INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

OPEN ACCESS

Received: 01/07/2025

Accepted: 30/09/2025

Published: 23/12/2025

Citation: Vinay Kumar V, Arunkumar TR (2025) Explainable AI for Mammography: A Review. Indian Journal of Science and Technology 18(46): 3675-3686. https://doi.org/10.17485/IJST/v18i46.1193

* **Corresponding author**.

arun@rcub.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Explainable AI for Mammography: A Review

V Vinay Kumar[1], T R Arunkumar[2]*

1 Research Scholar, Department of Computer Science, School of Mathematics and Computer Sciences, Rani Channamma University, Vidyasangama, PB-NH-4, Bhutaramanahatti, Belagavi, Karnataka, India
2 Assistant Professor, Department of Computer Science, School of Mathematics and Computer Sciences, Rani Channamma University, Vidyasangama, PB-NH-4, Bhutaramanahatti, Belagavi, Karnataka, India

Abstract

Background: Breast cancer is one of the top causes of death for women worldwide with an incidence rate of about one in eight women. Mammographic screening allows earlier detection and offers the best chance of cure. The advances in artificial intelligence through Explainable AI, Machine Learning and Deep Learning have bettered diagnostic capacities but mammography remains the primary method for screening. **Objectives:** To introduce a structured comparative framework mapping XAI methods against datasets, model architectures, and reported performance which is absent in prior studies. Furthermore, the study incorporates emerging paradigms such as curriculum learning, federated learning, and case-based reasoning in mammography interpretation, which have not been jointly reviewed before. **Method:** This systematic scoping review examines the use of XAI in breast cancer identification and risk estimation. We performed systematic searches on Scopus, IEEE Explore, PubMed, Springer nature and Google Scholar following a systematic search strategy. The search was conducted through 2019 to early 2025, in peer-reviewed studies of XAI methods in datasets of breast cancer. **Findings:** The implementation of Federated learning with hybrid ML methods enhances accuracy while preserving data privacy, while SHAP and LIME provide feature-based interpretability. Deep learning models like XGBoost and CNNs achieve high classification performance, and Grad-CAM aids radiologists by visually localizing tumours for better decision-making. **Novelty:** This review offers a comparative, metric-driven synthesis of XAI methods in mammography. It summarizes the comparison of datasets, model performance, and interpretability of the models. It also introduces emerging trends such as federated learning, curriculum learning, and human-centric AI, providing new perspectives on connecting algorithmic breakthroughs to clinical needs. **Significance:** XAI enhances the transparency, interpretability, fairness, and trustworthiness of the AI-enabled health system and medical devices, and more importantly, quality of care and health outcomes.

Keywords: Artificial Intelligence; Breast Cancer; Deep Learning; Explainable Artificial Intelligence (XAI); Grad-CAM; Interpretability; LIME; Machine Learning; SHAP; XGBoost

1 Introduction

Breast cancer stands as the most common cancer among female patients across the world while developing into one of the significant causes of mortality through cancer. The disease begins in breast duct or lobule epithelial cells before being detected through physical examination or imaging methods. The timely recognition of breast cancer proves vital because it leads to better treatment success rates together with longer patient survival expectations[(1)]. New diagnostic technologies have shown progress, but the ongoing detection problems demand better performance from and intelligence in screening methods.

Mammography remained essential for breast cancer screening due to its capability to find breast cancer indicators including micro-calcifications and architectural distortion patterns in early stages. Interpretation of mammograms is influenced by inconsistent viewing styles among radiologists which produces incorrect diagnosis results leading to delayed treatment or superfluous biopsy procedures[(2)]. New computational diagnostic tools currently develop to enhance radiologist's ability in providing both precise and uniform interpretation of medical images.

Over the past couple of decades, mammography has made great strides in machine learning (ML) which determines meaningful properties from large datasets. Machine learning algorithms, including support vector machines, random forests and logistic regression, have been employed to classify lesions, estimate the risk of malignancy and rank cases for review[(3)]. Through pattern recognition and probabilistic reasoning, these data-driven models help to reduce diagnostic errors and improve clinical decision-making.

Recent progress in computer aided detection (CAD) in mammography has been made possible by the ability of deep learning (DL) especially convolutional neural networks (CNNs), to automatically extract as well as classify feature directly from raw image. Whereas traditional ML approaches perform poorly on tasks like lesion detection and breast detection, DL models outperform them[(4)]. However, the trade-off of these models for large annotated datasets produces good generalizability and robustness in clinical practice.

More specifically, artificial intelligence (AI) can be viewed as the converged techniques of machine learning (ML) and deep learning (DL) and also offers a holistic basis with which to characterise a medical informatics system using artificial intelligence, that manages the automated, intelligent interpretation of data, and carrying out automated tasks in medical imaging. AI systems assist in pre-processing the mammograms, anomaly detection and technical support in diagnostic decisions in mammography. ML and DL have been integrated within AI pipelines to scalable, high throughput screening solutions with reduced workload and increase of diagnostic efficiency[(5)].

Even though AI models perform exceedingly well, their nature as "black boxes" has raised questions about how much clinicians can really trust them and whether we can hold anyone accountable for the decisions they make. As a potential remedy, the field of Explainable AI (XAI) has developed to enhance the transparency of these models and thus their trustworthiness, providing solutions that allow human users to better understand the reason behind the predictions made by the models. XAI methods like SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and plain old case-based reasoning are now in common use to interpret the outputs of these models[(6)].

The medical domain easily integrates data-driven investigations when it comes to AI applied to mammography, with the healthcare industry incorporating XAI methods actively in their AI workflows which help generate trustworthy and explainable explanations. Research results introduce hybrid models integrating ML algorithms and XAI methods to enhance explaining the clinical support[(7)]. Time distributed model training between institutes can be performed though Federated XAI frameworks that also preserves data privacy among entities[(8)]. Grad-CAM with remarkable maps shows the image areas that influence predictions by providing visual explanations that are in line with radiological logic[(9)]. Curriculum learning methods combined with patch-based techniques which reduce label data set requirements, strive to achieve explainer and interpretability standards[(10)].

1.1 Research Contributions:

This research work presents a comprehensive review on the research contributions related to mammography-based breast cancer detection and ensures explainability contributed through XAI. By synthesising a wide range of current studies (2019–2025) it offers an in-depth survey over methodological trends, datasets and architectures as well as explainability tools. This paper showcases the fusion of state-of-the-art deep learning models like Inception-ResNetV2, VGG along with hybrid ensemble frameworks elucidated by explainability techniques SHAP, LIME, Grad-CAM and Deep Taylor Decomposition to improve interpretability which in turn increases diagnostic confidence.

1.2 Research Gaps:

Research in Explainable AI (XAI) as it pertains to breast cancer detection remains an emerging topic but highlights a number of important gaps. Extend XAI methods for better interpretability of models, especially in combination with the positive and negative image classes; In addition to that tackle limited interpretable data problem in spatial data. The major problem is the data, which is still a serious bottleneck when it needs extensive pre-processing and augmentation to form large and diverse well-annotated dataset with multiple mammographic views of different races or age groups, from various hospitals and radiologists. To achieve this accuracy, we can further explore innovative deep learning approaches that leverage curriculum learning strategies as well as some of the most advanced algorithmic breakthroughs like neural networks.

1.3 Research Questions:

Some of the possible research questions which raised during the study are as follows:

- What improvements can be made in Explainable AI techniques to increase interpretability, clinical trust and decision transparency for breast cancer detection?
- How does greater diversity in datasets will improve model generalizability, fairness, and robustness?
- How federated Learning can be used along with other strategies to tackle productivity, data imbalance & bias challenges in medical AI apps?
- To what extent can standardized interpretability metrics and the latest XAI tools assist to better facilitate such collaboration between clinicians and AI, for improved diagnostic outcomes?
- How can advanced deep learning architectures, curriculum learning, and hybrid ensemble models be optimized for mammographic analysis?

2 Review Methodology

This review was conducted using systematic approach through extraction of literature from databases like PubMed, Scopus, IEEE Xplore, ScienceDirect, MDPI, Web of Science, DOAJ, ACM Digital Library, DBLP, Google Scholar, and SpringerLink along with additional references included from conference proceedings, arXiv preprints and official health organization reports.

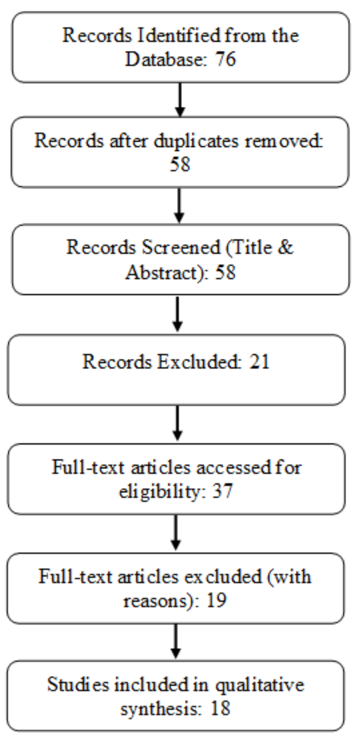

Fig 1. PRISMA workflow for study selection

The study was conducted with a total of 76 papers, after deleting 18 duplicate papers 58 papers were screened by referring abstract and title of the paper. Later 21 papers were excluded due to inapplicability to the current research. Remaining 37 eligible articles full texts was used for based on exclusion and inclusion criteria. After that 19 papers were excluded due to the following reasons: lack of XAI aspects (4), papers without experimental results – review papers (4), insufficient methodology and dataset (3), studies focused on image data not related to breast tumour (3), robust evaluation metrics (1), validation by peers in the form of a publication (2) and studies not focused on image dataset (2). Finally, 18 papers were selected which is relevant to my study by applying the inclusion criteria and using the keywords: "Breast Cancer", "Explainable Artificial Intelligence", "XAI," "Mammography", "Deep Learning", "SHAP", "LIME", and "GradCAM". The search spanned the years from 2019 to 2025. The inclusion criteria for the original research studies that applied XAI methods to assist in breast cancer diagnosis with mammography and established interpretable machine learning or deep learning models along with metrics for performance measurement and visualisation output relevant to clinic. Main parameters taken into account for concept development were related to dataset attributes, model architectures and eXplainable AI (XAI) methods, performance standards and clinical applicability.

3 Overview

The utilization of machine learning (ML) in combination with deep learning (DL) characteristics has significantly enhanced the reliability as well as the accuracy and effectiveness of mammography breast cancer screening techniques. Standard readings of mammograms cause radiologists to have inconsistency of readings that may cause false positive findings and a negative diagnosis. Mammography is a well-known application where support vector machine combined with random forests or logistic regression models has been successfully applied to analyse mammography images and detect the suspicious lesions and their potential malignancy[3]. Some researchers have achieved mammographic image analysis transition from DL with the assistance of CNN's modules, the automated feature detection and the enhancement of breast-unrelated identification tasks[4]. These models also exhibit potential for the discovery of ambiguous visual patterns that human radiologists cannot directly see for the radiologists to make confident decision-support[9].

XAI is an important part of the diagnostic tool on breast cancer in mammography by making AI predictions available for medical experts through a transparent explanation. Science has utilized some of these XAI techniques such as SHAP, LIME, Grad-CAM, as well as case-based reasoning to produce explanations for ML/DL model predictions Table 1.

Table 1. Different XAI Techniques used in Mammography

References	Dataset	XAI + ML & DL Method	Objective	Key Findings / Results	Limitations / Remarks
[3]	WBC, WDBC	SHAP, PDP, Permutation Importance.**ML** - SVM, ANN, and RF are utilized.	Classify and interpret breast cancer using model-agnostic methods	**WBC:** KNN - 97.7% accuracy & 98.2% precision **WDBC:** ANN - 98.6% accuracy & 94.4% precision Key features identified for malignancy prediction	Use of Model-agnostic tools will enhance usability and transparency in clinical workflows. Getting scalability issue for large datasets with SHAP methods. Feature importance analysis is a constraint in using XAI technique.

Continued on next page

Table 1 continued

[4]	Public dataset (ADMANI dataset of 1,048,345 examinations and 629,863 patients.The training set consists of 11,913 examinations, with 486 cancer cases.The dataset includes 28,911 instances of breast cancer)	CAM and Bounding boxes.**CNN Architectures:**VGG19, VGG16, and AlexNet	The paper illustrates the importance of augmentation techniques and data pre-processing for attaining high classification performance. Enhance classification using DL + synthetic data	Applying synthetic augmentation will improve model performance with CAM provided insights	Augmentation will boost DL reliability. Interpretability tools support visualization. Future investigation could explore using both positive and negative image classes. Limited classification performance due to low image resolution. Need for improved data pre-processing and augmentation techniques. One limitation of the study is that the images were scaled to a resolution of 256*256 and 512*512 pixels, which may have limited the classification performance.
[5]	Multiple (review paper)	Grad-CAM, ML, & DL	Review of DL and XAI methods for breast imaging ML for lesion detection and classification. DL methods for breast cancer risk prediction.	Summary of strengths and limitations of ML, DL, data augmentation, and XAI	Comprehensive survey. Useful in analysing trends in breast imaging using XAI. Search across multiple datasets on breast anatomy. Emphasis on overcoming data challenge using augmentation techniques. Augment deep learning-based approaches with more annotated databases.
[6]	Real clinical data with 4050 cases and Three public datasets: BCW with 683 cases. MM with 830 cases. BC with cases.	Case-Based Reasoning (CBR) method for visual reasoning. (scatter plot + rainbow boxes)	Provide visual analogical explanations	Competitive accuracy with k-NN of 80.3%. User study confirms clarity.	Bridges XAI and user-friendly visualization. Useful for clinical adoption. Treatment is difficult to distinguishing because of small sample size. Need for larger datasets to increase the classification accuracy.

Continued on next page

Table 1 continued

[7]	Public data from Indonesian women with 200 and without 200 breast cancer patients.	SHAP + ML **ML Methods** XGBoots, Logistic Regression, Random Forest SVM	Decision support tool with interpretability for early detection.	XGBoost performed best; Accuracy Train Data - 81.3% Test Data - 81% SHAP - Helped identify key features	Effective at individualizing risk factors. Supports early intervention. Analysing the imbalance of cases observed in healthy patients and those with breast cancer in the real world
[8]	WBC Dataset - 568 WDBC Dataset - 699	Federated Learning. ML Algorithms – SVM, RF, and XGBoost. SHAP (local computation in FL)	Enhance privacy-preserving explainability in diagnosis using XAI. Privacy risks in Federated Learning. Increase the interpretability and transparency of the model	Achieved 97% accuracy in both datasets for ANN as well as XGBoost within Federated Learning. Maintained privacy.	Shows high potential for federated settings without compromising model interpretability
[9]	CBIS-DDSM & DDSM	Grad-CAM. 3D Inception-ResNet V2	Integrate explainability into 3D CNN for classification	High performance;Accuracy - 98.53%,Recall - 98.53%,Precision - 98.40%,F1-score - 98.43%,AUROC - 0.9933,Efficacy of 3D visual explanations through Grad-CAM emphasizing the efficiency of distinction between benign and malignant scenarios.	Introduction of 3D CNN-based XAI in breast MRI. Further research will be directed at model validation and expansion. Address issues of interpretability, data imbalance, and generalisation. Lack of interpretability in deep learning models.
[10]	Internal -1976 External - 4276 images	Grad-CAM for explainable artificial intelligence. Patch-based (curriculum 20, curriculum 40, and curriculum 100 models) deep learning model for mammogram analysis.	Decrease annotation requirements under preserving of XAI in mammography	F1 is increased by curriculum learning from 80.55 to 83.95%. Grad-CAM explains predictions	Strong Generalization and Scaling with Distillation and Limited Supervision.
[11]	CBIS-DDSM	Deep Taylor Decomposition method. LRP-Epsilon technique. Guided Backprop method. LRP Preset A-Flat method. De-ConvNet method.	Early detection & Classification. Enhance trust and understanding of CNN models for DCIS detection	LRP and Deep Taylor more effective for neuron-level interpretations.	Shows clinical validation with experts. Limited to small test set.
[12]	Breast-XD (791 solid mass lesions from 752 different patients.)	SHAP and LIME **Classifiers:** SVM, K-NN, RF, DT, LR, XGBoost.	Compare XAI model vs radiologist in diagnosis	X2GAI model achieved an accuracy of 94.34% accuracy with few errors; better than the radiologists in in some significant areas.	Emphasis on the role of XAI toward avoiding unnecessary biopsies, errors. Include a variety of races and age groups in future studies.

Continued on next page

Table 1 continued

(13)	WBC	Grad-CAM LRP Attention Maps	Multi-class classification with heatmap-based localization	Achieved 92% accuracy; LRP added interpretability for model transparency	Localizes lesions through attention visualization. The network needs to be validated on larger datasets. Explore different deep learning models to improve model accuracy
(14)	WDBC (The dataset was sourced from Kaggle, containing 569 items with 33 features)	SHAP and LIME **ML Methods:** XGBoost, DT, AB, SVM, GB, and KNN.	Interpret ML-based predictions for classification.	LIME/SHAP offers clear local/global explanations. XGBoost: Accuracy - 99.30% in breast cancer classification.	Transparent models. Identifies most influential features. Lacking interpretability with currently used machine learning techniques applied in spatial data. Need to Train model with more data.
(15)	WDBC – 569 images with 212 malignant, 357 benign.	SHAP & LIME. Ensemble model with DT & SVM	Reduce features while retaining accuracy and interpretability. SHAP, LIME, and Skope Rules improved interpretability significantly.	Ensemble methods outperformed decision trees in feature importance. **DT:** Precision - 0.882 Accuracy - 0.9033 F1-Score - 0.8718 **Ensemble DT & SVM:** Precision - 0.9858 Accuracy - 0.9561 F1-Score - 0.9436 **SHAP** - revealed critical variables. **LIME** - Benign tumor output predicted with 0.99 probability	Simpler model without damaging performance. Study example-based XAI methods compared to SHAP and LIME. Investigate more complex machine learning methods such as Neural Networks for further improvement in performance ratios.
(16)	The study used a sample of 164 female patients undergoing Core Biopsy.	SHAP, LIME, PDP, and ICE. ANN	Model interpretation in neural networks, Risk prediction and classification	Enhanced model interpretability. LIME explanations aligned with clinical judgment	Enhances transparency in deep models

Continued on next page

Table 1 continued

[17]	Clinical dataset from Bangladesh of 500 patients collected from Dhaka Medical College & Hospital with 254 noncancerous and 246 cancerous cases	SHA + ML**ML Methods** -DT, RF, LR, NB, XGBoost	Cultural-context ML with XAI, Regional data-specific diagnosis Five machine learning methods were evaluated for classification accuracy.	SHAP analysis was applied to interpret XGBoost model predictions. XGBoost achieved the highest accuracy of 97%. Random Forest accuracy was 96%. Decision Tree accuracy was 91%. Naive Bayes accuracy was 94%. Logistic Regression accuracy was 93%.	Tackles data bias and advocates for XAI in underprivileged areas. Model adaptation for demographics of interest. Lack of a large and diverse dataset hinders generalization. Investigate newer machine learning and deep learning algorithms. ML for early detection of breast cancer. SHAP analysis interpreted model predictions and understand feature effects.
[18]	Mini-MIAS dataset with 322 mammograms from 161 patients.	Decision Tree, Random Forest	Fully automated interpretable CAD system. It employs interpretable methods for breast cancer detection and diagnosis.	95% accuracy; 100% precision/specificity on mini-MIAS	Visual descriptions, at each stage. Novel TbGC segmentation used. Absence of common metrics to evaluate interpretability on real-world data. Absolutely no mammogram dataset is available for thorough evaluation. The model will be validated on other datasets.

The authors are able to preserve their patient privacy as well as for localized explanations [8] and the system yields a higher prediction accuracy and is more interpretable which is ideal for early diagnosis, decision support [7] with the adoption of hybrid ML-XAI systems. Such case-based reasoning allows health practitioners to compare new cases to previously encountered ones and provides interpretability through developing an acceptable conceptual understanding of why decisions were made by AI systems [6].

The fusion of deep learning models specifically convolutional neural networks (CNNs) with explanation techniques Grad-CAM and attention maps enables the localization of the crucial regions in mammogram that influenced model prediction outcomes [9], [11], [13]. The image content will also help radiologists more confidence of the AI deductions and has educational value in medical environment. Curriculum-based learning allows patch-level explanation to reduce annotation workloads, as well as offering accurate and easy-to-interpret diagnosing results [10]. XAI-augmented systems output clinically meaningful explanations for their diagnoses at performance levels comparable to human radiologists based on direct comparisons between human and XAI-augmented systems output [12].

However, the interpretable ML models have been being tried with the goal of reducing for clinical use the input features in the classification process to make the diagnostic process open [15] [16]. Specific research on Bangladeshi community may inform on how XAI methods being used elsewhere can be adapted to address particular health requirements of a local population [17]. Joint APIs provide the Input APIs an ML/DL/Detection model which uses a computational system like XAI to maintain the quality of a medical diagnosis but without a trust-breaking problem like that medical AI of AI in Medicine. These findings collaborate to develop better breast cancer screening systems that are safer, better and open.

Explainable AI models have become integral elements of the mammography approach to breast cancer diagnosis not only to increase model explainability, but also to gain clinician trust. There are presently two of extremely used interpretation methods in the community, SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). These approaches provide visual explanations as well as analysis of individual features, enabling predictions to be easily understood by radiologists. The actual application of XAI allows gaining trust, transparency and ensuring traceability, as well as improved decision-making by breast cancer patients during diagnosis Table 2.

Table 2. Different Methods & Modalities of XAI

XAI Method	Model Type	Purpose / Output
SHAP (SHapley Additive Explanations) [7], [8], [14], [15], [17]	ML (Random Forest, XGBoost) / Federated Learning	Feature-attribution to prediction; explainable risk scores
LIME (Local Interpretable Model-Agnostic Explanations) [7], [14], [16]	ML models (SVM, Logistic Regression)	Local decision explanation with surrogate models
Grad-CAM (Gradient-weighted Class Activation Mapping) [9], [11], [13]	CNN / 3D Inception-ResNet V2	Visual heatmaps to indicate relevant regions in mammograms
Attention Mechanisms [10], [13]	Deep Learning / Patch-based CNNs	Spatial localization and visual description of region of interest
Case-Based Reasoning [6]	Visual Case Retrieval System	Interprets AI predictions by analogical past cases
Curriculum Learning + Patch-based XAI [10]	DL with reduced annotation effort	Interpretable patch-level decisions with efficient training
Feature Selection for Minimal XAI [15]	ML models (Decision Trees, SVM)	Simplified by translating them into input variables for transparent decision rules.
Federated Learning + XAI (SHAP) [8]	Distributed ML across institutions	Allows users to ensure privacy-preserving explainability in multicenter settings
Rule-Based Logic + XAI Visualization [18]	Hybrid ML-XAI pipelines	Produces interpretable decisions based on clinical reasoning
Model Performance Comparison (XAI vs Radiologists) [12]	Explainable AI vs Human Experts	Assess diagnostic performance and interpretability

SHAP allows a clinician to visually perceive the specific contribution of the input variables in the classification decision through feature-level attributions and can operate within federated learning settings for secure inter-institutional data sharing initiatives[7],[8],[14]. The LIME feature do partitioning of the black box models to simpler surrogate models for individual prediction explanation in clinical work[7],[14],[16].

The interpretability in deep learning-based mammography analysis provides an efficient visual explanation through integrating Grad-CAM (Gradient-weighted Class Activation Mapping) and attentional mechanisms in its interpretative approaches. Fine-grained level data from Grad-CAM creates visual heatmaps, which visually indicates where influences the prediction in classification and localization tasks[9],[11],[13]. The large number of convolutional mappings and attention mechanisms in CNN structures focus on significant diagnostic regions through the patch-format structures[10],[13]. Curriculum learning methods are another major development that reduces the dependence on annotations and generate interpretable recommendations using patch-based visual output[10].

Reasoning methods are not only visual explanation tools, but also robust diagnostic prediction methods, because the latest AI models are now incorporating case-based reasoning methods with little feature selection and rule-based logic systems. The CBR technique is helpful for medical diagnosis by allowing the doctor to see old cases, that are similar to the new ones, and it, therefore, helps improving both his understanding and his acceptance[6]. The quality of the predictive model depends on the use of important variables and as well as on retaining significant prediction ability based on feature selection methods[15]. XAI techniques integrated with rule-based logic resulted in successful models that simulate clinical decision making based on the findings in[18]. Research has shown that XAI - based models produce similar diagnostic outcomes as radiologists due to the existing evidence that interpretable outputs in XAI have the potential to help expert radiologists' decision in breast cancer screening[12].

3.1 Results and Findings

This review work outline, based on the necessity of transparent AI clinical systems that adhere to accountability expectations, we find the increasing adoption of XAI technologies in mammography-based breast cancer diagnosis[3] [6]. Some works used

SHAP and LIME to interpret as Random Forest, XGBoost and SVM machine learning classifiers [7] [8] [12] [14]. These techniques have demonstrated their capability of generating meaningful explanations ranging on the local and global description level. When applying SHAP and Permutation Importance Saliency Maps, it reached a classification accuracy of 97.7% (KNN) and 98.6% (ANN) matched with important features towards each Malignancy Prediction [3]. The diagnosis tools can prove highly beneficial for AI diagnostic trust and transparency but are plagued by issues of runtime efficiency and scalability across large datasets. [4] The study showed that data augmentation combined with VGG16 and AlexNet deep learning were successful in obtaining high-effective mammography detection results. Image resolution restriction in addition to the problem of class imbalance were considered as significant challenges toward this direction. The work by [6] followed a visual case-based reasoning approach to yield users the benefit of analogical interpretability that was empirically demonstrated in-user in line with a new way for clinicians/AI collaboration.

A study that could carry this out is the Federated Learning technology that allows both high privacy and accuracy during the construction of secure diagnostic networks as the work in [8] illustrates. Inception-ResNet V2 is a key 3D deep learning model that has [9] been used to study MRI scans because it can identify both diagnostic indicator as well as regional anatomical structures through Grad-CAM visualization. Two other approaches to improve annotation efficiency and control feature selection, namely patch-based curriculum learning that shares favourable results in [10] and hybrid system construction shown in [7]. Three key bottlenecks that still hinder DL model deployment are the difficult model interpretability pattern, imbalanced data and the lack of a unified interpretability criterion. The results of [12] confirm the critical necessity of XAI model evaluation within real world clinical practice by showing better diagnosis results than radiologists in some cases. The ability of the XGBoost algorithm stands out as most superior with 99.30% accuracy in breast cancer classification, and SHAP analysis to explain and interpret model predictions generated by the XGBoost model predictions [14]. Combining SHAP into LIME based evaluation method, we achieved a significant improvement of model interpretability and made transparent decision more powerful Table 3.

The study discussed obstacles like imbalanced data, small datasets, interpretability, generalisation, and collaboration gap between AI-Radiologist. These obstacles will make it difficulty of adopting XAI methods in diagnosing breast cancer for real world. But the study listed the outcomes by integrating quantitative metrics like F1 score, accuracy, precision, recall and AUC curve across different studies to reveal the trade-off between interpretability and performance prediction to make the readers to understand where and how the present XAI methods improve in clinical workflows.

3.2 Ethical, Legal and Regulatory Considerations

The ethical and regulatory challenges of implementing XAI in mammography range from bias reduction, data governance, accountability to clinical acceptance.

- ***Bias Mitigation:*** Bias in datasets needs to be examined through fairness audits, subgroup performance reports and mitigation techniques. The study suggests conducting regular bias audits with subgroup performance measures like sensitivity, and specificity across age, ethnicity and imaging device, data augmentation and re-sampling where appropriate, and prospective testing on representative cohorts.
- ***Data Governance:*** Data privacy and security demand solutions like enforcing regulations such as GDPR and HIPAA, federated learning, differential privacy, and transparent data-use agreements. Provide dataset provenance records such as Datasheets or DAIR-style metadata, enforce requirements for informed consent or ethical waivers, and have clear data access rules. This paper discusses federated learning as well as secure enclaves in terms of privacy-preserving schemes but considering the risks due to communication and aggregation.
- ***Accountability & Transparency:*** Accountability requires clear documentation in Model Cards, Dataset Datasheets and incident response procedures, and relevant regulation requires clarification of liability as well as certification and safety measures.
- ***Clinical Governance & Risk Management*:** Ultimately, human factors including clinicians training, trustworthiness and actionable explanations are needed to facilitate its real-world adoption to maintain iteratively transparent, ethical and reliable XAI systems for the clinic. A governance framework like multidisciplinary AI oversight committee is suggested for the pre-deployment clinical trials in order to access the domain-specific tasks and risk classification.
- ***Regulatory & Legal Context:*** The study briefly covers effects of data protection laws, medical device regulation, and emerging AI-specific regulation. It is emphasized that regulators will usually require that any such explainability is proportional to the risk involved, and there are growing calls for evidence of post market surveillance regarding performance drift.

Table 3. Summary of comparing performance metrics, interpretability, and clinical utility XAI techniques

XAI Technique	Performance Metrics (Accuracy/AUC)	Interpretability Strength	Clinical Utility	Limitations
SHAP	High Accuracy - up to 99%	Global + Local Explanations	Explains Feature Importance, Trusted in Diagnostics	High Computational Cost, Scalability Issues
LIME	Moderate Accuracy with 90-95%	Local Explanations	Quick, Lightweight Explanations for Radiologists	Less Stable on Complex Data
Grad-CAM	Preserve resolution in tumour segmentation and Provide visual explanations with high accuracy	Visual Heatmaps (Region-Specific) Grad-CAM aids interpretability	1. Tackles resolution loss and pixel imbalance. 2. Promotes visualization of deep models via Heatmaps	1. Test approaches on diverse medical image datasets. 2. Address elevated false positive rates in CAD systems.
Deep Taylor Decomposition	Strong Performance on CNN Models	Feature-wise Relevance Mapping	1. Detailed Attribution for Neural Networks. 2. Demonstrates clinical validation with experts.	Limited to small test set. Requires Large Training Data
Curriculum Learning	Improved Learning Efficiency	Adaptive Learning for Complex Data	Optimized Training for Diverse Mammography Views	Needs Broader Validation Across Datasets
Attention Mechanisms	Achieved 92% accuracy on lesion localization tasks (AUC up to 0.96)	Visual attention maps highlighting tumour regions	Aids radiologists in focusing on critical regions	1. The proposed network requires testing on larger datasets for validation. 2. Investigate various deep learning architectures to enhance model accuracy
Case-Based Reasoning	The best classification accuracy achieved was 80.3% using kNN	Visual analogical explanations + Case-based reasoning for interpretability	Supports clinical decision-making with case analogies. Useful for clinical adoption.	Difficulty in distinguishing treatments due to small sample size. Need for larger datasets to improve classification accuracy.
Curriculum Learning + Patch-based XAI	Curriculum learning improves F1 from 80.55 to 83.95% with patch-based models	Adaptive learning with localized heatmaps	Improves efficiency for mammography with diverse views	1. Patch diversity and computational cost constraints 2. Strong generalization and scalability with limited strong supervision
Federated Learning + XAI (SHAP)	Achieved 97% accuracy while preserving privacy using Federated Learning	SHAP values ensuring Enhance model interpretability and transparency	Enables cross-institutional collaboration without sharing raw data	Federated setup complexity and communication costs

4 Conclusion

XAI plays a major role in improving the transparency and interpretability of machine learning and deep learning models that are used to detect breast cancer from mammograms. XAI makes these algorithms more clinically acceptable. SHAP, together with LIME and Grad-CAM, which couple with attention mechanisms, have proved very effective at providing clear interpretations along the lines of clinical decision-making which help make sense of the model's predictions. Their use in the diagnostic process is often coupled with maintaining excellent accuracy levels. The clinical practice of XAI has yet to reach its full potential due to a number of continuing gaps that are waiting to be resolved. Current models aren't very generalizable because the datasets they use are so narrow, which causes problems when you try to apply them to different populations or when you use different imaging modalities. There are no standardized metrics in the present analysis for evaluating the clinical value of the explanations produced by XAI systems. Experienced physicians with clinical expertise cannot be matched by feature-based or visual XAI strategies. Thus, when creating better systems like human-in-the-loop (HITL), XAI should rely on feedback from medical professionals. Requirement for immediate and thorough examination of ethical and regulatory challenges of consents from patients and AI decision accountability in order to address them properly. These challenges could lead to confuse for the AI decision makers, both human and computer, and ultimately for our society. They could also lead to significant and negative costs on our healthcare system.

Future research should focus and struggle to gain performance benefits as well as access for clinical use with explainable functionalities, for mixed population and mixed scanner fully automatic computer-aided diagnosis. The study identified the opportunities like using federated learning, curriculum learning, XAI in multimodal pipelines, and human-centric AI will provide new perspectives on connecting algorithmic breakthroughs to clinical needs.

References

1) Hasan AH, Khalid UAR, Abid MA. Leveraging Artificial Intelligence in Breast Cancer Screening and Diagnosis. *Cureus*. 2025. https://doi.org/10.7759/cureus.79177.
2) Dang LA, et al. Impact of artificial intelligence in breast cancer screening with mammography. *Breast Cancer*. 2022;29(6):967–977. https://doi.org/10.1007/s12282-022-01375-9.
3) Khater T, et al. An Explainable Artificial Intelligence Model for the Classification of Breast Cancer. *IEEE Access*. 2025;13:5618–5633. https://doi.org/10.1109/ACCESS.2023.3308446.
4) Prodan M, Paraschiv E, Stanciu A. Applying Deep Learning Methods for Mammography Analysis and Breast Cancer Detection. *Applied Sciences*. 2023;13(7):4272. https://doi.org/10.3390/app13074272.
5) Mendes J, Domingues J, Aidos H, Garcia N, Matela N. AI in Breast Cancer Imaging: A Survey of Different Applications. *Journal of Imaging*. 2022;8(9):228. https://doi.org/10.3390/jimaging8090228.
6) Lamy JB, Sekar B, Guezennec G, Bouaud J, Séroussi B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*. 2019;94:42–53. https://doi.org/10.1016/j.artmed.2019.01.001.
7) Silva-Aravena F, Núñez Delafuente H, Gutiérrez-Bahamondes JH, Morales J. A Hybrid Algorithm of ML and XAI to Prevent Breast Cancer: A Strategy to Support Decision Making. *Cancers*. 2023;15(9):2443. https://doi.org/10.3390/cancers15092443.
8) Briola E, Nikolaidis CC, Perifanis V, Pavlidis N, Efraimidis P. A Federated Explainable AI Model for Breast Cancer Classification. 2024. https://doi.org/10.1145/3655693.3660255.
9) Talaat FM, Gamel SA, El-Balka RM, Shehata M, ZainEldin H. Grad-CAM Enabled Breast Cancer Classification with a 3D Inception-ResNet V2: Empowering Radiologists with Explainable Insights. *Cancers*. 2024;16(21):3668. https://doi.org/10.3390/cancers16213668.
10) Camurdan O, et al. Annotation-efficient, patch-based, explainable deep learning using curriculum method for breast cancer detection in screening mammography. *Insights Imaging*. 2025;16(1):60. https://doi.org/10.1186/s13244-025-01922-w.
11) Ferla ML, Montebello M, Seychell D. An XAI Approach to Deep Learning Models in the Detection of DCIS;vol. 677. 2023. https://doi.org/10.1007/978-3-031-34171-7_33.
12) Oztekin PS, et al. Comparison of Explainable Artificial Intelligence Model and Radiologist Review Performances to Detect Breast Cancer in 752 Patients. *Journal of Ultrasound in Medicine*. 2024;43(11):2051–2068. https://doi.org/10.1002/jum.16535.
13) Giammarco MD, et al. Explainable Deep Learning for Breast Cancer Classification and Localization. *ACM Transactions on Computing for Healthcare*. 2025;6(1):1–18. https://doi.org/10.1145/3702237.
14) Dutta M, Hasan KMM, Akter A, Rahman MH, Assaduzzaman M. An interpretable machine learning-based breast cancer classification using XGBoost, SHAP, and LIME. *Bulletin EEI*. 2024;13(6). https://doi.org/10.11591/eei.v13i6.7866.
15) Rashi A, Madamala R. Minimum relevant features to obtain AI explainable system for predicting breast cancer in WDBC. *International Journal of Health Sciences (IJHS)*. 2022;6(Special Issue 9):1312–1326. https://doi.org/10.53730/ijhs.v6nS9.12538.
16) Yahata E, Santos PDD, Pires MMDS, Suyama R, Simoes PW. Neural Networks and Explainable Artificial Intelligence for Breast Cancer Prediction and Classification. *Procedia Computer Science*. 2025;256:1159–1166. https://doi.org/10.1016/j.procs.2025.02.224.
17) et al Islam. Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. *Scientific Reports*. 2024;14(1):8487. https://doi.org/10.1038/s41598-024-57740-5.
18) Moroz-Dubenco C, Bajcsi A, Andreica A, Chira C. Towards an interpretable breast cancer detection and diagnosis system. *Computers in Biology and Medicine*. 2025;185:109520. https://doi.org/10.1016/j.compbiomed.2024.109520.