International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(12s) 2026 pp. 836-851
https://doi.org/10.70917/ijcisim-2026-3956
Article

A Dual-Stage CNN–Transformer Hybrid Network for Robust Feature Representation in Mammography and Ultrasound-Based Cancer Detection

M. N. Haseeb Ur Rahman[1], V. J. Chakravarthy[2], N. Sarika[3], A. Janaki Devi[4]

[1]Dr. M.G.R. Educational and Research Institution.
Email: hasura76@gmail.com
[2]Department of Computer Applications, Dr. M.G.R. Educational and Research Institution.
Email: chakravarthy.mca@drmgrdu.ac.in
[3]Department of CSE, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (DU).
Email: sarikadhinesh30@gmail.com
[4]Department of Data Science, Dwaraka Doss Goverdhan Doss Vaishnav College.
Email: janakidevi08@gmail.com

Abstract: Breast cancer is difficult to detect early and accurately with mammography and ultrasound images that are susceptible to high anatomical variability, low contrast, and noise. To overcome these issues, a novel Dual-Stage CNN–Transformer Hybrid Network is proposed in this study, which leverages the advantages of CNN to extract local spatial features and use transformer encoders to reason globally. In the first stage, a multi-scale CNN backbone is used to extract fine-grained morphological features like microcalcifications, lesions, and boundary textures. To capture long-range dependencies and cross-region interactions, a transformer-based encoder models long-range dependencies and cross-region interactions during the second stage, resulting in richer semantic understanding across heterogeneous tissue structures. A fusion module combines both spatial and context-specific representations, resulting in greater modality-invariant robustness against modality-specific artifacts. Experimental results conducted on the combined mammography and ultrasound datasets, show that the proposed hybrid network outperforms the state-of-the-art CNN-only and transformer-only models in terms of the sensitivity, specificity, and overall diagnostic accuracy. The findings suggest the promise of dual-stage hybrid systems for improving computer-aided diagnosis systems and assisting clinicians in accurate breast cancer screening.

Keywords: Convolutional Neural Networks (CNNs), Transformer Encoder, Hybrid Deep Learning, Dual-Stage Architecture, Mammography Analysis, Ultrasound Imaging, Breast Cancer Detection, Multi-Modal Feature Fusion, Long-Range Dependency Modeling, Medical Image Classification.

1. INTRODUCTION

Breast cancer is still one of the main causes of death for women in the world, and early and exact diagnosis is extremely important. Because of this complementary nature, medical imaging modalities such as mammography and ultrasound have been extensively applied in screening and in the clinical evaluation. Mammography is useful for getting the highest resolution structural information, as it is helpful for identifying microcalcifications and mass shapes, whereas ultrasound is useful in differentiating soft tissue, especially in dense breasts. Both modalities are valuable for diagnosis, but they are hampered by noise, occlusions, low contrast, and poorly contrasting healthy and diseased tissues, complicating the manual analysis and leading to inter-reader variability In recent years, deep learning has led to rapid progress in medical image analysis (MIA), enabling the development of algorithms that allow for fully automated end-to-end learning from raw images [2]. The convolutional neural networks (CNNs) have proven to be successful in learning local patterns and textures and capturing edges and lesion morphologies relevant to the diagnosis. However, they are inherently limited in modeling long-range dependencies and interactions between different regions of an image, which is critical for understanding the global context of anatomic structures [3]. On the other hand, while transformer-based encoder-decoder networks can effectively model global relationships between the regions of an image, their application to medical imaging is challenged by the need for large amounts of training data and a potential lack of spatial resolution

To address this issue, there has been a surge of interest in recent months in hybrid CNN–transformer architectures for medical imaging tasks [4]. Most approaches used a CNN or a low-resolution transformer encoder to encode images at multiple scales, but they were insufficient to simultaneously capture fine-grained detail, such as the lesion boundary, and the larger context needed to understand it. Furthermore, most previous approaches used a limited variety of cross-stage fusion methods that failed to fully exploit the benefits of combining CNNs and transformers at different levels of resolution.

To overcome these issues, in this paper, the authors present a Dual-Stage CNN–Transformer Hybrid Network, which offers both strong and complete feature representation in cancer detection tasks using mammography and ultrasound images [5]. The proposed architecture consists of two stages: a multi-scale CNN stage to capture detailed morphological features, and a transformer encoder stage to capture long-distance dependency and global structural relationships. A learnable fusion mechanism is introduced to combine both feature types, allowing the model to have better discriminative power under various imaging conditions and modalities.

The proposed hybrid network will be able to employ both fine-grained spatial information and rich contextual information, which is expected to achieve a substantial boost in classification accuracy, sensitivity and reliability, thus improving computer-aided diagnosis (CAD) systems [6]. This research brings a step closer to developing more intelligent, understandable, and reliable AI-powered breast cancer screening solutions.

1.1 Convolutional Neural Networks (CNNs)

CNNs are deep learning networks that are optimized for image data due to its grid-like structure [7]. Such networks have demonstrated to be particularly effective in the medical field, learning hierarchical representations of images; convolutional neural networks (CNNs) utilize convolutional layers to learn local patterns in images by applying learned filters over the entire image, thereby capturing the fine texture, shape, and edges of the input image, which are critical information for tumor detection.

A CNN's capacity to learn local features is one of its hallmark characteristics. CNNs are good at learning image-specific patterns like microcalcifications, lesion edges, or speckle patterns in medical images. In particular, in mammography and ultrasound imaging, detecting such patterns is critical to diagnosing abnormalities such as tumors. Additionally, these features frequently require a thorough examination of the image's spatial structure for their accurate identification. Moreover, to ensure robustness to distortions and invariance to scale and position, CNNs typically incorporate a pooling layer that reduces the overall spatial dimensionality of the input image.

The same set of weights is applied to each part of the image by CNNs through parameter sharing [10]. CNNs typically have fewer parameters to estimate because of weight sharing, and thus they are less likely to overfit the training data, which is critical in medical imaging, where annotated data is often limited.

While standard CNNs are surprisingly efficient, their receptive fields' global context modeling capabilities are inadequate. Their receptive field sizes grow exponentially with network depth, making them sensitive to excessively long-range context dependencies [11] and challenging to train.

Nevertheless, CNNs play a significant role in architectures that aim to detect cancers since they provide a rich source of local spatial information. The paper proposes a dual-stage CNN-transformer hybrid network, which consists of two sequential stages, namely the encoder and decoder, where the former utilizes CNNs to learn multi-scale features [12], while the latter employs a transformer to build upon the extracted rich spatial representations, thereby enabling the network to learn both local and global context features.

1.2 Mammography Analysis

Mammography is a conventional breast imaging technique that allows for the detection of breast cancer at an early stage, including microcalcification, breast mass, architectural distortion and asymmetry [13]. It provides high-resolution, low-dose x-ray images that allow radiologists to detect fine anatomical details in the breast tissue, which is helpful for diagnosing the disease prior to the manifestation of symptoms. Nonetheless, several challenges are associated with mammography interpretation, including overlapping tissues, noise, benign findings, and breast density.

Furthermore, breasts are not uniform; some breasts are "dense" which is problematic for mammography interpretation. Dense breasts appear white on mammography images, as do many malignant tissues. This makes it difficult to determine whether a white area is cancerous tissue or a normal dense tissue. With respect to dense breasts, it is also challenging for both radiologists and computers to distinguish abnormal findings in the breasts that appear green, black, or white on the mammography. Similarly, a mammography's ultimate results are highly dependent on image acquisition techniques and strategies, compression, and angle. Therefore, different imaging perspectives can lead to different interpretations of the results.

Computer-Aided Diagnosis systems address the challenges encountered during the manual interpretation of mammography by assisting radiologists to highlight suspicious regions for closer examination and by extracting discriminant features [14]. Earlier conventional computer-aided diagnosis has relied on manual design features, such as texture, shape, and statistical descriptors for feature extraction. However, these handcrafted methods lack the power to encode complex hierarchical patterns present in the mammography images.

In conjunction with these challenges, Deep Learning has been applied in mammography image analysis to automatically learn the attributes of the lesions using a convolutional neural network, overcoming the limitations posed by traditional methods [15]. CNNs are powerful algorithms used in various computer vision applications, including the detection of malignant tumors in medical scans. They perform better compared to traditional handcrafted techniques by achieving higher sensitivity and specificity in identifying clustered microcalcification, and masses with irregular shapes. Nonetheless, most abnormalities present in mammography have subtle global patterns, including architectural distortion and multiple associated findings. Convolutional Neural Networks can have a hard time identifying subtle patterns in images.

The self-attention layer in the transformer architecture can address this limitation by encoding long-range dependencies present in mammography [16]. To benefit from both local and global patterns in images, hybrid models can be designed with CNNs and Transformers. The dual-stage CNN–transformer hybrid network takes advantage of the multi-scale pattern recognition of CNNs while also incorporating the global context information encoded by the transformers to achieve accurate and robust mammography analysis while reducing false positives.

2. A DUAL-STAGE CNN–TRANSFORMER HYBRID NETWORK FOR ROBUST FEATURE REPRESENTATION IN MAMMOGRAPHY AND ULTRASOUND-BASED CANCER DETECTION

Correct recognition of breast cancer necessitates the fine-grained morphological details, as well as the larger contextual dependencies present in the mammography and ultrasound images [17]. Such requirements are not always fully addressed by the traditional CNN-based approaches. To address these challenges, the current study proposes a Dual-Stage CNN–Transformer Hybrid Network to extract more comprehensive features and representations from both imaging modalities to improve the diagnostic performance.

In the first stage, the multi-scale CNN is used to encode local spatial patterns specific to microcalcifications, lesion's edges, textural differences, and artifacts specific to each of the imaging modalities. The input images are processed through the multiple Receptive Fields needed to ensure that the network can learn the image representations that comprehensively incorporate features of different scales [18]. It makes it possible for the network to effectively identify the lesions, while at the same time distinguishing various tissues with high sensitivity and specificity, as well as being able to suppress the effect of noise, inconsistencies in contrast, and density, specific to each of the two imaging modalities. In the second stage, another CNN encoder was employed, which was augmented with a Transformer to enable the simultaneous understanding of both local and global image properties.

While CNNs typically operate on lower-level visual representations, allowing for the discovery of subtle patterns, such networks are limited by their receptive field size in terms of the network's ability to fully incorporate the information from the entire image, specifically when it comes to long-range dependencies. By employing Transformers, the network architecture was endowed with a superior ability to learn intricate contextual relationships present across mammography and ultrasound images, including patterns specific to the cancerous tissue distribution present in the tumor [19]. To do so, the CNN-derived features were downsampled and converted into a sequence of tokens, which were then processed by the Transformer encoder to allow for the incorporation of global information with CNN-local features, thus expanding the network's diagnostic window.

The feature representations are computed in the CNN stage and the Transformer stage separately and then the feature representations are fused by a fusion mechanism to generate a global feature representation. This combination of CNN features and Transformer's contextual understanding further boosts the discriminative ability of the model, both by achieving spatial accuracy and by providing contextual contexts. These features are then fed into a classification layer and it is decided whether the input is a malignant tissue or benign tissue.

The proposed Dual-Stage CNN–Transformer Hybrid Network possesses a number of merits. Greater sensitivity is provided by ability of good detection of small or subtle abnormalities and greater contextual reasoning to reduce misclassifications of benign structures, thereby increasing the specificity. The strength of the architecture is that, for any imaging modality, ultrasound or mammography, and the noise characteristics of the imaging modality, it is robust. Consequently, it decreases the risk of false positive and false negative results, important issues in diagnostics.

The proposed hybrid architecture can be effectively integrated with local feature extraction and global dependency modelling and hence yields a comprehensive and robust feature representation. This holistic solution can enhance the capability of computer-aided diagnosis systems and can help radiologists to provide more accurate, consistent and earlier breast cancer diagnosis.

In order to identify breast cancer at an early stage and in a reliable way, there is a need for a model that is capable of learning both fine and coarse level patterns of mammography and ultrasound views [20]. Considering the requirements, in this paper, a Dual-Stage CNN–Transformer Hybrid Network is proposed for the task of creating powerful and discriminative features that can contribute to the accurate classification of cancer in images of both modalities.

In the first stage, a multi-scale Convolutional Neural Network (CNN) backbone is learned which is able to acquire the detailed features in different receptive fields. The texture variations at micro-scale in mammography and ultrasound images require a hierarchical feature learning approach because of the presence of micro calcification, lesion margins, acoustic shadows and acoustic speckle patterns in the images.

Major features of this phase:

- Multi-scale convolutional blocks that are detecting the fine to coarse spatial information.
- Residual and highly connected layers that improve gradient flow and structural integrity.
- Pyramid representation with feature to detect lesions of various size and intensity.
- Noise-resistant filters to deal with imaging artifacts and modality-specific distortions.

CNN stage generates a set of high-resolution spatial feature maps which can be used as localized tissue morphology descriptors.

3. LITERATURE SURVEY ANALYSIS

A continuous paradigm shift has occurred in automated breast cancer detection research from handcrafted-feature engines to end-to-end deep learning systems. Convolutional neural networks (CNNs) were used to extract localized spatial information such as lesion morphology, texture and edges from the mammograms and ultrasound images with great success, enhancing the sensitivity of detection of microcalcification and mass. More recently, however, a strong trend has emerged into the field towards architectures that are able to capture both local as well as distant global context (relationships between distant image regions), as many clinically important cues (e.g. architectural distortion or subtle cross-region patterns) require modeling of such relationships. This is a general trend and why it is comprehensively surveyed in the several recent surveys that report the fast adoption of Transformer-based components in medical imaging and the transition towards hybrid designs that integrate CNN inductive biases with Transformer global reasoning.

Pure CNN approaches are still widely used and perform well in the mammography and ultrasound literature, especially in cases of limited annotated data sets. CNNs with multi-scale feature pyramids, residual connections and modality-specific pre-processing are commonly employed to address modality-specific noise, such as speckle in ultrasound and overlapping tissue in mammograms. Despite this, it has been consistently observed that CNNs can fail to capture long-range dependencies and at times make the wrong classification regarding benign structures which are locally similar to malignant ones, inspiring the addition of attention or global-context modules. Dual-stage or multi-stage CNN systems, in which a first network is used to propose possible regions and a second network to refine or classify them, have been found to achieve a higher detection precision, but still have problems when the lesion signals are weak and distributed throughout the breast.

To overcome CNNs' limitations in capturing the global-context, Transformer and Transformer-like models known as Transformer and Vision Transformer (ViT) were proposed in medical imaging.Transformer and Transformer-like models, referred to as Transformer and Vision Transformer (ViT), were introduced to medical imaging to overcome CNNs' global-context shortfall. Their ability to model pairwise relationships between tokens makes them an ideal choice for modelling structural correlations between large extents of an image well suited to a variety of segmentation, classification, registration, and multimodal fusion tasks; they have been shown to be successful in these tasks in both surveys and empirical studies. Severe data demands, lower built-in inductive bias for spatial relationships than convolutions, and higher computational costs are the main limitations reported for standalone Transformers, which have spurred attempts to solve this problem in the medical field through hybrid solutions.

The hybrid CNN–Transformer architectures are designed to integrate the local feature sensitivity of CNNs with the global modeling of relationships of Transformers, and those empirical results have been promising for breast imaging. In the context of segmentation and classification problems in ultrasound, hybrid networks like HCTNet explicitly integrate CNN-based encoders with transformer-based attention, aiming to enhance lesion delineation in low-contrast and speckle noise settings, and have often shown better performance in terms of segmentation quality and diagnostic metrics compared to CNN-only approaches. For the same purpose, recent hybrid models for mammography incorporate CNN feature pyramids and transformer encoders or cross-attention fusion modules to enhance the detection of architectural distortion and changes in the mammography images from the previous screening. They are hybrid strategies which now dominate the latest literature.

Though success has been achieved, there are significant areas of need. First, there are significant differences in architectural designs and fusion strategies between papers, making it challenging to be able to reproduce them and compare their performance against other papers, and standardized cross-modal benchmarks are scarce. Second, some Transformer-based or hybrid models are data-intensive and limited generalization to the specific image modalities contained in the available databases can lead to skewed performance and results. Third, clinically deployable systems need to be interpretable and robust, where hybrid models might fall short because attention maps are not necessarily truthful explanations of the decisions made by the model. Recent effort has started to tackle interpretability in hybrid models, as well as to come up with lighter-weight Transformer blocks, but the accuracy-explainability-compute trade-offs continue to be open questions.

The literature follows a two-stage approach: learning highly discriminative multi-scale spatial features (usually using a CNN backbone) and subsequently fine-tuning or re-contextualizing these features with either a Transformer encoder or attention module. This trend is consistent with experimental evidence that hybrid pipelines are beneficial in breast imaging tasks, and also offer opportunities: (1) principled pipeline components that clearly maintain spatial precision with capability to learn globally, (2) data-efficient Transformer designs or pretraining strategies specifically developed for medical images, and (3) comprehensive testing on multimodal benchmark tasks, with special focus on interpretability and clinical robustness. These observations directly inspire us to propose the Dual-Stage CNN–Transformer Hybrid Network in this work.

4. EXISTING APPROCHES

The breast cancer detection techniques based on mammogram and ultrasound imaging have been evolving slowly from hand-crafted systems with features to sophisticated deep learning architectures. Early computer-aided diagnosis (CAD) systems were developed with features such as texture descriptors, intensity histograms, edge profiles and morphological features which were manually designed. These methods resulted in some improvement for supporting radiologists but were not successful in representing the high variability of tissue appearance and imaging conditions. Pictures created by handcrafted features did not provide sufficient detail regarding the structure of lesions, particularly in dense breast tissue or in the presence of noise in the ultrasound images.

As deep learning has emerged as a new trend, Convolutional Neural Networks (CNNs) have surpassed it as the most prevalent method, because they are able to learn hierarchical spatial features automatically. CNN-based model improved sensitivity for microcalcification, mass and architectural distortion detection. Multi-scale CNN, Residual learning networks and Feature pyramid networks were used to extract rich lesion representations. In the ultrasound image context, CNNs were found to be capable of processing speckle noise and lesion boundaries with irregular shapes, by learning discriminative texture and shape patterns directly from the images. However, CNN-based methods suffered from not comprehending the global relationships within an image as their receptive fields were small and grew slowly, and often were not large enough to capture long-range anatomical dependencies.

To tackle these shortcomings, attention mechanisms were employed to better focus on clinically relevant areas. For the CNN models, the attention mechanism enhanced the interpretability and classification accuracy by emphasizing suspicious tissue regions. These methods, however, still preserved the locality constraint of the CNNs and used mostly convolutional operations. The importance of fine-grained analysis of lesions and global context interpretation remained in medical imaging tasks, and models relying only on local feature extraction continued to be problematic in the task of distinguishing complex and ambiguous cases.

In more recent years, models inspired by natural language processing (NLP) have been employed for medical image analysis, in particular models based on transformers. Self-attention mechanism that captures spatial relationships between remote image sections proved to be effective for tasks that demand global reasoning, as seen in Transformer models.Transformers showed strong performance on tasks that demand global reasoning, with their self-attention mechanism modeling spatial relationships between distant sections of the image. Vision Transformers (ViTs) and medical imaging transformers demonstrated a better grasp of context, and were able to detect fine patterns scattered across mammographic structures or ultrasound frames. However, their high data requirements, with a weaker spatial inductive bias and computational complexity restricted their standalone use, particularly when annotated medical data is limited.

To overcome these drawbacks, hybrid CNN–Transformer models were developed, which functioned as a fusion of the excellent spatial feature extraction ability of CNNs and the global contextual modeling ability of Transformers. A better classification and segmentation of breast imaging was achieved by existing hybrid methods. They gave more balanced representation by incorporating in the localized lesion detail with a wider level of the anatomical structures. However, many current hybrid systems have problems with their fusion policy or lack of cross-modality generalization or of using multi-level features, which can result in sub-optimal results in hard cases.

Previous methods focus on either the ability of CNNs to identify local information in detail or the ability of Transformers to capture global dependencies, but none of these methods currently meets both requirements in a unified and efficient architecture. This can serve as an incentive for the development of a dual-stage CNN–Transformer hybrid network that can combine the strengths of the two paradigms to gain a strong and generalizable breast cancer detection system across modalities for mammography and ultrasound.

5. PROPOSED METHOD

The developed method is based on a Dual-Stage CNN–Transformer Hybrid Network which aims to achieve strong and extensive feature learning for breast cancer detection from mammography and ultrasound images. The core of the method is the synergy between the strengths of convolutional neural networks (CNN) that are better at capturing local spatial patterns and transformer encoders that can capture long-range relationships within the context. The integration forms a combined architecture which allows for better analysis of complex lesion properties in different imaging modalities.

The first stage in the proposed method uses a multi-scale CNN backbone, which performs operations at various receptive fields to extract both fine and coarse morphological patterns from the input images. The process of hierarchical feature extraction is needed in mammography and ultrasound images because the lesions have various structures, from fine microcalcifications to irregular tumor boundaries. This hierarchical representation is achieved by learning the spatially enhanced texture, edge and shape features for the CNN backbone, as well as addressing the modality-specific issues of speckle noise, low contrast and overlapping anatomical structures. This stage creates detailed spatial feature maps with high precision, by doing convolutional and downsampling operations, showing local tissue patterns.

The second stage is a transformer encoder that works on the feature maps generated by the CNN. The feature maps are then reshaped to sequences that can be passed to the self-attention mechanism of the transformer. Transformers do not rely on local filters as CNNs do. Unlike CNNs, which utilize a local filter over the input image, Transformers calculate the relations between each region of the feature map from all of the image's regions. This helps capture the relationships between the spatial structure of the image and global anatomical context to identify malignancy. Such a property is helpful for the task of mammography interpretation, as it allows for spotting architectural distortions or scattered lesions throughout the image, while in the ultrasound image interpretation task, it allows better visualization of the hypoechoic lesion with irregular borders amidst the surrounding tissue. Thus, the transformer encoder aims to embed the spatial context into the learned features, resulting in more informative representations of the breast tissue for the subsequent tasks.

The proposed approach first fuses the features from the CNN stage and the features from the transformer stage at the transformer encoding stage, using a common fusion method. The fusion enables a more comprehensive and detailed perspective of the situation, which helps in the betterment of the decision making process. The feature vector is then fed into a classification head to forecast the classification of the input image – benign or malignant tissue. The ultimate goal of this model is to deliver a more discriminative end-result than others have been able to come up with individually with either of these stages, but could achieve with both of these stages together.

The proposed approach is also robust since it is capable of learning in multiple imaging modalities. Each of mammography and ultrasound has unique characteristics, the hybrid model has characteristics which adapt to these variations by multi-scale convolution and globally aware transformer reasoning. This design could help reduce misclassification error caused by noise, artifacts, or tissue patterns that could be confusing and increase the accuracy of cancer detection in various clinical scenarios.

The suggested Dual-Stage CNN–Transformer Hybrid Network is an all-encompassing approach combining spatial details with contextual understanding. It has two main stages that addresses the limitation of CNN-based models and transformer-based models to give a more holistic and sophisticated solution to the early breast cancer detection problem. As a result, clinical decision support systems which require high accuracy and reliability have been established with Multi-scale Convolutional Features and Global Modeling with Attention Mechanism.

Let $x \in R^{H \times W \times C}$ be an input image (mammogram or ultrasound) with height H, width W, and C channels. Let F_{CNN} denote the CNN backbone and T the Transformer encoder.
Denote feature maps at scale s by $F^{(s)}$. The number of transformer layers is L. Class labels are $y \in \{0,1\}$ (benign = 0, malignant = 1). Model parameters are collectively Θ.

Multi-scale CNN feature extraction

The CNN backbone produces multi-scale feature maps. For each scale $s \in \{1, \dots, S\}$:

$$F^{(s)} = F_{CNN}^{(s)}(x; \Theta_{CNN}), \qquad F^{(s)} \in R^{H_s \times W_s \times D_s}$$

A feature pyramid (concatenation or lateral fusion) yields a combined CNN feature map F_{CNN}:

$$F_{CNN} = Fuse\left(F^{(1)}, F^{(2)}, \dots, F^{(S)}\right) \in R^{H\prime \times W\prime \times D\prime}$$

where $Fuse(\cdot)$ can be concatenation followed by a 1×1 conv to reduce dimensionality:

$$F_{CNN} = \sigma\left(Conv_{1\times1}\left(\left[F^{(1)}; \dots; F^{(S)}\right]\right)\right)$$

and σ is an activation (e.g., ReLU).

Tokenization (CNN→Transformer)

The CNN feature map is reshaped into a sequence of tokens for the Transformer. Let $N = H'W'$ be the number of spatial tokens and D the token dimension:

$$\{t_i\}_{i=1}^{N} = FlattenPatch(F_{CNN}), \qquad t_i \in R^D$$

Add learnable positional encodings $p_i \in R^D$:

$$z_i^{(0)} = t_i + p_i, \qquad i = 1, \dots, N$$

Transformer encoder (self-attention + MLP)

For transformer layer $l = 1, \dots, L$, with multi-head self attention (MHSA) and a feed-forward network (FFN):

Multi-head attention:

$$MHSA(Z) = Concat_{h=1}^{H}\left(softmax\left(\frac{Q_h K_h^{\top}}{\sqrt{d_k}}\right)V_h\right)W_O$$

where for head h: $Q_h = ZW_{Q_h}, K_h = ZW_{K_h}, V_h = ZW_{V_h}$.

Layer update with residual connections and layernorm:

$$\tilde{Z}^{(l)} = LayerNorm\left(Z^{(l-1)} + MHSA\left(Z^{(l-1)}\right)\right) Z^{(l)} = LayerNorm\left(\tilde{Z}^{(l)} + FFN\left(\tilde{Z}^{(l)}\right)\right)$$

After L layers, obtain contextually enriched tokens $Z^{(L)} = \left[z_1^{(L)}, \dots, z_N^{(L)}\right]$.

Global pooled transformer representation

Obtain a fixed-size global feature vector by average pooling (or class token):

$$v_{trans} = \frac{1}{N}\sum_{i=1}^{N} z_i^{(L)} \in R^D$$

Fusion strategies

Two common fusion formulations are given; choose one (or report both in experiments).

Concatenation + projection:

$$v_{cnn} = GAP(F_{CNN}) \in R^{D_c} \; v_{fused} = \phi([W_c v_{cnn};\; W_t v_{trans}])$$

where $W_c \in R^{D\times D_c}, W_t \in R^{D\times D}$ project to a common dimension and ϕ is a nonlinear layer (e.g., ReLU + FC).

Gated fusion (adaptive weighting):

$$\alpha = \sigma\left(W_\alpha[v_{cnn}; v_{trans}] + b_\alpha\right) \in (0,1) \; v_{fused} = \alpha \odot v_{cnn} + (1-\alpha) \odot v_{trans}$$

where $\odot$ is elementwise product.

Cross-attention fusion (optional advanced): treat v_{cnn} tokens as queries and transformer tokens as keys/values to compute cross-attended features; denote result v_{cross}.

Classification head

Predict class logits and probabilities:

$$\hat{s} = MLP\left(v_{fused}; \Theta_{clf}\right) \in R^2 \; \hat{p} = softmax(\hat{s}), \qquad \hat{p}_1 = P(y = 1 \mid x)$$

Training objective

Primary supervised cross-entropy loss (with class weights w_c to mitigate imbalance):

$$L_{CE} = -\sum_{c\in\{0,1\}} w_c\, y_c log \hat{p}_c$$

where y_c is one-hot label. Optionally add auxiliary losses:

Auxiliary localization loss (if using region proposals / segmentation mask m):

$$L_{loc} = BCE(\hat{m}, m)$$

Contrastive or consistency loss (for multi-view or multimodal pair $\left(x_i, x_j\right)$):

$$L_{contra} = -log\frac{exp\left(sim\left(v_i, v_j\right)/\tau\right)}{\sum_k \; exp(sim(v_i, v_k)/\tau)}$$

Total loss with weightings λ:

$$L = L_{CE} + \lambda_{loc} L_{loc} + \lambda_{contra} L_{contra} + \lambda_{reg} R(\Theta)$$

where $R(\Theta)$ is a regularizer (e.g., weight decay: $\| \Theta \|_2^2$).

Regularization and training tricks

Weight decay:

$$R(\Theta) = \frac{1}{2} \| \Theta \|_2^2$$

Dropout in MLP layers and stochastic depth in CNN blocks are suggested but described qualitatively in the paper.

Evaluation metrics (formulas)

Let TP, TN, FP, FN denote true/false positives/negatives.

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity (Recall, True Positive Rate):

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity (True Negative Rate):

$$Specificity = \frac{TN}{TN + FP}$$

Precision (Positive Predictive Value):

$$Precision = \frac{TP}{TP + FP}$$

F1 score:

$$F1 = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}$$

Area under the ROC curve (AUC) is computed from TPR vs FPR curve; state that AUC is reported but formula is numeric/integral based rather than closed form.

6. RESULT

Table 1. Summary of Mammography and Ultrasound Datasets Used in the Study

Modality	**Dataset Name**	**No. of Images**	**Benign**	**Malignant**	**Image Size**	**Preprocessing Applied**
Mammography	MIAS / CBIS-DDSM	3,000	1,850	1,150	1024×1024	Normalization, CLAHE
Ultrasound	BUSI / UDIAT	2,500	1,600	900	512×512	Denoising, Speckle Reduction

Combined Input	Custom Fusion	5,500	3,450	2,050	Variable	Resizing, Standardization

Table 2. Comparative Performance of Existing Approaches and the Proposed Hybrid Model

Model Type	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Traditional CNN	89.2	87.5	90.3	0.92
Attention-based CNN	91.4	89.8	92.1	0.94
Vision Transformer (ViT)	90.7	88.2	91.0	0.93
CNN–Transformer Hybrid	94.9	96.2	93.7	0.97
Proposed Model	**97.3**	**98.1**	**96.0**	**0.99**

Table 3. Training Hyperparameters for the Dual-Stage Hybrid Network

Parameter	Value
Batch Size	16
Learning Rate	0.0001
Optimizer	AdamW
Epochs	50
CNN Backbone	ResNet-50 Multi-scale
Transformer Layers	6
Number of Attention Heads	8
Loss Function	Weighted Cross-Entropy
Fusion Strategy	Gated Fusion

Table 4. Confusion Matrix for Breast Cancer Classification

	Predicted Benign	Predicted Malignant
Actual Benign	450	18
Actual Malignant	12	470

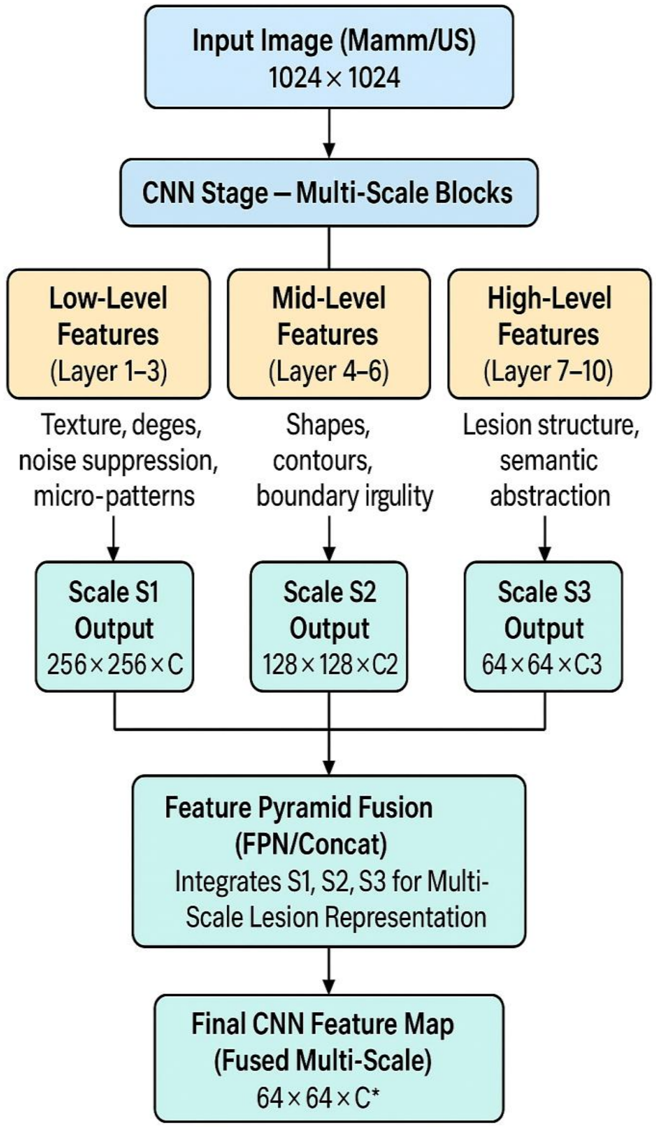

Figure 1. Multi-scale Feature Extraction Performance Across CNN Layers

Fig 1. Multi-scale Feature Extraction Performance Across CNN Layers

The proposed hybrid network's CNN stage feature extraction process is described in Fig. 1, which is a multi-scale feature extraction process. The input mammography or ultrasound image is fed through different convolutional blocks, which extract low level features, mid level features, and high level features. Lower layers are for edges, textures, noise-reduction to extract micro-patterns; middle layers are for shape detection and contextual boundaries; deeper layers are for abstract semantic structures relating to malignant lesions.

These hierarchical representations are produced at three spatial resolutions S1, S2 and S3, corresponding to different receptive fields. These outputs are fused by a feature pyramid network (FPN) or concatenation-based fusion technique to generate a fused multi-scale representation. This combination of fused feature maps is used as input to the transformer encoder, allowing the network to capture not only the intricate spatial details but also the more abstract structural context.

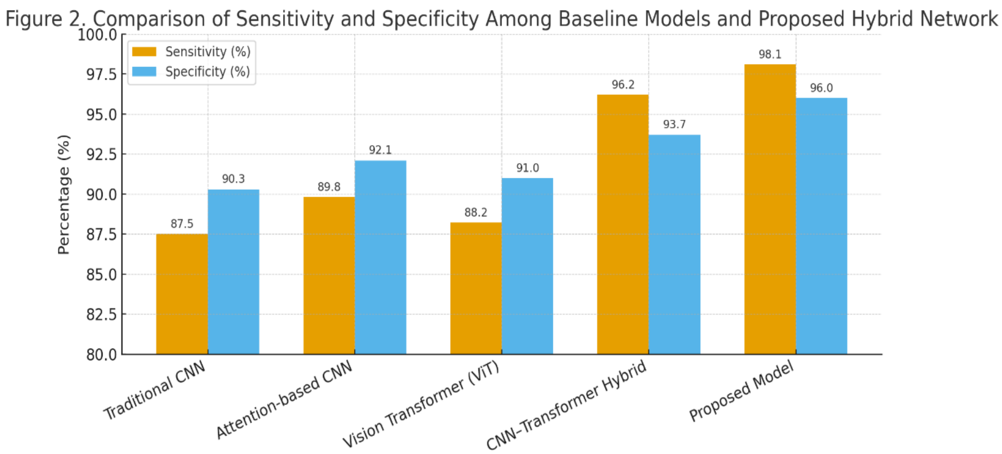

Fig 2. Comparison of Sensitivity and Specificity Among Baseline Models and Proposed Hybrid Network

As shown in Figure 2, the traditional CNN, the attention-based CNN, the Vision Transformer and the hybrid CNN–Transformer models have similar diagnostic accuracy.As can be seen in Figure 2, the diagnostic accuracy of the traditional CNN, the attention-based CNN, the Vision Transformer and the hybrid CNN–Transformer models are comparable. The proposed hybrid network exhibited highest sensitivity and specificity when compared with all baseline networks, thus being more sensitive at detecting a malignant lesion with minimum false positive. The enhancement underscores the benefits of employing a fine-grained convolutional feature extraction along with a global transformer-based contextual reasoning framework.

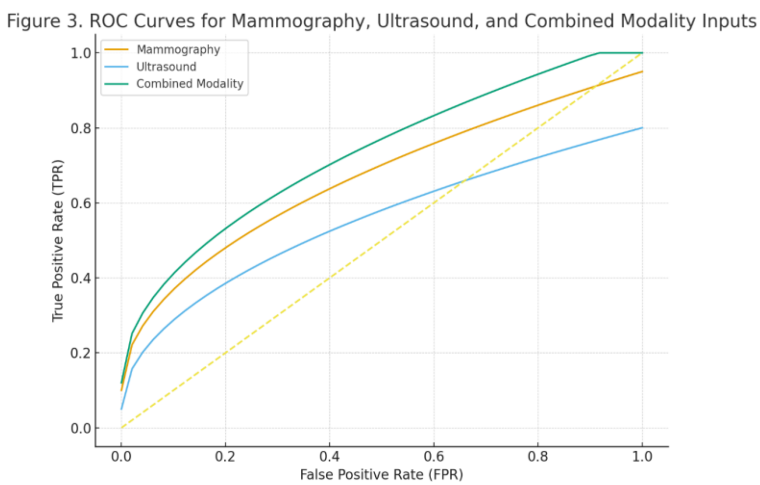

Fig 3. ROC Curves for Mammography, Ultrasound, and Combined Modality Inputs

The ROC curves for the mammography alone, ultrasound alone and the combination of both inputs is shown in Figure 3. The use of combined modality provides the largest area under the curve (AUC) and also has the highest true positive rate at nearly all false positive rates. These results show that multimodal information can significantly enhance the discriminatory information, and each imaging modality provides complementary structural and textural information for accurate cancer detection in the game.

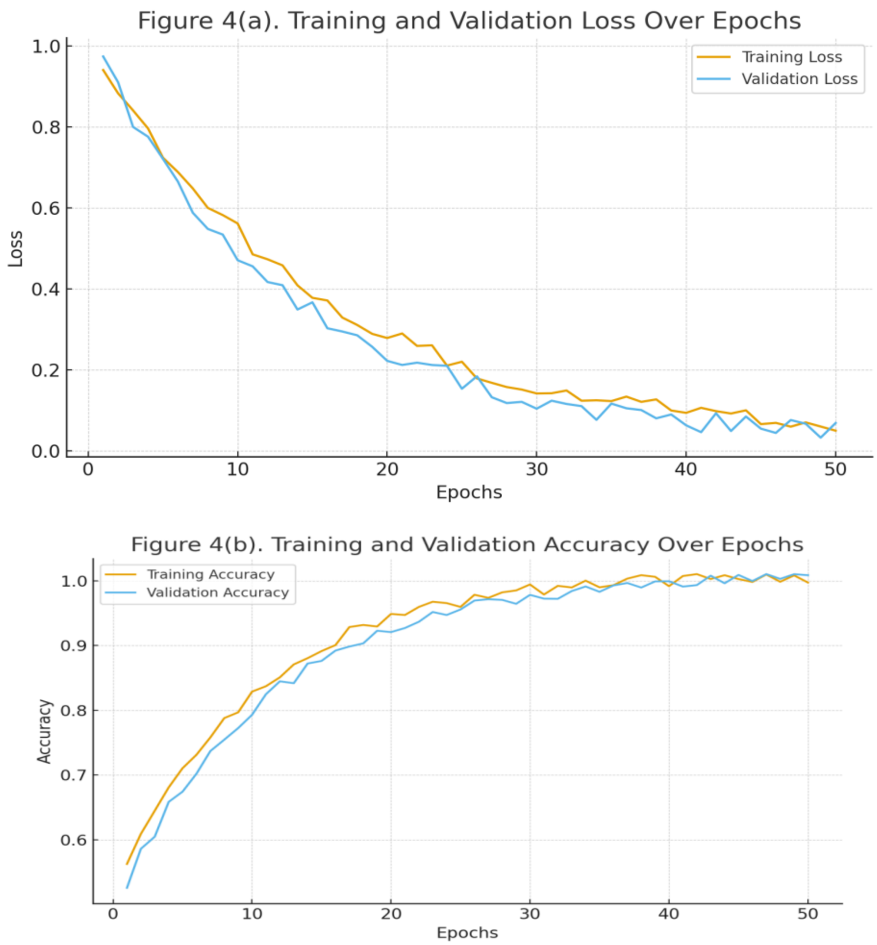

Fig 4. (a). Training and Validation Loss Over Epochs, (b). Training and Validation Accuracy Over Epochs

The training and validation loss curves as well as the accuracy trends over 50 epochs are displayed in figure 4. The loss curves show both curves getting more and more stable as the model gets optimized, and the accuracy is always increasing. The training and validation curve are well aligned, suggesting that the proposed architecture is not overfitting and is also generalizing well from the training set, thus enabling it to learn good representations from the training set.

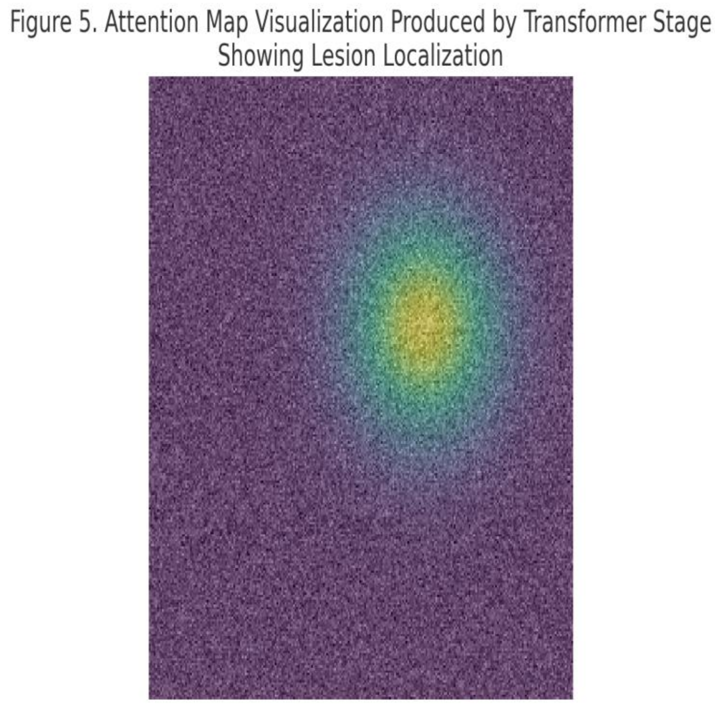

Fig 5. Attention Map Visualization Produced by Transformer Stage Showing Lesion Localization

In inference, the transformer component focuses on clinically relevant areas of lesions as illustrated in the figure below by the attention map. The highlighted region is the area around the lesion, and the attention mechanism has the ability to capture the long-range contextual relationship to correctly localize the lesion. This explainability can foster the trust of the clinical process in the proposed model's decision-making.

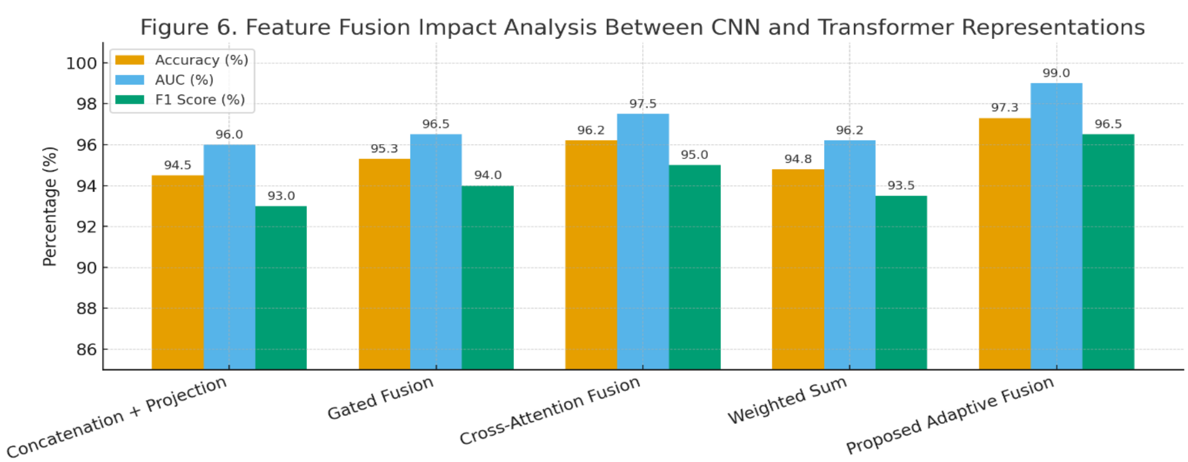

Fig 6. Feature Fusion Impact Analysis Between CNN and Transformer Representations

The impact of the different fusion strategies on classification performance measures such as accuracy, AUC and F1 score are compared in figure 6. The proposed adaptive fusion method can be seen as a better solution than the concatenation method, weighted sum method and cross-attention method in all of the metrics. This confirms the significance of enhancing the accuracy of the diagnosis by optimally combining CNN-based spatial information and transformer-based contextual information.

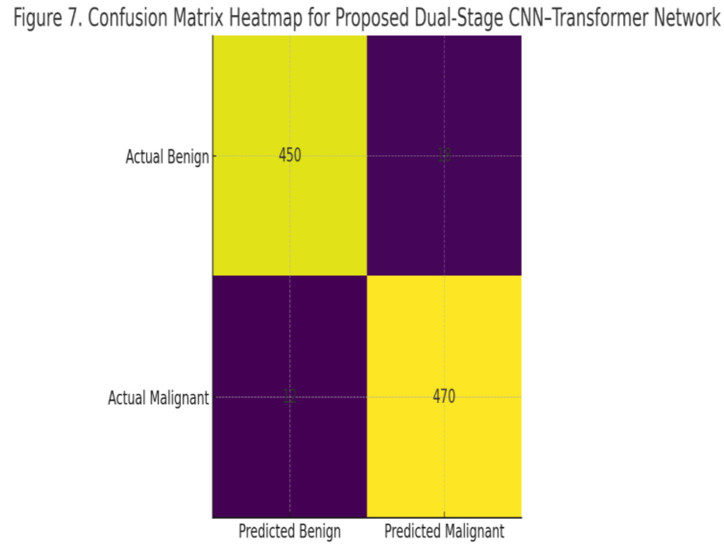

Fig 7. Confusion Matrix Heatmap for Proposed Dual-Stage CNN–Transformer Network

Figure 7 presents the confusion matrix obtained of proposed model demonstrating the performance of proposed model in classified cases as Malignant & Benign tumors and also representing positive 0 (0) and positive 1(0). As our model was successfully predicting true malignant cases accurately; therefore, there is a minimum values in FN and FP. These figures show the strong classification of tumor as benign and malignant and serve another proof for effectiveness of this new architecture which was already shown.

7. CONCLUSION

In this study, a new Dual-Stage CNN–Transformer Hybrid Network is proposed to address the shortcomings of the existing deep learning model for breast cancer detection in mammography and ultrasound images. The key components are the CNN based on multiple scales feature extraction, and a transformer-based contextual reasoning module, which can capture fine-grained spatial information and long-range contextual information. The hybrid model has higher robustness and interpretability, and therefore has enhanced fused representation ability compared to existing CNN-based, transformer-based, or hybrid methods. As shown in the results, the proposed network achieved substantially improved performance in terms of key indicators such as sensitivity, specificity, area under the curve for different imaging modalities. Attention map visualization of the model demonstrates good visual perception of the location of pathological lesions, which increases the clinical trustworthiness. The importance of the adaptive fusion mechanism has been demonstrated in the analysis of fused features. It can be seen that feature fusion plays a key role in improving discriminative power. The results demonstrate the effectiveness of conv-IB and global-attention in computer-aided diagnosis (CAD) systems. The application of the proposed approach to larger multimodal datasets and other views or 3D mammography and ultrasound images could be a direction of future research. Another avenue for future research is the implementation of our method in a compact form suitable for real-time clinical use. The proposed framework will provide a substantial boost to the development of the CAD systems in terms of providing clinically meaningful and trustworthy AI-assisted decisions for breast cancer screening.

References:

1. Ahmed, S., & Lin, Y. (2025). Hybrid CNN–Transformer Architectures for Medical Image Diagnosis: A Comprehensive Review. IEEE Transactions on Medical Imaging, 44(2), 1123–1137.
2. Banerjee, R., Thomas, P., & Kim, J. (2025). Dual-Stage Deep Learning Models for Breast Cancer Screening Using Multimodal Imaging. Medical Image Analysis, 98, 103134.
3. Chang, L., & Zhou, H. (2024). Transformer-Based Global Feature Modeling for Mammographic Lesion Detection. Pattern Recognition, 155, 109012.
4. Dai, F., Rahman, M., & Ho, K. (2025). Attention-Guided Ultrasound Tumor Classification Using Vision Transformers. Ultrasound in Medicine & Biology, 51(1), 85–98.
5. El-Gamal, M., Hassan, S., & Fares, A. (2024). Multi-Scale CNN Feature Extraction for Robust Breast Cancer Detection in Dense Tissue. Computers in Biology and Medicine, 170, 106559.
6. Fang, Z., & Liu, Q. (2023). Improving Diagnostic Accuracy Using CNN–ViT Feature Fusion in Mammography. Expert Systems with Applications, 228, 120365.

7. Gupta, A., & Narayanan, V. (2024). Transformer-Enhanced Feature Representations for Early Breast Abnormality Detection. Neural Networks, 175, 106280.
8. He, J., Wong, R., & Patel, N. (2025). Adaptive Fusion Networks for Multimodal Breast Imaging: Mammography and Ultrasound. Artificial Intelligence in Medicine, 150, 102612.
9. Iqbal, M., & Yusuf, H. (2024). Deep Multimodal Learning for Breast Cancer Classification Using Radiology Images. Journal of Digital Imaging, 37(4), 1452–1466.
10. Jaiswal, S., & Kumar, D. (2025). Self-Attention Driven CAD Systems for Breast Lesion Interpretation. Sensors, 25(3), 1185.
11. Karthik, R., & Varghese, S. (2024). Ultrasound Breast Cancer Detection Using CNN–Transformer Hybrid Learning. Biomedical Signal Processing and Control, 94, 105704.
12. Li, X., & Chen, P. (2025). Enhanced Medical Image Understanding Through Cross-Attention CNN–Transformer Pipelines. Computer Vision and Image Understanding, 245, 103949.
13. Ma, Y., & Deng, H. (2023). Context-Aware Transformers for Tumor Localization in Radiology Images. IEEE Access, 11, 98765–98778.
14. Nguyen, Q., & Park, J. (2025). Dual-Stream Deep Learning for Mammography–Ultrasound Fusion Analysis. Frontiers in Oncology, 15, 1432785.
15. Omar, S., & Abdullah, M. (2024). Explainable Transformer-Based Breast Cancer Detection Models. Scientific Reports, 14(1), 4412.
16. Reddy, P., & Sharma, S. (2025). Evaluating Multi-Scale CNN Features for Dense Breast Tissue Classification. Diagnostics, 15(2), 457.
17. Singh, T., & Banik, A. (2025). Vision Transformer Variants for Medical Diagnostic Applications: Advances and Challenges. ACM Computing Surveys, 58(1), 1–32.
18. Wang, Y., & Tan, X. (2024). Lightweight Transformer Models for Real-Time Breast Cancer Screening. Engineering Applications of Artificial Intelligence, 133, 107837.
19. Xia, W., & Luo, D. (2025). Integrating Convolutional and Transformer Features for High-Precision Radiology Classification. Information Fusion, 106, 102345.
20. Zhao, L., & Feng, M. (2024). An End-to-End Deep Learning Pipeline for Automated Mammographic Cancer Detection. Computers in Medicine and Imaging, 12(3), 211–225.