ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 16, Issue, 07, pp.70745-70749, July, 2026
https://doi.org/10.37118/ijdr.31079.07.2026

RESEARCH ARTICLE OPEN ACCESS

CROSS-MODAL ATTENTION-BASED MULTIMODAL DEEP LEARNING FRAMEWORK FOR BREAST CANCER DIAGNOSIS USING MAMMOGRAPHY AND ULTRASOUND IMAGING

[1]*Indu P. K., [2]Dr. Beni, G. and Dr. Rene Dev, D. [3]

[1,2]Dept. of Computer Science and Engineering, Noorul Islam Centre for Higher Education, TamilNadu, India[1]
[3]Dept. of Electronics and Communication Engineering, T John Institute of Technology, Bengaluru, [3]

ARTICLE INFO

Article History:
Received 20th April, 2026
Received in revised form
18th May, 2026
Accepted 25th June, 2026
Published online 30th July, 2026

Key Words:
Breast Cancer Diagnosis, Multimodal Deep Learning, Mammography, Ultrasound Imaging, Cross-modal Attention, Swin Transformer, Medical Image Analysis.

***Corresponding author: Indu P. K.,**

ABSTRACT

Breast cancer continues to be one of the top causes of cancer-related death for women globally, highlighting the significance of prompt and precise detection. Single-modality imaging is often used by conventional computer-aided diagnosis systems, which restricts their capacity to provide additional diagnostic data. In order to diagnose breast cancer automatically, this study suggests a multimodal deep learning framework that combines ultrasound and mammography imaging. The suggested approach learns discriminative representations from heterogeneous imaging modalities by combining an ultrasonic encoder based on Swin Transformer and a mammography encoder based on ResNet101 with a cross-modal attention fusion technique. To increase lesion visibility during preprocessing, contrast enhancement utilizing Contrast Limited Adaptive Histogram Equalization (CLAHE) was used. The classification of benign and malignant conditions was done using the fused multimodal characteristics.Utilizing publicly accessible mammography and ultrasound datasets, an experimental evaluation was carried out utilizing a pseudo-paired multimodal learning approach. The suggested model demonstrated the efficacy of multimodal fusion for breast cancer diagnosis with an accuracy of 88%, an F1-score of 0.87, and an Area Under the ROC Curve (AUC) of 0.903. The findings show that cross-attention greatly enhances diagnostic performance when convolutional and transformer-based representations are combined. Future multimodal medical imaging applications and intelligent clinical decision support systems have great potential with the suggested paradigm.

Citation: Indu P. K., Dr. Beni, G. and Dr. Rene Dev, D. 2026. *"*Cross-Modal Attention-Based Multimodal Deep Learning Framework for Breast Cancer Diagnosis Using Mammography and Ultrasound Imaging"*. International Journal of Development Research*, 16, (07), 70745-70749.

INTRODUCTION

One of the most common cancers in women worldwide, breast cancer continues to be a serious public health issue. Treatment results and patient survival rates are greatly enhanced by early detection and precise diagnosis. Clinical breast cancer screening and diagnosis rely heavily on medical imaging modalities like mammography and ultrasound. While ultrasound imaging offers supplementary soft tissue information and better lesion visibility in dense breast tissues, mammography is very good at detecting calcifications and structural abnormalities. Many current computer-aided diagnosis systems rely on single-modality imaging, despite deep learning's impressive achievement in medical image processing. These methods might not be able to record complimentary data from other imaging modalities. In order to integrate heterogeneous medical imaging data for better diagnostic accuracy, multimodal learning has become a potential research avenue.

Representation learning in medical imaging has been significantly improved by recent developments in transformer topologies and convolutional neural networks (CNNs). While transformers can model global contextual linkages, CNNs are better at capturing local texture and edge-based information. The performance of breast cancer diagnosis may be significantly enhanced by incorporating these structures into a multimodal fusion framework. In order to classify breast cancer utilizing mammography and ultrasound pictures, a cross-modal attention-based multimodal deep learning architecture is presented. The suggested system uses a Swin Transformer encoder for learning ultrasound representations and a ResNet101 encoder for extracting mammography features. Complementary multimodal features are fused using a cross-modal attention mechanism for the classification of benign and malignant conditions.

This study's primary contributions can be summed up as follows

- A multimodal paradigm for diagnosing breast cancer that combines ultrasound and mammography.
- ResNet101 and Swin Transformer encoders are combined in a hybrid CNN-transformer architecture.
- A method for learning complementary multimodal representations using cross-modal attention fusion. A preprocessing pipeline that uses enhancement based on CLAHE to improve lesion visibility.
- A thorough analysis of multimodal classification metrics, such as AUC-ROC.

Related Work

Using medical imaging modalities like mammography, ultrasound, and magnetic resonance imaging, deep learning has shown great promise in the diagnosis of breast cancer. In lesion classification and detection tasks, CNN-based architectures such as ResNet, DenseNet, and EfficientNet have demonstrated impressive results. Because mammography-based diagnosis techniques are so good at identifying calcifications and micro-lesions, they are routinely used. However, in dense breast tissues, mammography might not be enough. Complementary information on tissue properties and lesion shape is provided by ultrasound imaging. A number of multimodal strategies have been put out to increase diagnostic precision by combining various imaging modalities. Conventional fusion approaches frequently depend on early fusion methods or feature concatenation, which might not accurately represent inter-modal interactions. Because transformer topologies may collect long-range contextual information, they have recently attracted interest in medical picture analysis. By utilizing hierarchical attention mechanisms, Swin Transformer has proven to perform well in picture categorization challenges. Despite recent progress, limited research has explored multimodal fusion using cross-attention mechanisms between mammography and ultrasound modalities. This paper proposes a cross-modal attention-based multimodal framework that integrates CNN and transformer representations in order to overcome this constraint.

MATERIALS AND METHODS

Dataset Description: Two publicly accessible medical imaging datasets—the BUSI Dataset for ultrasound imaging analysis and the CBIS-DDSM dataset for mammography analysis—were used to assess the suggested multimodal breast cancer diagnosis paradigm. These datasets were chosen because they offer clinically relevant imaging samples for the classification of benign and malignant breast lesions and are frequently utilized in studies on computer-aided diagnosis for breast cancer.

The suggested system may learn complimentary diagnostic representations from various imaging modalities thanks to the utilization of multimodal datasets. While ultrasonic imaging offers improved view of soft tissue architecture and lesion shape, mammography mainly records structural abnormalities and calcifications. By including these modalities, the suggested deep learning architecture becomes more strong and capable of diagnosis.

Mammography Dataset: Mammography analysis was done using the Curated Breast Imaging Subset of the Digital Database for screening Mammography (CBIS-DDSM) dataset. Deep learning research on breast cancer makes substantial use of CBIS-DDSM, a standardized and curated version of the original DDSM dataset. The dataset includes pathology-confirmed diagnosis labels, expert radiologist annotations, and digital mammography images gathered from clinical screening exams. Important lesion features found in the mammography pictures include, architectural deformities, masses, calcifications, and asymmetries. These characteristics are very important for characterizing lesions and diagnosing breast cancer. Additionally, the dataset offers information on breast density, pathology labels, ROI annotations, and lesion localization**.** Only benign and malignant lesion categories were chosen for binary classification in this investigation. The mammography dataset utilized in this study included shown in Table 1

Table 1. Mammography dataset

Class	Number of Images
Benign	772
Malignant	785

The proportion of benign and malignant samples is comparatively balanced, which improves classification stability and lessens class imbalance during model training. Because it can identify microcalcifications and early-stage abnormalities, mammography is one of the most essential breast cancer screening methods. However, women with thick breast tissue may do worse on mammograms, which is why supplementary imaging modalities like ultrasonography are necessary.

Ultrasound Dataset: Ultrasound imaging analysis was done using the Breast Ultrasound Images (BUSI) dataset. The BUSI dataset is a publicly accessible breast ultrasound imaging dataset that was gathered for studies on computer-aided diagnosis of breast cancer. Because it offers complementing soft tissue information that is frequently difficult to see in mammography images, especially in dense breast tissues, ultrasound imaging is essential to the diagnosis of breast cancer. The BUSI dataset includes - benign lesions, and malignant lesions. Deep learning-based feature extraction can benefit from the dataset's lesion variety and clinically significant textural variations. To ensure consistency with the mammography classification setup, only benign and malignant ultrasound images were used for binary classification studies in this investigation; normal cases were not included.

The ultrasound dataset utilized in this study included shown in Table 2.

Table 2. Ultrasound Dataset

Class	Number of Images
Benign	440
Malignant	210

Important diagnostic data is contained in the ultrasound images, such as: lesion boundaries, internal structural patterns, tissue heterogeneity, lesion texture, and acoustic shadowing. The structural information gleaned from mammography pictures is enhanced by these features.

Binary Classification Configuration. The suggested framework was created as a binary classification method to differentiate between both malignant and benign breast tumours. The definitions of the class labels were: Label 0 for Benign and Label 1 for Malignant Because binary classification is a clinically significant diagnostic challenge for computer-aided diagnostics systems and early breast cancer detection, it was chosen.

Pseudo-Paired Multimodal Strategy: The scarcity of publicly available datasets with same-patient mammography and ultrasound pictures is a significant obstacle in multimodal breast cancer research. This work used a pseudo-paired multimodal learning technique to overcome this restriction. Using this strategy-Malignant mammography images were paired with malignant ultrasound images, while benign mammography images were paired with benign ultrasound images. This method allows for multimodal feature learning while maintaining class-level consistency across imaging modalities.The pseudo-paired strategy allows the network to learn: shared diagnostic traits across modalities, complementary multimodal representations, and cross-modal feature interactions. Before multimodal pairing, dataset balancing was carried out because the mammography and ultrasound datasets had different sample counts. To guarantee balanced multimodal correspondence, the bare minimum of samples across modalities was chosen for every class.The balancing process is summarized below Table 3

Table 3: Balancing process

Class	Mammogram Samples	Ultrasound Samples	Selected Paired Samples
Benign	772	440	440
Malignant	785	210	210

Therefore, the final multimodal paired dataset consisted of , shown in Table 4

Table 4. Final paired dataset

Class	Final Paired Samples
Benign	440
Malignant	210

This method of balancing was beneficial: maintain class consistency across modalities, enhance optimization stability, and lessen training bias. When same-patient multimodal datasets are not available, the pseudo-paired multimodal learning approach—which may be thought of as a weakly supervised multimodal fusion strategy is frequently used.

Image Preprocessing: To increase lesion visibility, lower noise, improve image quality, and standardize the inputs for deep learning model training, image preparation was carried out. Medical pictures often include low contrast, modality-specific artifacts, intensity fluctuations, and acquisition noise. Preprocessing is therefore crucial for enhancing classification performance and representation learning.

The preprocessing pipeline consisted of:

1. CLAHE enhancement,
2. image resizing,
3. And intensity normalization.

CLAHE Enhancement: Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to increase lesion visibility and improve local contrast in both mammography and ultrasound images. CLAHE is a sophisticated contrast enhancement method that targets specific local picture areas rather than the entire image. By avoiding undue noise amplification, this method enhances the visibility of tiny lesion features. In mammography images, CLAHE improves: tissue abnormalities, lesion edges, microcalcifications, and minute density changes. In ultrasound imaging, CLAHE improves: lesion borders, diverse tissue architecture, texture visibility, and auditory patterns. Deep feature extraction capability is improved and diagnostically significant regions are much more visible when CLAHE is used.

Image Resizing: All images was scaled to 224 x 224. Pixels prior to being supplied to the deep learning system, resizing the image was required since the suggested architectures: Swin Transformer and ResNet101,need inputs of a defined size. Standardizing the dimensions of images also enhances batch processing efficiency, lowers computational complexity, and stabilizes network training. The chosen image size provides balance between: computational effectiveness, memory usage, and diagnostic picture detail preservation.

Normalization: Pixel intensities were pre-processed and then standardized to the following range: [0,1] utilizing normalization with min-max. Normalization enhances convergence speed, gradient optimization, and numerical stability during training. Furthermore, normalization improves multimodal representation learning and feature fusion performance by lowering intensity-scale differences across ultrasound and mammography modalities. Prior to multimodal fusion and deep feature extraction, the pretreatment pipeline guarantees that both imaging modalities are standardized.

Proposed Methodology

Overview

The proposed multimodal framework consists of the following stages:

1. Image preprocessing
2. Mammography feature extraction
3. Ultrasound feature extraction
4. Cross-modal attention fusion
5. Classification

Mammography Encoder: A ResNet101 architecture pretrained on ImageNet was utilized for mammography feature extraction. The CNN encoder captures local texture patterns, lesion boundaries, and calcification features.

Ultrasound Encoder: A Swin Transformer architecture was employed for ultrasound feature extraction. The transformer encoder captures global contextual information and hierarchical lesion representations.

Cross-Modal Attention Fusion: A cross-modal attention mechanism was utilized to integrate mammography and ultrasound features. The attention mechanism enables one modality to guide feature refinement in the other modality.

The attention operation is defined as:

Attention(Q,K,V)=Softmax((QK^T)/sqrt(d_k))V

where Q represents query vectors, K represents key vectors, and V denotes value vectors.

Classification Layer: The fused multimodal representations were passed through fully connected layers with ReLU activation and dropout regularization for benign and malignant classification.

Experimental Setup

Implementation Details: The proposed framework was implemented using Python and PyTorch.

Hardware Configuration

- NVIDIA GPU
- CUDA acceleration
- Jupyter Notebook environment

Training Parameters

Parameter	Value
Batch Size	8
Learning Rate	1e-4
Optimizer	AdamW
Scheduler	Cosine Annealing
Epochs	20
Image Size	224 × 224

Evaluation Metrics

The proposed framework was evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- AUC-ROC

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is defined as:

AUC = ∫TPR(FPR)d(FPR)

RESULTS AND DISCUSSION

Quantitative Results

The proposed multimodal framework achieved promising diagnostic performance.

Confusion Matrix

Actual / Predicted	Benign	Malignant
Benign	86	2
Malignant	14	28

Classification Performance

Metric	Value
Accuracy	88%
AUC	0.903
F1-score	0.87
Benign Recall	0.98
Malignant Recall	0.67

The experimental results demonstrate that the proposed multimodal framework effectively integrates complementary information from mammography and ultrasound modalities.

DISCUSSION

The proposed framework achieved strong classification performance, particularly in benign lesion detection. The high AUC score indicates strong discriminative capability of the multimodal fusion strategy. The lower malignant recall suggests that additional improvements such as self-supervised pretraining, focal loss, and advanced augmentation techniques may further improve sensitivity. The combination of CNN and transformer architectures enables the framework to capture both local lesion characteristics and global contextual representations. Cross-modal attention fusion demonstrated superior capability in learning complementary multimodal relationships compared to traditional feature concatenation strategies.

CONCLUSION

This study presented a cross-modal attention-based multimodal deep learning framework for breast cancer diagnosis using mammography and ultrasound imaging. The proposed framework integrates ResNet101 and Swin Transformer encoders with a cross-attention fusion mechanism to learn complementary multimodal representations. Experimental results demonstrated that the proposed framework achieved strong classification performance with an AUC of 0.903 and accuracy of 88%. The findings indicate that multimodal fusion significantly enhances breast cancer diagnosis performance compared to single-modality approaches. Future work will focus on incorporating self-supervised learning strategies such as Masked Autoencoders (MAE) and DINO, clinical metadata integration, lesion segmentation, and explainable AI techniques including Grad-CAM and attention visualization. The proposed framework shows promising potential for intelligent computer-aided diagnosis systems and future clinical deployment.

REFERENCES

He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Liu, Z., Lin, Y., Cao, Y., *et al*. Swin Transformer: Hierarchical Vision Transformer using shifted windows.

Dhungel, N., Carneiro, G., & Bradley, A. Automated mass detection in mammograms using cascaded deep learning.

Al-antari, M. A., Al-masni, M. A., Kim, T. S. Deep learning computer-aided diagnosis for breast lesion classification in ultrasound images.

Dosovitskiy, A., *et al*. An image is worth 16x16 words: Transformers for image recognition at scale.

Carion, N., *et al*. End-to-end object detection with transformers.

Litjens, G., *et al*. A survey on deep learning in medical image analysis.

Esteva, A., *et al*. A guide to deep learning in healthcare.

****