Nature Switzerland, pp. 148–161, Feb. 2025.
13. Mubasshira, Rahman, M. M., Mondal, J., Parvez, M. M. H., Uddin, M. N., & Akter, L. (2026). Artificial Intelligence (AI)-Assisted Treatment of Breast Cancer. In *Nano Theragnostics in Breast Cancer: Advances, Challenges, and Future Prospects* (pp. 659-705). Singapore: Springer Nature Singapore.
14. Sharma, N., & Sharma, S. (2026). Explainable multimodal deep learning for breast cancer diagnosis. In Recent Advances in Computational Methods in Science and Technology (pp. 205-210). CRC Press.