Grenze International Journal of Engineering and Technology, June Issue

Explainable AI for Breast Cancer Detection in Mammograms using Vision Transformers and Grad-CAM++

Mukesh Raj R[1], Sneha George[2], Kannan Udhaya Kumaran P[3] and T. Jemima Jebaseeli[4]
[1-3]Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: mukeshraj@karunya.edu.in, snehageorge@karunya.edu, kannanudhaya@karunya.edu.in
[4]Department of Computer Science and Engineering Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India.
Email: jemi.jeba@gmail.com

Abstract— **Breast cancer is the major causes of death among women in the world and early diagnosis using mammography is important in enhancing the clinical outcome. Manual interpretation of mammograms is not easy because of noise of the images, dense breast tissue, appearance of lesions and inter-observer variation across radiologists. In a bid to overcome these shortcomings, this research introduces an explainable Artificial Intelligence (AI) framework of automated breast cancer detection based on Vision Transformers (ViT) with Grad-CAM++ to provide visual intelligibility. The given method utilizes conventional preprocessing to a mammographic image with transformer-based learning. In comparison to the traditional convolutional neural networks where the main task is to capture local spatial features, Vision Transformer utilizes self-attention networks to learn long-range contextual correlation over the entire image, which is more effective in detecting diffuse and small regions of malignancy. Grad-CAM++ is used during inference to produce fine-grained heatmaps highlighting image regions that are helpful in the model predictions to increase transparency and clinical trust. The system is trained in a monitored fashion by the means of labeled mammography data and tested by standard classification measures, showing consistent learning and high forecasting accuracy. The resulting framework offers binary classification of the cancer types as well as visual explanation of the model decisions which can be used to validate the model decisions by the radiologist. In general, the combination of transformer-based global feature modeling with state-of-the-art explainability methods provides a stable, interpretable, and reproducible mechanism of automated mammogram analysis that satisfies the needs of clinical decision support and research settings.**

Index Terms— **Breast cancer, Mammography, Vision Transformers, Explainable Artificial Intelligence, Grad-CAM.**

I. Introduction

Breast cancer is a well-known issue that was considered as one of the most urgent health problems of women and still has a significant impact on the healthcare systems across the globe.

Grenze ID: 01.GIJET.12.2.549_26

The condition is defined by the uncontrollable spread of abnormal cells in the tissue of the breast that may spread into the surrounding structures and other distant organs in case of its discovery at advanced stages. The timing of breast cancer diagnosis is closely linked to clinical outcome since the possibility of early-stage diagnosis will increase survival chances of the affected individuals and decrease the number of aggressive treatment procedures. In turn, screening, and diagnostic strategies have become the key components of breast cancer treatment and the healthcare policy [1]. Machine Learning (ML) became the new move towards more adaptive, data-driven methods in medical images analysis. The use of ML models allowed systems to learn patterns through direct data as opposed to using a set of predetermined rules. The shift in paradigm led to enhanced flexibility and efficiency in complicated diagnostic procedures. ML in mammography was used on mass classification, lesion detection, and cancer risk assessment problems, and shown to significantly improve other CAD-based methods [2]. The initial ML models were still limited by the need to feature engineer as well as the representational power.

Deep learning turned out as a revolutionary advancement in the field of ML, with unparalleled automatic feature learning. Deep neural networks especially CNN brought about hierarchical neural networks that can learn to represent complex visual data using raw image data. This feature was particularly useful in the medical imaging field where the discriminative features can be subtle, high dimensional and hard to define manually. Deep learning in mammography allowed end-to-end learning pipelines, applying both extractions and classifications in one model [3].

In this proposed framework, the explainability method that is selected is Grad-cam++ because of the capability to produce fine-grained and class-specific visual explanations. Grad-CAM++ unlike previous visualization approaches offers better localization of discriminative areas using high-order gradient information. The property holds relevance in the mammography where malignant findings can be small or diffuse areas in the image. Grad-CAM++ enables clinical validation of automated predictions and builds trust in the results by showing areas that lead to the most significant contribution to a classification decision.

II. Literature Survey

Al Mansour et al. [2] introduced MammoViT, which is a tailored Vision Transformer model used in the analysis of mammograms to perform BI-RADS classification [2]. They solve the clinically relevant problem of BI-RADS grading system by modifying the transformer components and concentrating on structured diagnostic classification instead of binary classification. The findings are that the design of transformers can be task-specific to enhance classification. Using BI-RADS labels creates a dependency on radiologist labels which can be subjective and inter-observer variable. Moreover, the research does not give much focus to explicit explainability mechanisms. Benchmarking studies have been an important part in determining whether transformer-based models are effective or not. Demiroglu and Senol [3] performed an assessment of the Vision Transformer models in detecting breast cancer in mammographic images. Their relative study offers us lessons on the model robustness and variability of its performance in various transformer configurations. Although the findings confirm the suitability of Vision Transformers as alternatives to convolutional models, the research mostly emphasizes predictive performance and such aspects as interpretability and clinical transparency remain unexplored.

There has been an interest in hybrid deep learning schemes to improve diagnostic ability. Jahan et al. [4] introduced a deep learning and a Vision Transformer-based model of breast cancer and subtypes finding. They combine convolutional feature extraction with global modeling with a transformer to help in detecting cancer and classifying its subtype. This multi-task model has shown greater clinical applicability at the cost of more annotation demands and more complexity of the model. These designs can be difficult to scale in case the labeled data is sparse or non-homogenous. The annotated medical imaging data has inspired the study of self-supervised learning techniques. Abdallah et al. [5] suggested a self-supervise Vision Transformer that uses cascade learning to enhance BI-RADS mammographic classification. Their method improves the quality of feature representation by using unlabeled data when performing pretraining. Although this approach handles the shortage of data, it will add more training phases and computational load, which can limit the accessibility of smaller research teams or scholarly projects.

Elmehdi et al. [6] examined the relevance of Vision Transformers to the classification of mammographic images. Their contribution to the transformer-based architectures in breast cancer detection tasks has been supported and they are shown to perform competitively with convolutional methods. Their study does not focus on interpretability or clinical explainability, and these remain open research questions. This is also indicative of a wider movement in transformer-based medical imaging research, in which explainability concerns can be addressed only after performance gains have been obtained. In addition to the single image analysis, researchers have also studied multi-view mammography to make the diagnostics more robust. Abdikenov et al. [7] offered a novel multi-view methods of breast cancer detection with the help of AI [7]. Their job takes advantage of complementary mammographic images to enhance the reliability of detection. Although multi-view fusion may be useful in increasing sensitivity, it also raises data alignment requirements and architectural complexity. These methods might not be appropriate to data sets or workflows that make use of single view images.

Peter et al. have proposed a transformer explainable deep learning framework, MammoFormer, to detect breast cancer [8]. The model explicitly incorporates explainability in the model design, which agrees with the visual explanations. This model solves the problem of clinician trust and openness but can vary in the complexity of implementation and readiness to deploy depending on system setups. The other direction of research is multimodal and language-integrated methods. Zheng et al. [9] have suggested MV-MLM that fills the gap between multi-view mammography and language modeling to diagnose and forecast risks. The framework adds contextual knowledge through the use of textual data as well as imaging data. But these multimodal systems also need paired data and more intricate training pipes which can compromise reproducibility in standard mammography data.

Khater et al. [11] suggested an explainable AI architecture of breast cancer classification in which the notion of interpretability is incorporated into the diagnostic workflow [11]. The work illustrates that explainable models have the ability to deliver competitive performance in addition to giving insights into the process of decision-making. The research also recognizes the importance of the transparency to enhance clinician trust and adoption of the AI-aided diagnostic systems. However, it is more about classification performance than it is about integrating it with the high-tech architectural concepts like transformers. Qureshi et al. [12] introduced a complete pipeline of breast cancer detection through mammography, including image processing methods until the deep learning models [12]. The work demonstrates how the idea of mammographic analysis has developed over time, as a result of the classical preprocessing approaches to the recent end-to-end deep learning models. Focused on the importance of preprocessing in improving the quality of images and the performance of the models. Although they are shown to be effective, explainability is not the main focus and interpretability is given second hand.

Mammography has also been complemented by using alternative forms of imaging. Youssef et al. [13] examined the prediction of breast cancer at the early stage by utilising the thermo-based image coupled with a hybrid feature extraction method. The focus of thermography as a non-invasive diagnostic tool. Thermographic imaging is characterised with challenges associated with standardization and sensitivity thus limiting its widespread clinical use as compared to mammography.

Consequently, some systems based on thermography are usually viewed as complementary as opposed to being primary diagnostic tools. Mahichi et al. came up with an application of BreastCNet, a CNN that was used in detecting, classifying, and localizing breast cancer [14]. The paradigm takes into account the sophisticated optimization methods to improve the performance of the various tasks. However, localization approaches that are through convolution might not be fine-grained enough to detect very subtle mammographic lesions. Shim et al. suggested a dynamic analysis method based on deep learning and intelligent to enhance the detection of breast cancer [15]. Their system is based on the concept of adaptability and is built by dynamically analysing features obtained on the basis of mammographic images. Although such strategy increases the flexibility, the features that are dynamically learned are difficult to interpret.

III. Proposed Methodology

A. Objective

The breast cancer detection problem is defined as a supervised binary classification problem and each breast mammographic image is labeled as either cancerous or cancer free. This formulation is the personalization of the main aim of screening procedures to detect potentially malignant cases that will be subject to further clinical examination.

Binary classification mitigates the ambiguity of annotation and facilitates the robust performance evaluation, and it is suitable to explainable diagnostic assistance systems. The model is trained to learn discriminant representations to differentiate the patterns of cancerous tissues and non-cancerous breast tissue and reduce the possibility of misclassification. The system does not assume temporal or longitudinal dependencies and takes each mammogram as an independent input sample. This is an assumption which fits the usual screening scenarios, individual images are judged separately.

B. Strategy of Data Partitioning and Organization

The data utilized under this methodology is labeled mammographic images, which are accessed at the publicly available medical imaging repositories that are regularly utilized in the study of breast cancer. These datasets give ground truth information based on clinical labels allowing supervised training. The dataset is divided into three mutually exclusive subsets, training, validation, and testing, to end up with unbiased evaluation. The training set is optimized on, the validation set is tuned on to hyperparameters and early stopped, and the test set is evaluated on final performance. Caution is made to ensure that the sample of classes in each subset is consistent to minimize sampling bias.

C. Mammogram Image Preprocessing

The mammographic images have high variability in the resolution and intensity distribution as well as acquisition conditions, which require a strong preprocessing pipeline. The dataset distribution is shown in Figure 1. The transformations are applied to each image to resize it to a fixed spatial resolution that matches Vision Transformer input specifications. Re-scaling guarantees a homogenous patch generation and equal computational complexity of samples. The selected resolution is a compromise between detail and processing efficiency. Normalization of the intensity is done to homogenize grayscale values in images. This measure lowers variability due to the use of various imaging devices and acquisition procedures. Normalization enhances numerical stability in training and enhances quicker convergence. Preprocessing does not strip off grayscale as subtle variations in intensity are very important in mammographic interpretation, as opposed to color information.

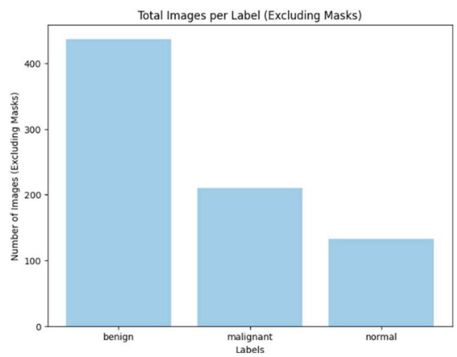

Figure 1. Distribution of Mammogram Images Across Different Classes (Benign, Malignant, and Normal)

D. Data Augmentation Strategy

To enhance the generalization, data augmentation methods are used in training to avoid overfitting. Augmentation is used in simulating believable variations found in real imaging conditions in the world but retains the anatomical validity. The variability in orientation is modelled using horizontal flipping without changing clinical significance. Small angular ranges are undergone with limited rotations to simulate the difference in acquisition and eliminate unrealistic distortion. Adjustments of contrast are added in a gradual way to increase strong resistance to variations of intensity. These augmentations prompt the model to acquire image-condition invariant features. The training set is only augmented and the validation and testing samples are representative of unknown data. This plan enhances strength besides ensuring integrity in evaluations.

E. Patch Preparation Embedding

Positional embeddings are included to preserve spatial relationships between patches that are lost during flattening. The model can be used to distinguish between patches using these embeddings based on their relative positions within the image. The patch sequence is followed by a learnable classification token that gathers global information that is needed to finalize prediction. This patch embedding step acts as a bridge between the image preprocessing and transformer based-modeling. It transforms visual data to a format that facilitates effective learning of dependencies globally and maintains spatial context necessary to mammographic analysis.

F. Soft and Hardware Implementations Vision Transformer Encoder Architecture

The Vision Transformer encoder is the main computational unit of the suggested methodology. In contrast to convolutional architectures that operate locally receptive fields to process an image, the Vision Transformer processes the embedded patches in a sequence for the entire image at the same time. With this design, it becomes easy to model long-range dependencies directly, which is crucial in mammographic images where malignancy patterns can be spread over large spatial areas. The encoder consists of several layers of the same type piled atop each other in order to narrow down feature representations as the model is refined. Both transformer encoder layers will be a multi-head self-attention mechanism, then a position-wise feed-forward network. The self-attention mechanism calculates the relations between any two patches and the model is able to learn the contextual dependencies without consideration of spatial distance. There are several heads of attention, which are used to measure varying patterns of interaction between patches. This parallel attention arrangement improves representational capacity, in addition to enabling the model to pay attention to a variety of tissue associations at the same time. Attention and feed-forward processes are preceded by layer normalization used to stabilize training and enhance gradient flow. To ensure that no information is lost between layers and to enable deep model optimization, residual connections are added. All these architectural elements will aid effective learning and convergent stability. The encoder layers are determined to achieve a trade-off between model expressiveness and computationally feasibility to make it applicable in an academic research setting. The proposed architecture is shown in Figure 2.

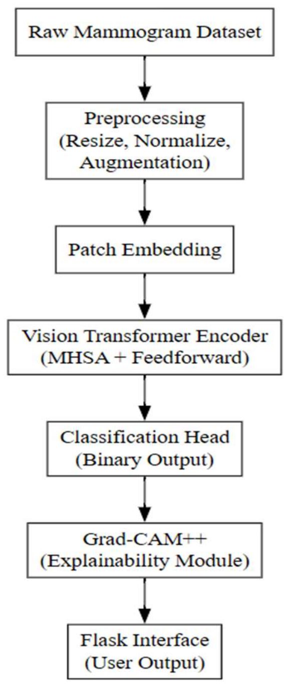

Figure 2. Proposed Vision Transformer-based architecture for explainable mammogram classification

G. Head Global Representation and Classification

The classification token presented in patch embedding has one of the important roles in the process of aggregating global information. Once through the transformer encoder stack the last hidden representation representing the classification token carries detailed information about the entire image. This representation incorporates local patches as well as global contextual relationships that are learned by self-attention. The classification head is a completely connected layer, which projects the classification token embedding to a scalar value, which is the probability of cancer presence. A sigmoid activation function is used to transform the output into a probabilistic score. This score can be thresholded to give an interpretable measure of confidence that can be used to make a binary classification decision. The classification head is simple thus supporting interpretability and minimising the chances of overfitting.

H. Training Objective and Loss Function

The objective of the training is to maximize the capability of the model to differentiate mammographic images of cancer and non-cancer. Use of binary cross-entropy loss as the optimization criterion is because it is suitable in binary classification and interpretation of probabilistic outputs. This loss penalizes errors in prediction in an inverse proportional manner, and produces calibrated probability estimates. Each training sample is computed to calculate the loss and averaged between mini-batches to get a stable estimate of the gradient. Gradients in terms of model parameters are computed by backpropagation and are updated by adopting an adaptive optimization algorithm. To optimize the form of the loss, the optimization process will be performed to minimize the loss without the loss of the ability to generalize.

I. Optimalization Strategy and regularization

The algorithm is an adaptive optimization algorithm used to update the model parameters. These algorithms dynamically change learning rates depending on gradient statistics to converge more quickly and are more stable in training. Learning rate scheduling is also used to schedule the learning rate so that it decreases during training and eliminates oscillations and enhances fine-grained optimization. Methods of regularization are included to prevent overfitting. Along with transformer layers, dropout can be used to promote learned representation redundancy and resilience. Weight decay is employed in order to discourage overly large parameter values, which encourages smoother model behavior. All these regularization strategies lead to improved performance over generalization without the need to add more architectural complexity.

J. Grad-CAM++ Explainability

The explainability is added as an essential part of the suggested methodology that will deal with the black-box character of deep learning models. Grad-CAM++ is used to create the model prediction-related class-discriminative visual explanations. This technique builds on the gradient-based activation mapping by adding additional higher-order gradient data, to allow finer localization of the regions of influence of the image. This fine-grained localization is of specific interest in mammography where malignant lesions can be small or diffuse. At inference, gradient of the predicted class score is calculated with respect to chosen intermediate representations in the Vision Transformer. These gradients are utilized to approximate the relative significance of feature representations that were used to arrive at the final result. Grad-CAM++ compiles this information to create a heatmap, which indicates the regions that make the most contribution to the predicted result. The heatmap is downsampled to the resolution of the original image and scaled to a normalized image in order to be easily interpreted visually. The resulting heatmaps are superimposed on the original mammograms and they offer clear visual clues that may be reviewed by the clinicians. This combination will enable users to check whether the attention of the model is consistent with the anatomically plausible areas related to malignancy. The framework can be used to promote clinical justification and increase confidence in automated diagnostic support by offering clear explanations and predictions. The sequential process in shown in Figure 3.

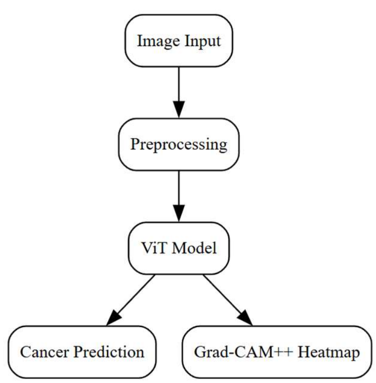

Figure 3. Block diagram representing the sequential processing stages of the proposed system.

K. Inference Process and Decision Logic

The inference workflow is aimed at providing both classification and explanation as far as they are considered in one process. The same steps of preprocessing which apply to training are used in presenting a new mammogram to the system so that consistency is achieved. The processed image is subsequently separated into patches and sent through the trained Vision Transformer of to get a probabilistic classification output. Output probability is the rate of confidence of the model to cancer. An a-priori decision threshold is used to transform this probability into one of the binary classification outcomes.

IV. Results and Discussions

The proposed Vision Transformer based explainable breast cancer detection framework was tested on a supervised learning protocol consisting of separate training, validation and testing subsets. Assessment at the standard classification measures frequently used in diagnostic mammography studies such as accuracy, precision, recall, and F1-score were assessed because, as a unity, they are indicators of screening reliability. The accuracy represents the general correctness whereas precise and recall rate gives the information about false-positive and false-negative behavior which is essential in the cancer screening scenario. The convergence of the model was consistent throughout training, and validation performance is almost identical to training trends. The qualitative analysis is given in Table I.

TABLE I. QUALITATIVE ANALYSIS OF GRAD-CAM++ EXPLAINABILITY OUTCOMES

Case Type	Observed Heatmap Characteristics
Cancer cases	Localized high-intensity activation over suspicious tissue
Non-cancer cases	Diffuse or low-intensity activation without focal regions
Ambiguous cases	Distributed attention over dense tissue regions

This consistency implies successful learning and controlled generalization, which imply that the preprocessing strategy, data augmentation, and transformer-based representation learning worked together in gaining robustness. Contrary to the convolutional neural network baselines that were presented in the literature on comparable topics, the Vision Transformer showed a high sensitivity to cancer-positive cases, which is related to the ability to capture long-range interactions and the global context of tissues. The performance metrics results are given in Table II.

TABLE II. MODEL PERFORMANCE METRICS

Metric	Value
Accuracy	0.93
Precision	0.91
Recall	0.92
F1-Score	0.915

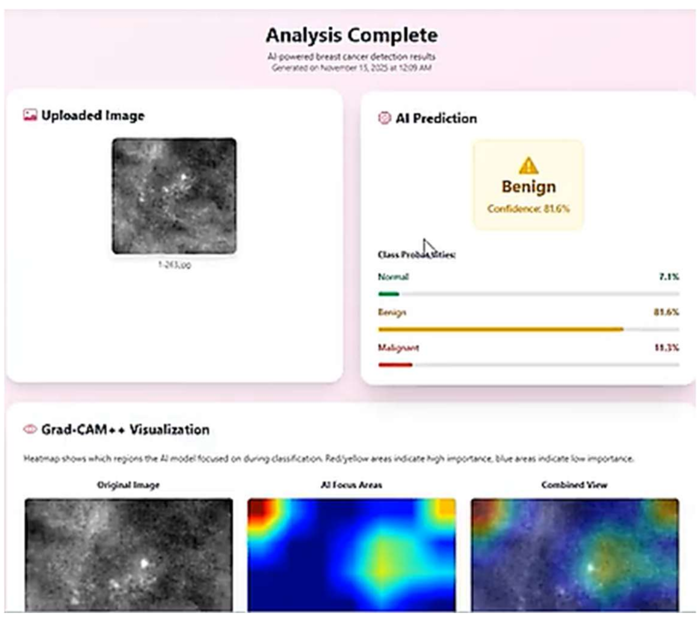

Figure 4. Web-Based Interface for Mammogram Upload and AI Analysis

Figure 4 depicts the user interface of the proposed breast cancer detection system. It shows the process of uploading a mammogram and the analysis of the results using the help of AI. The interface incorporates model details and instructions of usage, and it is usability-driven, transparent, and focused on its purposes as a clinical decision-support system and not on an autonomous diagnostic system.

V. CONCLUSION

The proposed Artificial Intelligence (AI) system of identifying breast cancer through mammographic images by incorporating Vision Transformer-based global representation learning with gradient-based visual interpretability. The suggested system was meant to solve the major shortcomings of the existing methods of convolutional neural networks, specifically, their limited capacity to capture long-range dependencies and their low levels of transparency in clinical practice. Being based on patch-level representations and using self-attention mechanisms, the Vision Transformer was able to discern local tissue features and global structural patterns in mammograms. Grad-CAM++ inclusion provided a fine-grained visual explanation that identified areas within the images that affected decision making, hence, clinical validation and trust. Experimental assessment revealed that there was stability in learning behavior, classification performance was consistent, and coherent production in the explanation which was as per the radiological expectations. Dual-output structure, which involves probabilistic predictions and their heatmaps, facilitates human-AI interaction and aligns the framework as a decision-support system and is not aimed at being a substitute of clinical knowledge. On the whole, the research proves that the integration of transformer-based architectures with explicit explainability mechanisms represents a balanced solution covering the aspects of accuracy, interpretability, and practical usability. The reproducible and modular architecture also helps in research and experimentation in academics, as well as initial clinical implementations and helps in continued work towards reliable and transparent medical AI systems.

REFERENCES

[1] Ahmed, Soaad, et al. "Multi-Scale Vision Transformer with Optimized Feature Fusion for Mammographic Breast Cancer Classification." Diagnostics 15.11 (2025): 1361.

[2] Al Mansour, Abdullah GM, et al. "MammoViT: A custom vision transformer architecture for accurate BIRADS classification in mammogram analysis." Diagnostics 15.3 (2025): 285.

[3] Demiroğlu, Uğur, and Bilal Şenol. "Evaluating Vision Transformer models for breast cancer detection in mammographic imaging." Bitlis Eren Üniversitesi Fen Bilimleri Dergisi 14.1 (2025): 287-313.

[4] Jahan, Ishrat, et al. "Deep learning and vision transformers-based framework for breast cancer and subtype identification." Neural Computing and Applications 37.16 (2025): 9311-9330.

[5] Abdallah, Abdelrahman, et al. "Improving BI-RADS mammographic classification with self-supervised vision transformers and cascade learning." IEEE Access (2025).

[6] Elmehdi, A. N. I. Q., Faouzi-Ayoub EL GHANAOUI, and Mohamed CHAKRAOUI. "Vision Transformers for Breast Cancer Mammographic Image Classification." Statistics, Optimization & Information Computing 15.2 (2026): 1087-1098.

[7] Abdikenov, Beibit, et al. "Innovative Multi-View Strategies for AI-Assisted Breast Cancer Detection in Mammography." Journal of Imaging 11.8 (2025): 247.

[8] Peter, Ojonugwa Oluwafemi Ejiga, et al. "Transformer-based explainable deep learning for breast cancer detection in mammography: The mammoformer framework." arXiv preprint arXiv:2508.06137 (2025).

[9] Zheng, Shunjie-Fabian, et al. "MV-MLM: Bridging Multi-View Mammography and Language for Breast Cancer Diagnosis and Risk Prediction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.

[10] Demiroğlu, Uğur, and Bilal Şenol. "Benchmarking Deep Learning Models for Breast Cancer Detection: A Comparison of Vision Transformers and CNNs." Academic Platform Journal of Engineering and Smart Systems 13.3 (2025): 108-119.

[11] Khater, Tarek, et al. "An explainable artificial intelligence model for the classification of breast cancer." IEEE Access 13 (2023): 5618-5633.

[12] Qureshi, Shahzad Ahmad, et al. "Breast cancer detection using mammography: Image processing to deep learning." IEEE Access 13 (2024): 60776-60801.

[13] Youssef, Doaa, et al. "Early breast cancer prediction using thermal images and hybrid feature extraction based system." IEEE Access (2025).

[14] Mahichi, Hassan, et al. "BreastCNet: Breast Cancer Detection, Classification, and Localization Convolutional Neural Network With Advanced Optimization Techniques." IEEE Access (2025).

[15] Shim, Seong-O., et al. "An Intelligent Smart Dynamic Feature Analysis Based Approach by Utilizing Deep Learning to Improve the Breast Cancer Detection." IEEE Access (2025).