BİTLİS EREN ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

ISSN: 2147-3129 / e-ISSN: 2147-3188

Article Type : Research Article
Received : August 25, 2025
Revised : November 19, 2025
Accepted : December 15, 2025
DOI : 10.17798/bitlisfen.1772185

Year : 2025
Volume : 14
Issue : 4
Pages : 2616-2638

IMPROVING BREAST CANCER DIAGNOSIS USING ATTENTION-ENHANCED HYBRID CNN–TRANSFORMER MODEL

Asli Nur POLAT [1,*], Hussein Mahmood Abdo MOHAMMED [2]

[1] *Atatürk University, Electrical and Electronics Engineering Department, Erzurum, Türkiye*

[2] *Turk Telekom, Energy and Cooling Systems Department, Ankara, Türkiye*

**** Corresponding Author:*** *asli.omeroglu@atauni.edu.tr*

ABSTRACT

Breast cancer is the most common cancer among women and the most frequently diagnosed cancer worldwide. Recent advancements in deep learning have led to significant improvements in tumor detection from breast ultrasound (BUSI) images, enhancing the diagnostic accuracy of breast cancer screening. Although deep convolutional neural networks (CNNs) and transformer-based architectures have individually yielded promising results, challenges such as low contrast, spatial variability, and irregular tumor shapes continue to hinder the robustness of current methods. Therefore, in this study, a novel hybrid CNN–Transformer framework is proposed to improve discriminative feature extraction for BUSI cancer analysis. The network employs a dual-branch architecture, integrating features extracted from both CNN and transformer models. In the first branch, the Swin Transformer is combined with a Triplet Attention to strengthen its ability to learn long-range dependencies and global contextual information. The Triple Attention module processes feature maps along three orthogonal axes, enabling a more effective representation of both spatial and channel-level relationships. The second branch incorporates the Efficient Net architecture augmented with an Efficient Channel Attention (ECA) module, which facilitates adaptive channel-level feature recalibration. This design allows the model to emphasize diagnostically salient regions within ultrasound images. High-level features from both branches are fused for final classification. Experimental results on the BUSI dataset demonstrate that the proposed architecture achieves superior performance, with 97.4% accuracy, 97.9% precision, 97.9% sensitivity, and a 97.9% F1-score. These outcomes confirm the effectiveness of the proposed hybrid CNN–Transformer design in improving automated breast cancer diagnosis using ultrasound imaging.

Keywords: Breast cancer, Ultrasound imaging, Transformer networks, CNNs, Attention mechanisms, Hybrid deep learning.

1 INTRODUCTION

Breast cancer is the most prevalent cancer diagnosed in women globally, and it is the main cause of cancer-related deaths in 112 nations and 157 countries, respectively [1]. Although the mortality rate has decreased in recent years with early diagnosis and appropriate treatment, breast cancer is still a major threat to women's health. Classical diagnostic methods are based on imaging techniques such as ultrasound, MRI, mammography, histopathology and CT. In addition to these techniques, biopsy is sometimes used for a more definitive diagnosis. During the biopsy procedure, tissue samples are collected and examined by pathologists. This method, which largely depends on the expertise of the pathologists, is invasive and time-consuming.

Among the available imaging modalities, mammography is widely regarded as the standard technique for breast cancer screening and diagnosing. However, this procedure is both painful and exposes patients to X-rays. It also does not detect all breast cancers and takes a long time to get the result [2]. Although noninvasive breast ultrasonic imaging (BUSI) is considered the primary complementary technique to mammography for cancer diagnosis, studies indicates that it yields superior diagnostic performance compared to mammography [3]. BUSI offers several benefits, including being noninvasive, portable, real-time, cost-effective, and free from radiation risks [4]. Due to these advantages and higher sensitivity, ultrasound imaging is widely used for tumor characterization and early-stage screening and detection of breast cancer compared to other imaging techniques. Studies suggests that incorporating ultrasound into breast screening programs can lower breast cancer mortality rates by 30%–40%. Despite this progress, it has been reported that 10% to 30% of cancers are missed and 30% are recalled for further assessment of possible abnormalities [5]. The complexity of biological structures and the subtlety of pathological semantics require intensive labor in manual examination. The subjective nature of this examination leads to errors in image interpretation, delays in early diagnosis, tumor progression, and prevents timely intervention. Poor probe contact, noise, low contrast, and inappropriate pressure in ultrasound imaging make accurate interpretation of ultrasound images difficult and time-consuming [6].

Considering these difficulties, it has become increasingly clear that breast cancer diagnosis requires automated, reliable, and more efficient systems that can support clinicians and provide more accurate insights into disease detection. Computer-Aided Diagnosis (CAD) systems overcome these difficulties in classifying various types of breast cancer and show promising results in automatic image analysis tasks [7]. Deep learning-based transformers have

recently been used in breast ultrasound imaging (BUSI) and have shown impressive success in simulating long-range dependencies [8], but they still have some drawbacks. For efficient training, transformers usually need large-scale datasets and tend to ignore important spatial and local feature information. Furthermore, their ability to learn characteristics across different scales is limited since they rely on a token-based attention mechanism at a single scale. Because of this, transformers have failed to capture inter-channel feature relationships, which can be difficult when working with various sizes of pathologies.

On the other hand, CNNs are famous for their capability of extracting varied image detail levels and are effectively used in numerous fields, including medical imaging. However, since CNNs are mainly focused on the local receptive field, they are not so effective in extracting broader contextual or global knowledge. For breast ultrasound imaging, this is a significant drawback because the overall contextual consideration contributes profoundly to the accurate analysis of the disease (Wu et al., 2025). Thus, combining the global representation power of transformers with the local feature extraction capabilities of CNNs appears to be a potential approach to improve classification performance in breast ultrasound since CNNs and transformers have complementary characteristics [9], [10].

In the existing literature, most hybrid CNN–ViT architectures commonly adopt conventional spatial and channel attention mechanisms to enhance feature representation and global context modeling [11]. Although such hybrid designs show promising results, their attention mechanisms lack the ability to capture synergistic interactions between spatial and channel domains. In contrast, the proposed hybrid CNN-Transformer architecture offers a novel and complementary attention integration by combining the Triplet Attention and Efficient Channel Attention (ECA) modules in a simultaneous and unified framework. To meet this, a dual-branch hybrid network is proposed to integrate CNNs and transformer-based attention mechanisms to detect breast cancer. In this method, the Swin Transformer and a Triplet Attention module are used within the first branch which allows the network to better capture inter-channel and contextual data. By focusing attention along three orthogonal axes, this module helps the network better capture inter-dimensional spatial dependencies. Simultaneously, the second branch is built on an Efficient Channel Attention (ECA) module integrated with an Efficient Net-B0 architecture. The ECA mechanism models inter-channel dependencies with minimal computational overhead, allowing the network to assess the relative importance of channel-wise features. The complementary features extracted from both branches are concatenated to form a high-level joint representation, which is subsequently fed into the

final classification layer. The integration of attention mechanisms across both spatial dimensions and channel dimensions allows the network to concentrate on the most informative regions of the input, thereby enhancing its ability to learn both fine-grained local patterns and comprehensive contextual information. As a result, the proposed hybrid CNN-transformer architecture achieves improved accuracy and generalization capability in the task of breast cancer diagnosis. The following are the main contributions of the study:

- A novel hybrid CNN–Transformer architecture enhanced with attention mechanisms is introduced, aiming to effectively integrate both global and local feature representations to achieve superior classification performance.
- The proposed framework uniquely incorporates two emerging attention strategies—triplet attention and efficient channel attention modules. To the best of our knowledge, this is the first study to investigate the combined application of these modules in the classification of breast cancer using ultrasound imaging.
- Specifically, triplet attention modules are integrated with the Swin Transformer to enhance inter-dimensional spatial feature representation, while efficient channel attention modules are combined with Efficient Net to strengthen the model's ability to capture and emphasize critical inter-channel dependencies across multiple branches.
- The effectiveness of the proposed method is validated on the BUSI dataset, where it surpasses state-of-the-art approaches and achieves a classification accuracy of 97.43%.

2 RELATED WORK

In the literature, breast lesion classification approaches using BUS images are based on hand-crafted and CNN-based methods. Table 1 provides a summary of the studies focusing on machine learning and deep learning approaches, especially in recent years. In machine learning approaches based on hand-crafted methods, low-level feature extraction methods that include manual extraction of texture, shape, and Histogram of Oriented Gradients (HOG) features have been commonly used [12]–[14]. Since the quality of these low-level features varies significantly depending on changes in ultrasound imaging conditions, device settings, and noise, these features negatively affect the classification accuracy and weaken the reliability of diagnostic models.

Recently, various deep learning architectures, generative models, and optimization strategies have been developed to improve diagnostic accuracy, eliminate class imbalance, and

increase the reliability of diagnostic models. [15] proposed a meta-ensemble learning model combining the features of architectures such as Inception, ResNet50, and DenseNet121 and showed that this model outperformed traditional deep learning models like Convolutional Neural Networks (CNNs) in classifying BUS images by achieving 90% accuracy. Similarly, [16] made a comprehensive comparison of the performance of deep neural networks and proposed MobileNet for feature extraction, Genetic Algorithm (GA) for feature selection, and ensemble model based on weighted voting scheme for classification. The proposed model achieved 96.53% accuracy and F1 score, while providing about 4% and 9% accuracy improvement for BUSI and UDIAT datasets, respectively.

Unlike classical deep learning models, hybrid and transformer-based architectures have been developed recently. Mehmood et al. (2025) proposed a hybrid CNN–Transformer framework that incorporates regional and boundary feature extraction to improve model performance. The extracted features were combined with global attention mechanisms and achieved 95.63% accuracy, 95.57% F1 score, and 94.79% sensitivity. Similarly, [17] proposed a hybrid model combining Convolutional Neural Networks(CNNs) and Graph Neural Networks (GNNs) and achieved an AUC score of 0.911 on the BUSI dataset. The proposed model enabled feature extraction from spatial-level images and geometric-level graphs and showed reduced labeling costs. [18] proposed a binary classification filter to ensure analysis of only meaningful data. A hierarchical two-layer classification architecture was developed using NASNet and BUS images and achieved 92.7% accuracy. [19] first studied the effect of meta-learning methods on the BUSI dataset and achieved 88.2-88.9% accuracy using prototypical networks and model-free meta-learning (MAML) algorithms. This accuracy demonstrated the effectiveness of meta learning approaches by providing a 6-7% improvement over the baseline models.

Recently, new encoder models have been developed as competitors to traditional CNNs and Image Transformers (ViTs) architectures. [20] developed a Mamba-based model and achieved an average AUC of 87.50 ± 12.08 in the diagnosis of breast cancer by capturing long-range dependencies in BUS images. Furthermore, generative models have been introduced as a strategy to address the challenge of inter-class data imbalance in breast cancer diagnosis. [4] introduced a baseline generative model BUSGen trained on more than 3.5 million BUS images and improved the diagnostic performance by achieving an AUC of 0.954. It achieved a 16.5% improvement in sensitivity, outperforming board-certified radiologists in early breast cancer detection. [21] used CNN with 2D Variational Mode Decomposition (2D-VMD) to reduce the

computational cost and increase the feature representation. With their proposed method, they achieved 14.47% improvement in accuracy and 10% improvement in areas under precision recall curves compared to the state-of-the-art methods for the general dataset. Explainable AI models are also among the approaches that have attracted attention recently. [22] proposed an explainable AI model combining Densenet121 with Grad-CAM (Gradient Weighted Class Activation Mapping) to classify breast cancer and achieved 89.87% accuracy on the BUSI dataset. In the literature, multiple datasets have been used to improve breast cancer diagnosis performance using ultrasound breast images. Evaluations based on a single dataset carry the risk of overfitting the model to features specific to that dataset, while testing on multiple datasets with different features and distributions decrease this risk. Recent studies on combining two different datasets, including ultrasound and other imaging modalities, for breast cancer diagnosis rely on classical deep learning algorithms [23]–[26] and hybrid transformer-CNN methods [27]. Based on the current literature review, it appears that attention mechanisms in computer-aided breast cancer diagnosis using only BUS images have not yet been adequately addressed in the hybrid transformer-CNN architecture. Therefore, this study aims to comprehensively examine the attention mechanisms in the hybrid transformer-CNN architecture.

Table 1. Comparative summary of deep learning models for breast ultrasound image classification.

Reference	Methods	Notes	Performance
[15]	Meta-learner ensemble of ImageNet-pretrained CNNs	Meta-learner enhances base CNN predictions	ACC: 90%
[20]	Mamba-based vision encoder, CNN, ViT	The imbalance in the dataset has not been addressed during the statistical significance analysis.	ACC: 87.50% ± 12.08
[19]	Few-shot ProtoNet & MAML	Improved performance with ~6–7% over baseline	ACC: 88.2%–88.9%
[16]	MobileNet + Genetic Algorithm (GA)	The ensemble model improves the classification performance.	ACC: 96.53%, F1-score: 96.53%, Recall: 96.54%, Precision: 96.60%
[17]	CNN + Graph neural network (GNN),	Outperforms conventional methods.	AUC: 0.911 ACC:87.6%
[5]	CNN Vision-Transformer	Outperforming existing ViT and CNN methods	ACC: 95.63% F1-score: 95.57%, Sensitivity: 96.42%, Precision: 94.79%
[21]	Convolutional neural network (CNN)	Outperforms state-of-the-art techniques in accuracy by a mean (SD) of 14.47% (8.42%).	ACC: 93%
[22]	Explainable AI-based framework	Uses multiple breast ultrasound image dataset.	ACC: 89.87%

Table 1 (continued). Comparative summary of deep learning models for breast ultrasound image classification

Reference	Methods	Notes	Performance
[18]	NASNet	Eliminating irrelevant images at the filtering stage allows the system to focus on useful data and improves overall performance.	ACC: 92.7%
[23]	Modified InceptionV3, GoogLeNet, ShuffleNet, AlexNet, VGG-16, and SqueezeNet	Uses multiple breast ultrasound image dataset	ACC: 99.10% Recall: 98.90% Precision: 99.00% F1-score: 98.80%
[24]	EfficientVGG-Net V1	Uses multiple breast ultrasound image dataset	ACC: 99.27 % Precision: 99.87 % Recall: 99.87 % F1-score: 99.24 %
[27]	SwinEff-AttentionNet	Uses multiple breast ultrasound image dataset	ACC: 98.50% Precision: 98.20% Recall: 98.80% F1-score: 97.60%
[25]	ResNet-18	Uses multiple breast ultrasound image dataset.	ACC: 99.7% Precision: 99% Recall: 100% F1-score: 99.5%
[26]	VGG16+ResNet50+ Support Vector Machine (SVM)	Uses multiple breast ultrasound image dataset	ACC: 99.36%

3 MATERIALS AND METHODS

The proposed model adopts a dual-branch architecture enriched with attention mechanisms, as shown in Figure 1. This design brings together the complementary information of convolutional neural networks (CNNs) and transformer-based architectures to improve breast ultrasound image classification. In the first branch, Efficient Net is used as the backbone for convolutional feature extraction. To make it more effective, it is paired with the Efficient Channel Attention (ECA) module, which adaptively highlights the most relevant channel relationships. This setup helps the CNN branch capture fine details in the images, such as tumor edges, textures, and small variations in breast tissue. The second branch is built on the Swin Transformer, which is particularly effective in modeling long-range dependencies and capturing global contextual information. To refine these representations, the Triplet Attention module is incorporated. This module strengthens feature learning by modeling interactions across three dimensions—height, width, and channel—thereby ensuring that the most discriminative features are highlighted while suppressing irrelevant background noise. Once features are extracted, the outputs of both branches are brought together in a shared feature space. In this way, the CNN contributes precise local information while the Transformer contributes broader

context, and their fusion creates a richer and more complete representation of each ultrasound image. The fused features are then passed through a simple classification head: a fully connected layer with ReLU activation and a dropout layer to reduce overfitting. The final output gives the diagnostic category of the image (normal, benign, or malignant).

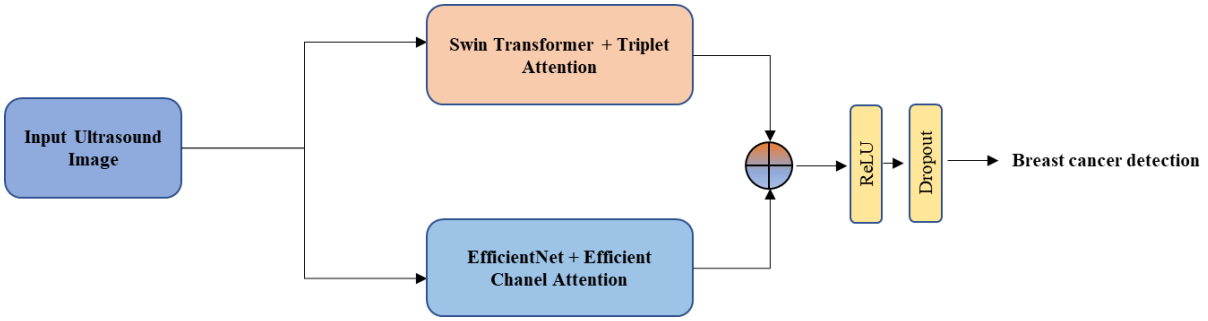


Figure 1. The proposed breast cancer diagnosis framework has two branches: Swin Transformer–Triplet Attention branch and an EfficientNet–ECA branch. The ultrasound image is simultaneously processed by dual-branch. The extracted multi-scale features are fused through concatenation and fed into ReLU and Dropout layers for final breast cancer classification.

The following subsections describe each component of the proposed architecture in detail, outlining the roles of Efficient Net with ECA, the Swin Transformer with Triplet Attention, the feature fusion process, and the classification stage.

3.1 The Swin Transformer

In the proposed model, the Swin Transformer branch is mainly responsible for capturing long-range dependencies and global contextual cues from breast ultrasound images. First the input images are processed by the backbone, the Swin Transformer Tiny (Swin-T) model (Z. Liu et al., 2021), which generates deep feature representations by modeling spatial relationships through its shifted window-based attention mechanism. To further refine these representations, the extracted features are passed through the Triplet Attention Module (Misra et al., 2021). This module enhances the feature maps by modeling interactions across the height, width, and channel dimensions, allowing the network to emphasize the most informative and discriminative regions while suppressing irrelevant responses. In this branch, the Triplet Attention complements the Swin Transformer by strengthening the balance between global context and fine-grained local details. Finally, the refined outputs from this branch are fused with the complementary features obtained from the Efficient Net + ECA branch. This fusion produces a rich and diverse feature space, which is subsequently passed to the classification layer for the final prediction.

Vision Transformers process images in fixed-sized embeddings by splitting them into a series of non-overlapping patches. This approach cannot suitably represent the complex specifics of an image despite successfully representing local and global characteristics. Spatial hierarchies along with local-global relations within the images can be represented in a better way due to the hierarchical organization in the Swin Transformer along with shifted windowing. Also, in contrast to flat models such as the Vision Transformer (ViT), Swin Transformers offer a lower computationally expensive hierarchical organization for high-resolution images to save on the computational cost. Figure 2 provides a summary of the Swin Transformer architecture.

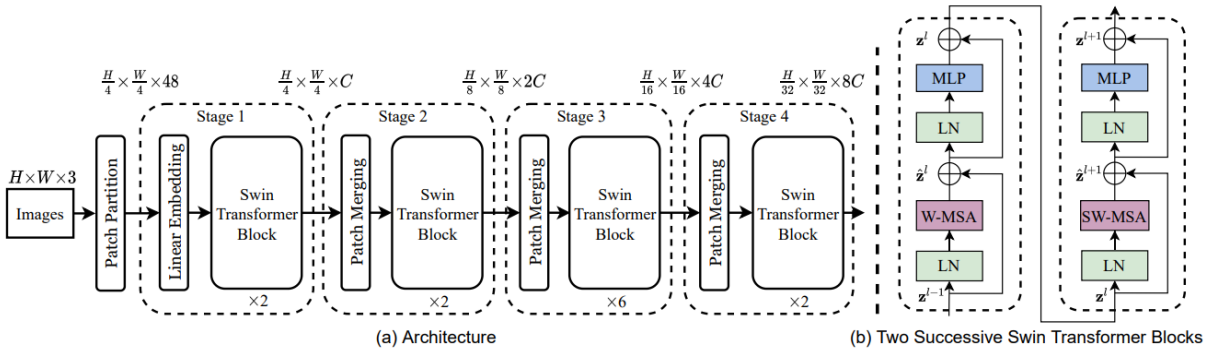


Figure 2. Swin Transformer model [28]

The Swin Transformer model, shown in Figure 2, begins by splitting the input image into non-overlapping regions (patches). In this study, a patch size of 4×4 is used, which results in a feature vector of 4×4×3=48. Every patch is considered as a token, and its representation comes directly from the raw pixel RGB values. These values are then mapped into a new feature space of dimension C using a linear embedding step. Once embedded, the tokens are processed by a sequence of Transformer blocks that use a modified self-attention mechanism. Together with the embedding stage, these form the first stage of the model.

As the network goes deeper, a patch-merging layer is introduced in order to reduce the number of tokens and to build a hierarchical feature representation. In the first merging stage, every group of four neighboring patches (2×2) is concatenated and then passed through a linear projection layer. This operation reduces the number of tokens by a factor of four, producing what is referred to as Stage 2. The same merging and transformation procedure is repeated in the following stages, resulting in Stage 3 and Stage 4. As shown in Figure 2b, the Swin Transformer block consists of window-based self-attention mechanisms (W-MSA and SW-MSA), a multi-layer perceptron (MLP), and Layer Normalization (LN) layers.

3.2 Triplet Attention Module

The architecture of the Triplet Attention module is shown in Figure 3. It is built from three parallel branches. The module takes as input a tensor $X \in \mathbb{R}^{CxHxW}$, where C denotes the number of channels, H the spatial height, and W the spatial width (Misra et al., 2021). Each branch is designed to capture relationships across different dimensions of the input. Two of the branches focus on cross-dimensional interactions, modeling how the channel dimension C relates to either the height H or width W. The third branch, similar in design to the CBAM module (Woo et al., 2018), is used to generate spatial attention. The final feature representation is obtained by taking the simple average of the outputs from all three branches.

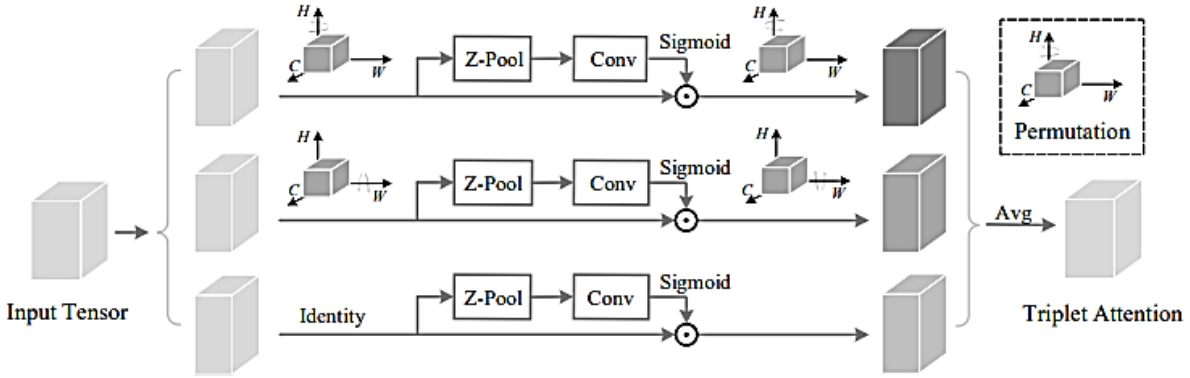


Figure 3. The Triple Attention Module architecture consists of three branches: the first branch computes the channel dimension C and the interactions in the W, the second branch computes the C and the interactions in the H, and the third branch computes the attention weights to capture the interactions between H and W. The average of the feature maps obtained from the three branches constitutes the final output [29].

The first branch of the Triplet Attention module, whose three-branch structure is shown in Figure 3, captures the interaction between the H and C dimensions. First, a new tensor ($\hat{X}_1$) is created by rotating the input tensor X 90 degrees counterclockwise along the height (H) axis. Z-pooling, which consists of averaging and max-pooling operations, is applied to this resulting tensor along the spatial dimensions of the tensor $\hat{X}_1$ (W × H × C dimensions). This results in a feature map of 2 × H × C. A convolutional layer and a batch normalization layer are applied to these resulting feature maps, respectively. In this study, the kernel size of the convolutional layer is selected as 7 × 7. This final feature map is fed to a sigmoid function to obtain the attention weights. Finally, the input map and attention maps are multiplied element-wise in the first branch to obtain the importance of each channel and its relationship to the spatial content and rotated along the spatial height axis to return the original input dimension.

The second branch captures the interaction between the W and C dimensions. The input tensor X is rotated 90 degrees counterclockwise along the W axis to obtain a new tensor ($\hat{X}_2$). A similar z-pooling operation is applied to this new tensor as in the first branch. The tensor $\hat{X}_2$ is then passed through a sigmoid function to obtain the attention weights, and the input map is multiplied element-wise. Finally, the attention-weighted tensor, which contains the important features extracted along the W axis, is rotated along the spatial expansion axis to obtain the original input size.

In the third branch, spatial attention maps are generated by performing Z-pooling across the input tensor channels, resulting in the $\hat{X}_3$ tensor. As in the other branches, Z-pooling and convolution are applied. Attention maps are obtained by passing the generated tensor through a sigmoid activation function and multiplying it element-wise with the input map. Finally, the C × H × W tensors obtained from each of these three branches are averaged to obtain the final feature map, which represents the relationship between the different dimensions of the input:

$$y = \frac{1}{3}\left(\overline{\hat{\chi}_1\sigma\left(\psi_1(\hat{\chi}_1)\right)} + \overline{\hat{\chi}_2\sigma\left(\psi_2(\hat{\chi}_2)\right)} + \chi\sigma\left(\psi_3(\hat{\chi}_3)\right)\right) \quad (1)$$

Here, ψ_1, ψ_2 and ψ_3 represent the convolution operation with kernel size k on three branches. $\overline{\hat{\chi}_1\sigma\left(\psi_1(\hat{\chi}_1)\right)}$ and $\overline{\hat{\chi}_2\sigma\left(\psi_2(\hat{\chi}_2)\right)}$ represent a 90◦ clockwise rotation.

3.3 Efficient Net

The other branch of the proposed model is based on the Efficient Net architecture [30] and is enhanced with an ECA (Efficient Channel Attention) block [31]. In this branch, the input ultrasound images are first processed through the Efficient Net model to extract deep feature representations. These extracted feature maps are then passed through the ECA block to apply a channel attention mechanism. This allows the model to emphasize informative channel-specific features, which are subsequently fused with the features obtained from the other branch before being passed to the classification layer.

Efficient Net, one of the CNN-based architectures, is a deep neural network model that is considered as a high-performance model that achieves high accuracy on the ImageNet benchmark while preserving computational efficiency. Unlike other CNN architectures, Efficient Net utilizes a compound scaling method using a set of fixed scaling coefficients to uniformly scale the network in three dimensions: width, depth, and resolution. Additionally, it utilizes the Swish activation function [30].

The Efficient Net family includes eight versions ranging from B0 to B7, with the number of parameters increasing with each version. The baseline model of the family is Efficient Net-B0. To minimize the number of trainable parameters, Efficient Net-B0 is adopted in the proposed study. The model consists of a total of 18 convolutional layers (D = 18), with each layer using either a 3×3 or 5×5 kernel. The core building block of the architecture is the MBConv module, which consists of a layer that first expands the channels and then compresses them. A Batch Normalization layer and a nonlinear activation function are used after each layer except the last fully connected layer. Depthwise convolutions are used in each layer to perform downsampling and a global average pooling operation is applied in the final layer before the fully connected layer to reduce the spatial resolution to one.

3.4 Efficient Channel Attention (ECA)

Figure 4 shows the structure of the Efficient Channel Attention (ECA) block. The input feature map is represented by $X \in \mathbb{R}^{C\times H\times W}$. C, H, and W denote the number of channels, height, and width of the feature map, respectively. First, global average pooling (GAP) is applied to the input feature map to compute the average value of all pixels in each channel, capturing channel-wise statistics. Next, a one-dimensional convolution with a kernel size of 1×k is used to model local cross-channel interactions, where the parameter k defines the extent of channel interaction. The kernel size k determined by:

$$k = \psi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{2}$$

where C and k represent channel dimension and kernel size respectively and $\gamma = 2, b = 1$. The suitability of the locality window for each layer is ensured by the ψ mapping. In the ECA module, the kernel size k determined as 5, in accordance with the dynamic kernel selection strategy. After adaptively determining the kernel size k, a 1D convolution is employed to obtain channel attention. The convolution result is then passed through a sigmoid function:

$$s_i = \sigma\left(\sum_{j=i-\lfloor k/2\rfloor}^{i+\lfloor k/2\rfloor}\right) \tag{3}$$

σ is the sigmoid function, w_j stands for convolutional weights, and z_j represents spatial locations from GAP. The output vector generated from this step are applied back to the feature map through element-wise multiplication, producing a refined feature map that emphasizes informative channels [31].

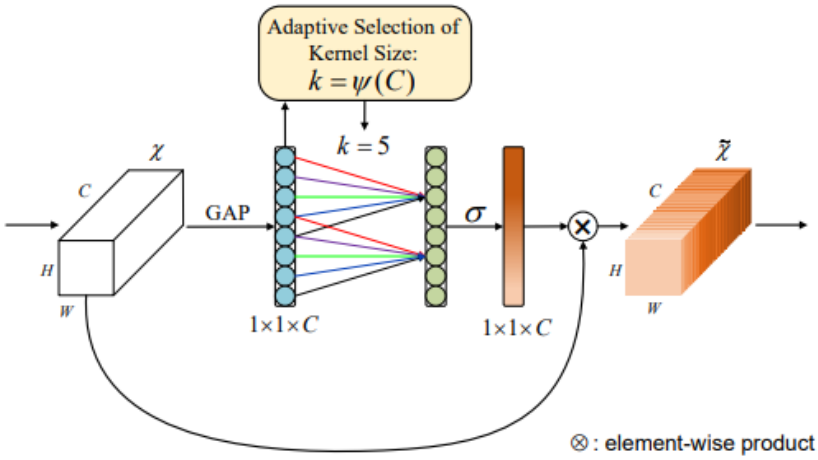


Figure 4. Efficient channel attention (ECA) module. Channel weights are obtained by applying a 1D convolution with a kernel size of k to the features obtained by global average pooling (GAP). The k-dimensional channel weights are determined adaptively, and the attention weights are fed into the sigmoid activation and multiplied element-wise with the original feature map to obtain the output [31].

3.5 Fusion Strategy

Features obtained from each branch of the proposed multi-branch deep learning architecture were combined using the concatenation approach. Highly representative feature vectors of 128 dimensions were obtained from the input image from the Swin Transformer and EfficientNet branches. Complementary channel and spatial features captured by these two branches were combined along the channel axis, resulting in a fused feature representation with a total dimension of 256. The proposed fusion approach contributes to overall model performance by combining information captured by both architectures at different scales and structurally into a single, high-dimensional representation. The concatenation fusion approach was chosen as a powerful fusion mechanism because it is easy to implement and allows for straightforward fusion in hybrid architectures without information loss.

3.6 Classification

The high-dimensional combined feature vector obtained from the fusion phase is classified using a linear classifier. First, the combined features from both branches are passed through a fully connected (linear) layer to optimize their dimensions, followed by the ReLU activation and dropout layers. ReLU activation is used to add nonlinearity, while the dropout layer is used to prevent overfitting. Finally, the classification process is performed by generating logit values corresponding to the number of classes using the second linear layer. Through this layer, the ultrasound images are classified, enabling the diagnosis of breast cancer.

4 EXPERIMENTAL SETUP AND RESULTS

4.1 Dataset

To evaluate the performance of the proposed model, the Breast Ultrasound Images Dataset (BUSI), a publicly available dataset introduced by [32], was employed. The dataset includes breast ultrasound scans collected from 600 female patients aged 25-75. In total, 780 images are provided, with an average resolution of approximately 500 × 500 pixels. Each image is assigned to one of three categories: normal, benign, or malignant. The distribution of images across these categories is reported in Table 2, and representative samples are shown in Figure 5.

Table 2. The sample distribution of label categories in the breast ultrasound images dataset.

Class	Number of images
Benign	437
Malignant	210
Normal	133
Total	780

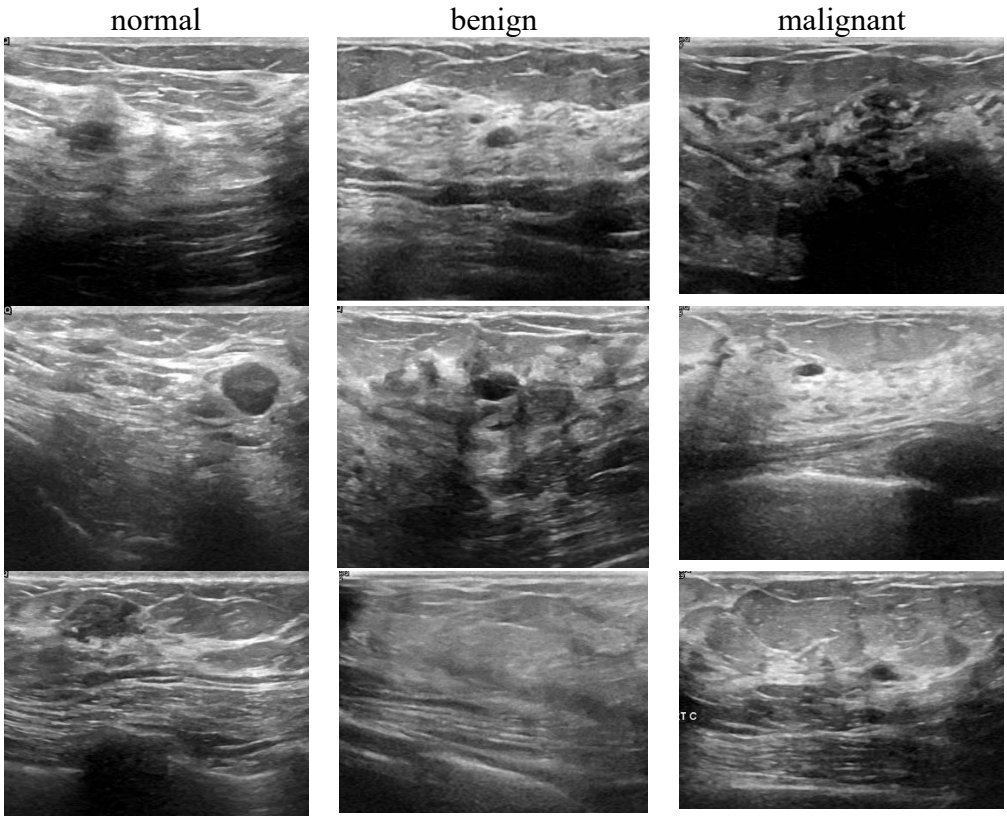


Figure 5. The example images from BUSI dataset: a) normal, b) benign and c) malignant.

4.2 Performance Metrics and Implementation Details

To provide a comprehensive comparison with other research that used the BUSI dataset, the data were randomly split into two subsets: 80% was allocated for training and 20% for testing. In this study, the classification model's effectiveness was assessed using widely adopted evaluation measures, namely accuracy (ACC), sensitivity (SEN), precision (PRE), and the F1 score. The definitions of these metrics are presented below.

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

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

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

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

In this study, PyTorch was used to train and evaluate a proposed framework for classifying breast ultrasound images. During the training stage, all input images were resized to 224 × 224 pixels and normalized based on the ImageNet mean and standard deviation values. The batch size was set to 32 for training. Model parameters were updated using the AdamW optimization algorithm with a learning rate of 1e-4 and a momentum of 0.99. Cross-Entropy Loss was utilized as the loss function. All experiments were conducted on three NVIDIA GeForce RTX 2080 Ti GPUs, each with 12 GB of memory. The proposed model contains about 31.8M trainable parameters to be updated during the implementation.

4.3 Experimental Results

In this study, various experiments have been performed to evaluate the effect of each module and branch on the performance of the proposed method. The obtained results are provided in Table 3. As can be seen in Table 3, by utilizing a single-branch architecture with only Efficient Net for breast cancer diagnosis, the lowest classification performance was achieved, achieving an accuracy of 82.1%, a precision of 82.3%, a sensitivity of 80.8%, and an F1 score of 81.5%. When the Efficient Channel Attention (ECA) block was added to the Efficient Net model, the performance increased by 4.63%, 4.50%, 5.45%, and 5.03% in terms of accuracy, precision, recall, and F1 score, respectively. This increase demonstrates the positive effect of channel attention mechanisms, particularly on CNN-based feature extraction.

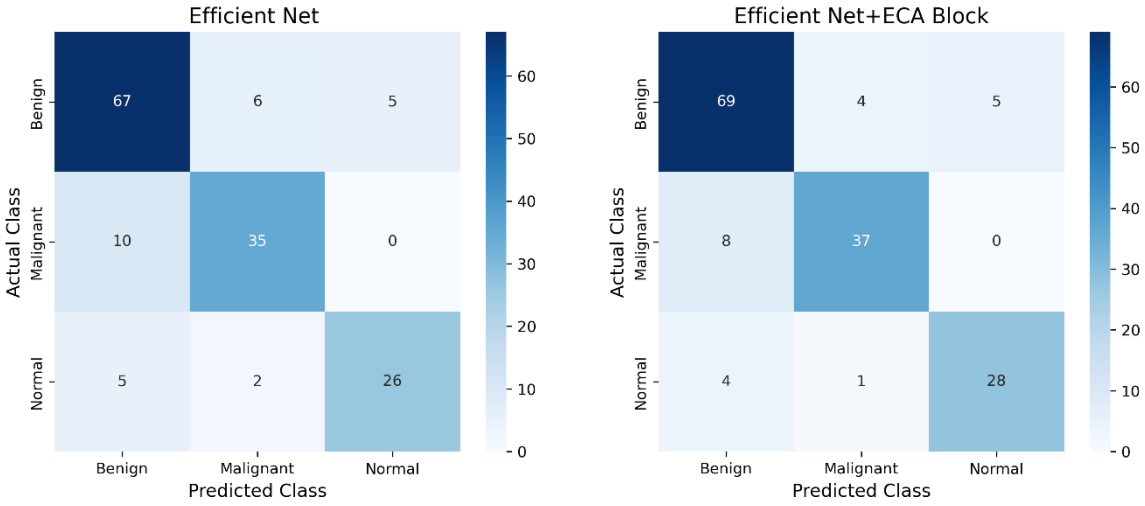
Table 3. Comparison of the classification performance of the proposed method with various modules.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Parameters
Efficient Net	82.1	82.3	80.8	81.5	4,011,391
Efficient Net+ECA Block	85.9	86.0	85.2	85.6	4,011,394
Swin Transformer+ Efficient Net	91.7	91.6	92.3	91.9	31,855,865
Swin Transformer	91.7	91.5	92.7	92.0	27,521,661
Swin Transformer+Triplet Attention	94.2	93.1	94.1	93.5	27,521,762
Proposed	97.4	97.9	97.9	97.9	31,855,969

Similarly, the single-branch architecture employing only the Swin Transformer model achieved a diagnostic accuracy of 91.7%. When combined with the Triplet Attention module, accuracy increased to 94.2%. The diagnostic performance showed improvements of 2.73%, 1.75%, 1.51%, and 1.63% in precision, recall, and F1 score metrics, respectively. The best results were obtained by the proposed architecture, which fuses these two-attention module-based branches. This structure integrates both the Swin Transformer + Triplet Attention and Efficient Net + ECA components, providing a rich feature representation. The proposed model achieved superior performance with 97.4% accuracy, 97.9% precision, 97.9% sensitivity, and a 97.9% F1 score, outperforming other methods. These results demonstrate the high effectiveness of the proposed approach in classifying breast ultrasound images.

The analysis of results from Table 6 demonstrates the impact of different modules on the performance of the VTCNet model. The baseline model achieves a good level of performance. Removing the Focus module has a minimal effect on performance, while excluding the SPPF module leads to a slight decline. The absence of the C3 module also has a modest impact. The optimal performance is achieved when all three modules (Focus, SPPF, and C3) are integrated, showing significant improvements in accuracy, precision, recall, and F1 score. These findings highlight the importance of the combined modules in enhancing the model's ability to classify and detect targets in cervical cancer images

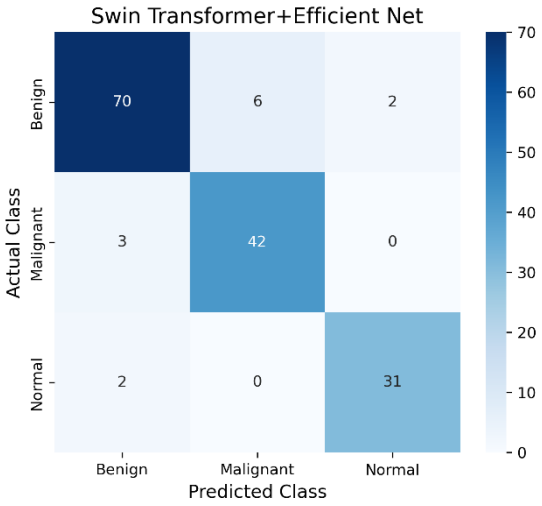
As the final demonstration of the obtained results from the ablation studies, the confusion matrices are presented for each model in Fig. 6. Confusion matrices allow comparison of the distribution of correct and incorrect classification patterns across classes obtained from ablation studies, thus revealing the contribution of each model component. Furthermore, the confusion matrices clearly demonstrate the class-by-class improvements achieved by the proposed model in difficult-to-separate classes such as benign and malignant.



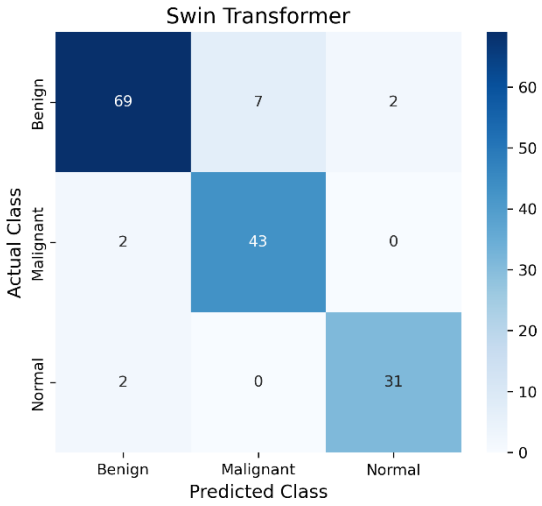
(a)

(b)

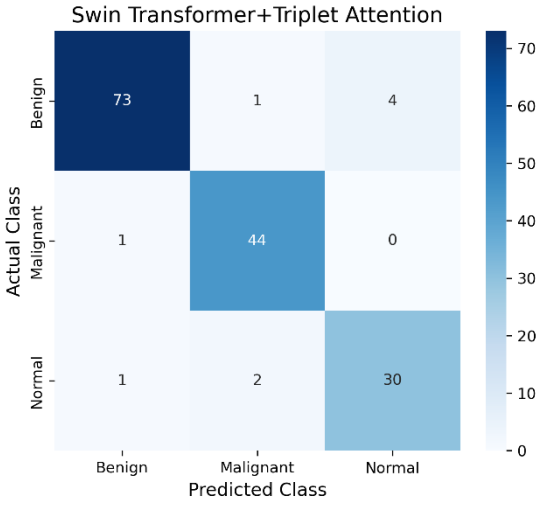
Figure 6. Confusion matrices of the ablation studies.



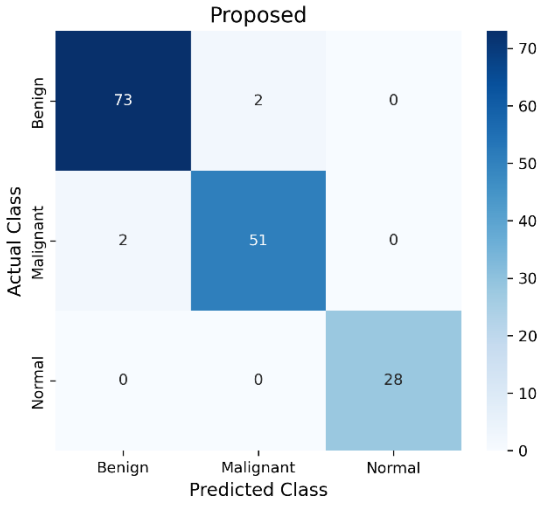
(c)



(d)



(e)



(f)

Figure 6 (continued). Confusion matrices of the ablation studies.

4.4 Comparison with the state-of-the-art methods

The comparison of the proposed method with other deep learning-based approaches reported in the literature is presented in Table 4. All of these methods relied on the same dataset (BUSI) for the classification task involving benign, malignant, and normal samples. Based on the experimental results in Table 4, the studies by [17] and [19] achieved the lowest average accuracy among all compared works. [22], employing an Explainable AI-based framework, and [15], utilizing meta-learning to improve model generalizability, attained accuracies around 90%. On the other hand, the integration of a Mamba-based encoder with CNN and Vision Transformer architectures increased classification accuracy by 3.08%. Furthermore, the diversity of pretrained networks used in these studies (NasNet, MobileNet, layered-structure CNN) indicates that different CNN models demonstrate superior performance in classifying breast ultrasound images [18], [20], [21]. A model combining CNN and Vision Transformer architectures achieved a classification accuracy of 96%, reflecting a 2.83% improvement over pure CNN architectures [5]. This clearly highlights the superior performance of hybrid CNN-Transformer models. Pretrained CNN models [16] and Genetic Algorithm-based approaches [33], which applied various preprocessing procedures, obtained nearly comparable results to the transformer-based model proposed by [5]. As evident from Table 4, the proposed method enhanced performance over different deep learning-based models presented in the literature. The proposed approach attained an accuracy of 97.4%, precision of 97.9%, recall of 97.9%, and an F1-score of 97.9%, outperforming its closest competitors. The results indicate that the proposed architecture benefits greatly from the joint use of Swin Transformer and Efficient Net, as it is able to capture both local and global information from breast ultrasound images. Efficient Net contributes to learning fine-grained spatial patterns such as textures and tumor boundaries, while the Swin Transformer complements this by modeling long-range dependencies and contextual relationships across the entire image. This complementary design explains the model's ability to maintain subtle local details while also considering broader contextual cues that are critical for accurate diagnosis. The integration of attention mechanisms further enhances performance. The Efficient Channel Attention (ECA) module improves channel-wise feature selection, ensuring that relevant features are emphasized, while the Triplet Attention module strengthens cross-dimensional interactions by jointly modeling height, width, and channel dependencies. These modules help the network concentrate on the most informative and diagnostically relevant regions, which in turn improves both robustness and interpretability.

Overall, the integration of these components results in consistently high performance achieved across all evaluation metrics. The strong results suggest that the hybrid CNN–Transformer framework is well suited to the complexity of breast ultrasound images and has the potential to reduce misclassification by balancing accuracy, robustness, and interpretability. Furthermore, the results clearly demonstrate that attention mechanisms enhance the robustness of this hybrid framework and significantly impact diagnostic accuracy by focusing on significant lesion regions in complex ultrasound images.

However, despite promising results, the proposed model also has some limitations. One of the main limitations is that it cannot fully represent the diversity and complexity of real-world clinical data in the BUSI dataset. Including multimodal data, such as demographic or genetic information, could significantly increase the generalizability and accuracy of the model. Furthermore, the proposed model exhibits some computational complexity due to its reliance on transformer branch. The multi-branch hybrid architecture allows for further optimization of the model architecture, improving model performance, but it also leads to high training time and computational costs. Therefore, future studies will limit the extension of network layers and utilize more efficient networks to reduce the computational complexity of the model. By combining CNN and Swin Transformer architectures with attention mechanisms in an innovative approach, the model demonstrated superior performance compared to state-of-the-art models in breast cancer detection. Despite limitations such as the use of single-modal data and computational costs, this study has the potential to contribute to the early diagnosis of breast cancer. The proposed approach improves both the model's detection accuracy and interpretability.

Table 4. Classification result comparison of different deep learning models on BUSI dataset.

Reference	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1- score (%)
[17]	CNN + Graph neural network (GNN),	87.6	-	83.3	83.3
[19]	Few-shot ProtoNet & MAML	88.2	-	88	87.2
[22]	Explainable AI-based framework	89.87	91.11	89.87	90
[15]	Meta-learner ensemble of ImageNet-pretrained CNNs	90	90	89.5	89.5
[20]	Mamba-based vision encoder, CNN, ViT	89.06 ± 3.72	-	-	-
[21]	Two-dimensional variational mode decomposition (2D-VMD) + Convolutional neural network (CNN)	93	86	92	89

Table 4 (continued). Classification result comparison of different deep learning models on BUSI dataset.

Reference	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1- score (%)
[18]	NASNet	93.1	95.1	88.4	91.6
[5]	CNN+Vision-Transformer	95.63	94.79	96.42	95.57
[33]	Transfer Learning	96.2	96.2	96.1	96.1
[16]	MobileNet + Genetic Algorithm (GA)	96.53	96.60	96.54	96.53
Proposed	**CNN+Swin Transformer+Attention**	**97.4**	**97.9**	**97.9**	**97.9**

5 CONCLUSION

In this study, a multi-branch deep learning algorithm based on attention mechanisms was proposed to effectively combine the advantages of CNNs and vision transformers for ultrasound breast image classification. It combines the Efficient Net boosted by the Efficient Channel Attention (ECA) within one branch, and the Swin Transformer incorporating the Triplet Attention within the other branch. By combining the complementary advantages of the CNNs in extracting fine-grained local information and transformers in capturing the long-range dependencies, the network proposed was able to produce a rich and very discriminative feature representation. The experimental results on the BUSI dataset confirmed the effectiveness of this design. Each added module contributed positively to performance: the ECA block improved the baseline CNN by emphasizing relevant inter-channel dependencies, while the Triplet Attention mechanism boosted the transformer branch by enhancing cross-dimensional feature interactions. When combined, these modules allowed the final model to achieve superior results, reaching an accuracy of 97.4%, precision of 97.9%, recall of 97.9%, and an F1-score of 97.9%, thereby outperforming other state-of-the-art approaches. Beyond the raw numbers, these findings demonstrate the importance of hybrid architectures in medical imaging tasks. CNNs alone may overlook global dependencies, while transformers without attention refinement can underutilize local spatial cues. Our results show that integrating the two, guided by attention mechanisms, yields a balanced and more powerful diagnostic tool. While the BUSI dataset provided valuable benchmarking, future research could extend validation to larger and more diverse datasets to further confirm the model's robustness. Additionally, incorporating clinical metadata and adapting the network for efficient real-time deployment represent

promising directions toward practical application in clinical environments. This study highlights the potential of hybrid CNN–Transformer architectures enriched with attention mechanisms as a promising approach for improving the breast cancer diagnosis. With further refinement and validation, such models may play an important role in supporting early detection and reducing diagnostic errors in clinical practice. Future work will focus on testing the model's generalizability to a wider range of datasets including other imaging modalities.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics.

Artificial Intelligence (AI) Contribution Statement

AI-based tools were employed only to enhance language quality and proofreading. No AI tool was used for data generation, analysis, or decision-making. All content, including text, data analysis, and figures, was solely generated by the authors.

Contributions of the Authors

Asli Nur Polat: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft, Supervision. **Hussein M. A. Mohammed:** Software, Validation, Data curation, Visualization, Writing – review & editing.

REFERENCES

[1] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024.

[2] H. A. Geuzinge, E. A. M. Heijnsdijk, I.-M. Obdeijn, H. J. de Koning, M. M. A. Tilanus-Linthorst, and F. S. Group, "Experiences, expectations and preferences regarding MRI and mammography as breast cancer screening tools in women at familial risk," *The Breast*, vol. 56, pp. 1–6, 2021.

[3] Y. Wang *et al.*, "Comparison of ultrasound and mammography for early diagnosis of breast cancer among Chinese women with suspected breast lesions: A prospective trial," *Thorac. cancer*, vol. 13, no. 22, pp. 3145–3151, 2022.

[4] H. Yu *et al.*, "A Foundational Generative Model for Breast Ultrasound Image Analysis," *arXiv Prepr. arXiv2501.06869*, 2025.

[5] A. Mehmood, Y. Hu, and S. H. Khan, "A Novel Channel Boosted Residual CNN-Transformer with Regional-Boundary Learning for Breast Cancer Detection," *arXiv Prepr. arXiv2503.15008*, 2025.

[6] J. Liu *et al.*, "Speckle noise reduction for medical ultrasound images based on cycle-consistent generative adversarial network," *Biomed. Signal Process. Control*, vol. 86, p. 105150, 2023.

[7] S. Degadwala and D. Vyas, "A review on machine learning and deep learning methods on medical image classification," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 3, pp. 546–555, 2024.

[8] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[9] Q. He, Q. Yang, and M. Xie, "HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation," *Comput. Biol. Med.*, vol. 155, p. 106629, 2023.

[10] H. Zhang *et al.*, "HAU-Net: Hybrid CNN-transformer for breast ultrasound image segmentation," *Biomed. Signal Process. Control*, vol. 87, p. 105427, 2024.

[11] Y. Haruna, S. Qin, A. H. A. Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, "Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey," *Eng. Appl. Artif. Intell.*, vol. 144, p. 110057, 2025.

[12] S. Bourouis, S. S. Band, A. Mosavi, S. Agrawal, and M. Hamdi, "Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images," *Front. Oncol.*, vol. 12, p. 834028, 2022.

[13] A. K. Mishra, P. Roy, S. Bandyopadhyay, and S. K. Das, "Breast ultrasound tumour classification: A Machine Learning—Radiomics based approach," *Expert Syst.*, vol. 38, no. 7, p. e12713, 2021.

[14] M. Wei *et al.*, "A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images," *Comput. Math. Methods Med.*, vol. 2020, no. 1, p. 5894010, 2020.

[15] M. D. Ali *et al.*, "Breast cancer classification through meta-learning ensemble technique using convolution neural networks," *Diagnostics*, vol. 13, no. 13, p. 2242, 2023.

[16] M. F. Dar and A. Ganivada, "Deep learning and genetic algorithm-based ensemble model for feature selection and classification of breast ultrasound images," *Image Vis. Comput.*, vol. 146, p. 105018, 2024.

[17] J. Ru, Z. Zhu, and J. Shi, "Spatial and geometric learning for classification of breast tumors from multi-center ultrasound images: a hybrid learning approach," *BMC Med. Imaging*, vol. 24, no. 1, p. 133, 2024.

[18] C. Kormpos, F. Zantalis, S. Katsoulis, and G. Koulouras, "Evaluating Deep Learning Architectures for Breast Tumor Classification and Ultrasound Image Detection Using Transfer Learning," *Big Data Cogn. Comput.*, vol. 9, no. 5, p. 111, 2025.

[19] G. Işık and İ. Paçal, "Few-shot classification of ultrasound breast cancer images using meta-learning algorithms," *Neural Comput. Appl.*, vol. 36, no. 20, pp. 12047–12059, 2024.

[20] A. Nasiri-Sarvi, M. S. Hosseini, and H. Rivaz, "Vision mamba for classification of breast ultrasound images," in *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, 2024, pp. 148–158.

[21] M. Saini, S. Hassanzadeh, B. Musa, M. Fatemi, and A. Alizad, "Variational mode directed deep learning framework for breast lesion classification using ultrasound imaging," *Sci. Rep.*, vol. 15, no. 1, p. 14300, 2025.

[22] M. R. Alom *et al.*, "An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images," *Sci. Rep.*, vol. 15, no. 1, pp. 1–34, 2025.

[23] S. A. Chelloug, A. S. B. Mahel, R. Alnashwan, A. Rafiq, M. S. A. Muthanna, and A. Aziz, "Enhanced breast cancer diagnosis using modified InceptionNet-V3: a deep learning approach for ultrasound image classification," *Front. Physiol.*, vol. 16, p. 1558001, 2025.

[24] H. A. Helaly, M. Badawy, E. M. El-Gendy, and A. Y. Haikal, "Early breast cancer detection, affected cell classification, and segmentation framework," *Eng. Appl. Artif. Intell.*, vol. 162, p. 112598, 2025.

[25] M. Abbadi, Y. Himeur, S. Atalla, and W. Mansoor, "Interpretable deep transfer learning for breast ultrasound cancer detection: A multi-dataset study," *arXiv Prepr. arXiv2509.05004*, 2025.

[26] N. Muzoglu, "Breast Cancer Classification in Ultrasound Imaging Using Cost-Sensitive Learning and K-Means SMOTE on the Imbalanced BUSI Dataset with Deep Feature Extraction," *Bitlis Eren Üniversitesi Fen Bilim. Derg.*, vol. 14, no. 2, pp. 755–776.

[27] I. Nissar, S. Alam, and S. Masood, "SwinEff-AttentionNet: a dual hybrid model for breast image segmentation and classification using multiple ultrasound modality," *Biomed. Signal Process. Control*, vol. 112, p. 108795, 2026.

[28] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[29] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.

[30] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.

[31] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11534–11542.

[32] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Br.*, vol. 28, p. 104863, 2020.

[33] F. Taheri and K. Rahbar, "Improving breast cancer classification in fine-grain ultrasound images through feature discrimination and a transfer learning approach," *Biomed. Signal Process. Control*, vol. 106, p. 107690, 2025.