Image Anal Stereol 2024;43:185-194 doi: 10.5566/ias.3116
Original Research Paper

A VISION TRANSFORMER NETWORK WITH WAVELET-BASED FEATURES FOR BREAST ULTRASOUND CLASSIFICATION

CHENYANG HE, YAN DIAO, XINGCONG MA, SHUO YU, XIN HE, GUOCHAO MAO, XINYU WEI, YU ZHANG AND YANG ZHAO✉

The Comprehensive Breast Care Center, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710004, China
e-mail: kenhcyxjtu@sina.com, dy971203@163.com, cameo1190@163.com, yushuo@xjtu.edu.cn, 132277809001@163.com, gchmao@163.com, maoxinwey@stu.xjtu.edu.cn, shishizheige@163.com, szhaoy@xjtu.edu.cn
(Received December 27, 2023; accepted April 16, 2024)

ABSTRACT

Breast cancer is a prominent contributor to mortality associated with cancer in the female population on a global scale. The timely identification and precise categorization of breast cancer are of utmost importance in enhancing patient prognosis. Nevertheless, the task of precisely categorizing breast cancer based on ultrasound imaging continues to present difficulties, primarily due to the presence of dense breast tissues and their inherent heterogeneity. This study presents a unique approach for breast cancer categorization utilizing the wavelet based vision transformer network. To enhance the neural network's receptive fields, we have incorporated the discrete wavelet transform (DWT) into the network input. This technique enables the capture of significant features in the frequency domain. The proposed model exhibits the capability to effectively capture intricate characteristics of breast tissue, hence enabling correct classification of breast cancer with a notable degree of precision and efficiency. We utilized two breast tumor ultrasound datasets, including 780 cases from Baheya hospital in Egypt and 267 patients from the UDIAT Diagnostic Centre of Sabadell in Spain. The findings of our study indicate that the proposed transformer network achieves exceptional performance in breast cancer classification. With an AUC rate of 0.984 and 0.968 on both datasets, our approach surpasses conventional deep learning techniques, establishing itself as the leading method in this domain. This study signifies a noteworthy advancement in the diagnosis and categorization of breast cancer, showcasing the potential of the proposed transformer networks to enhance the efficacy of medical imaging analysis.

Keywords: Breast cancer, Convolutional neural networks, Deep learning, Ultrasound, Vision-Transformer..

INTRODUCTION

Breast cancer is a prevalent malignancy among women on a global scale, and its incidence is progressively elevating, positioning it as the second most significant contributor to cancer-related deaths [1]. The timely identification of breast cancer is of utmost importance in the diagnosis and subsequent management of the illness (Wang, 2017). At present, a variety of diagnostic imaging modalities are utilized to identify anomalies in a patient's breast, including mammography, ultrasound (US), magnetic resonance imaging (MRI), and computer tomography (CT). Breast ultrasound (BUS) is a widely employed imaging modality for the characterization of breast tumors. The use of breast ultrasound (BUS) has been considered as a viable alternative to mammography in cases when individuals have thick breast tissue. Furthermore, several investigations have demonstrated that breast ultrasound (BUS) has superior diagnostic capabilities in comparison to mammography (Duffy *et al.*, 2002). The use of BUS presents several advantages, including its non-invasive nature, portability, real-time imaging capabilities, rapid results, cost-effectiveness, and absence of ionizing radiation. Nevertheless, the BUS technique does possess several limitations, which encompass the presence of artifacts like as shadows, diminished contrast, and the occurrence of speckle noise.

In the domain of medical imaging, significant progress has been made by recent advancements in deep learning methodologies, including convolutional neural networks (CNNs) and vision-Transformers (Litjens *et al.*, 2017), (Shamshad *et al.*, 2023). These techniques have demonstrated notable achievements. For example, several convolutional neural network (CNN) architectures, such as VGG, were utilized to identify breast cancers (Kalafi *et al.*, 2021), (Luo *et al.*, 2022). CNNs are preferred in this context because to their ability to acquire strong feature representations from breast ultrasound (BUS) images. While certain methods based on CNNs have reached a classification

[1] https://www.who.int/news-room/fact-sheets/detail/breast-cancer

accuracy of about 90%, it is important to acknowledge their limits. CNNs address long-range dependencies by increasing the size of the convolution kernel, which can result in decreased system speed and improved feature representation. In practical applications, the computing cost of the minor resource system is too high, hence constraining its potential to generalize. Vision-Transformers, in turn, enable the extraction of long-range dependencies by utilizing the self-attention process. Several research in the existing literature have utilized a vision-Transformer model to perform breast tumor classification on BUS (Gheflati and Rivaz, 2022) and (Hassanien *et al.*, 2022). Several research have employed the technique of combining several convolutional neural network (CNN) models to improve the accuracy of breast tumor categorization in breast ultrasound (BUS) images.

In addition, the majority of currently available breast tumor categorization methodologies yield a categorical designation, specifically benign or malignant. In this study, we propose a novel approach for breast cancer categorization utilizing the wavelet based vision transformer network. The main objective of the proposed model is to address three key challenges: enhancing feature representation by eliminating imaging artifacts, validating across various BUS datasets, and improving classification results. The model employs Discrete Wavelet Transform (DWT) at the input stage to extract essential features in the frequency domain while preserving spatial representation. Using DWT, detailed image texture information is retained through multi-frequency feature representations. The transformer network that has been presented exhibits the capability to effectively capture intricate aspects of breast tissue, leading to precise classification of breast cancer with a notable degree of accuracy and efficiency. The proposed approach being presented aims to combine the strengths of attention mechanism in order to effectively handle the uncertainties present in BUS images caused by factors such as shadows, low contrast, and speckle noise. The effectiveness of the proposed model is demonstrated by a thorough and detailed analysis of experimental outcomes using two publicly accessible BUS datasets.

RELATED WORK

This section provides an overview of the prior studies conducted in the field of breast tumor classification in ultrasound, focusing on classical methods, CNN-based approaches, and Transformer-based networks.

CLASSICAL METHODS

Numerous studies have employed conventional manual feature extraction methods in the classification of breast cancers shown in ultrasound imaging. The present study shows a categorization framework for breast anomalies utilizing seven Nakagami parametric images derived from ultrasound radio frequency (RF) data (Chowdhury *et al.*, 2022). Various morphometric, elemental, and hybrid properties were derived from each parametric images. The author utilized the aforementioned characteristics and employed a support vector machine (SVM) classifier for their classification. (Wei *et al.*, 2020) proposed the utilization of a manually designed feature extractor that incorporates several techniques such as local binary patterns (LBP), histogram of oriented gradients (HOG), gray-level co-occurrence matrices (GLCM), and shape features for the analysis of breast tumor characteristics in ultrasound imaging. The researchers employed Support Vector Machines (SVM) and Naive Bayes (NB) algorithms to perform breast tumor classification. The classification scores obtained from these algorithms were combined using a weighted fusion technique. The resulting classification attained an accuracy of 91.11%.

(Nemat *et al.*, 2018) proposed the utilization of a computer-aided diagnostic system (CAD) that incorporates a preprocessing operation to improve the quality of ultrasound depicting breast cancers. Subsequently, the use of the watershed method was employed for the purpose of segmenting the breast tumor. Ultimately, the authors integrated the logistic regression methodology into their study for the purpose of distinguishing between malignant and benign tumors. The CAD method presented by (Abdel-Nasser *et al.*, 2017) has four primary stages, including super-resolution calculation, region of interest extraction, feature extraction, and classification. The researchers employed a set of five manually designed features derived from several image analysis techniques, including GLCM (Gray-Level Co-occurrence Matrix), LBP (Local Binary Patterns), HOG (Histogram of Oriented Gradients), phase congruency-based LBP, and pattern lacunarity spectrum. These features were extracted from a BUS (Breast Ultrasound). The collected attributes were utilized to classify tumors into two categories: malignant and benign. The conventional approach exhibits many shortcomings. The approach is computationally time-consuming, less resilient, and necessitates particular feature choices and preprocessing activities.

CNN-BASED METHODS

A plethora of deep learning-based methodologies have been devised for the purpose of categorizing breast tumors into benign and malignant classifications. In their study, (Kalafi *et al.*, 2021) implemented a modification to the VGG16 network by incorporating an attention mechanism. This modification aimed to enhance the network's ability to extract pertinent characteristics and emphasize crucial pixel information pertaining to the target tumor in ultrasound images, while distinguishing it from the backdrop. The researchers employed a composite loss function comprising of binary cross-entropy and the logarithm of the hyperbolic cosine loss. The technique that was proposed demonstrated an overall accuracy rate of 93% in the classification of benign and malignant tumors in ultrasonic imaging. (Fan *et al.*, 2023) suggested an innovative model that combines localization and classification of breast masses using attention mechanisms and a sequential semi-supervised learning approach.

(Zourhri *et al.*, 2023) proposed system that utilizes Transfer Learning, approach enabling the repurposing of pre-trained models for breast tumor classification task in the US. Specifically, four pre-trained models—VGG16, VGG19, MobileNetV2, and ResNet50V2—were employed. (Luo *et al.*, 2022) proposed a deep learning approach for the segmentation and classification of breast cancers utilizing ultrasound imaging. Initially, the segmentation network produced a binary segmentation map. The subsequent stage involved the utilization of two parallel networks, each with two inputs, including the original images and the segmented image. The feature aggregation network, which incorporates channel attention, was proposed to enhance the classification performance by combining the retrieved features. Nevertheless, a significant drawback of this study is its failure to function in a comprehensive manner that incorporates increased training duration and complexity. (Byra, 2021) proposed a novel transfer learning technique called deep representation scaling (DRS) layers, which involves incorporating additional features between the pre-trained convolutional neural network (CNN) layers to improve performance. The use of this approach successfully decreases the number of trainable parameters inside the network, resulting in a notable enhancement of classification accuracy by 91.5%.

TRANSFORMER-BASED METHODS

Limited research has been undertaken in the field of ultrasonography to assess the efficacy of Transformer techniques in the detection of breast cancer. (Gheflati and Rivaz, 2022) introduced Transformer-based techniques for the classification of breast cancers using two ultrasound datasets. The utilization of a pre-trained Vision Image Transformer (ViT) model by the author served the purpose of mitigating overfitting and enhancing the acquisition of more effective feature representations on very limited ultrasound datasets. The researchers conducted a comparison between the findings obtained from the Vision Transformer (ViT) model and the current leading Convolutional Neural Network (CNN) approaches, and found that the ViT model achieved similar classification performance. In their study, (Ge *et al.*, 2023) employed a combination of Convolutional Neural Network (CNN) and Transformer models to enhance the acquisition of more effective feature representations. These representations were then utilized for the classification of breast masses into benign and malignant categories, using ultrasound images as the primary dataset. The researchers utilized a dataset consisting of 4128 images of breast ultrasound (BUS), which were further categorized into 2064 samples of benign nature and 2064 samples of malignant kind. The strategy that has been advised has yielded an Area Under the Curve (AUC) value of 97.5%. In their study, (Mo *et al.*, 2023) introduced a novel approach called the anatomy-aware HoVer-Transformer model. This model was designed specifically for the purpose of extracting anatomical information from ultrasound images, with the ultimate goal of accurately identifying breast cancers. The aforementioned methodology employed three distinct BUS datasets in order to attain cutting-edge outcomes. The radiomics approach proposed by (Hassanien *et al.*, 2022) involves using ultrasound sequences of the breast for feature extraction using the ConvNext network. Additionally, the method incorporates a pooling mechanism to calculate a malignant tumor score.

MATERIAL AND METHOD

DATASET

The present analysis utilized two publicly accessible datasets, namely UDIAT and Baheya Hospital, located in Egypt. The specifics pertaining to each dataset are explained in the subsequent sections.

– **The UDIAT BUS dataset**: The samples were collected at the UDIAT Diagnostic Centre, which is a part of the Parc Tauli Corporation located in Sabadell, Spain (Yap *et al.*, 2017). The dataset known as UDIAT comprises a collection of 163 ultrasound images specifically depicting breast cancers. In the provided samples, there are 109

cases of benign breast tumors and 54 cases of malignant breast tumors. The ultrasound images include a mean resolution of 760 pixels in width and 570 pixels in height.

– **The Baheya Hospital dataset**: The breast ultrasound samples utilized in this study were obtained from Baheya Hospital, located in Egypt (Al-Dhabyani *et al.*, 2020). The dataset has a total of 780 samples, which are classified into three categories: normal, benign, and malignant. The normal category consists of 133 samples, the benign category consists of 487 samples, and the malignant category consists of 210 samples. The typical resolution dimensions of an ultrasound sample are 500 pixels by 500 pixels. This study specifically eliminates samples from normal classes and focuses solely on samples classified as benign or malignant.

The proposed framework for breast tumor malignancy prediction is illustrated in Fig. 1. We employed wavelet based vision-Transformer architectures in order to assess the malignancy ratings of breast tumors based on ultrasound images.

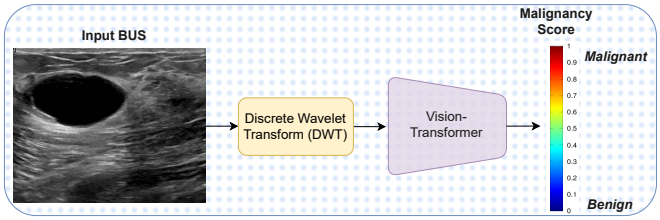

Fig. 1. *The proposed framework is being suggested for the prediction of breast tumor malignancy. The provided wavelet based input image of a BUS is fed into the vision-Transformer model, which subsequently produces a malignancy score.*

WAVELET BASED VISION TRANSFORMER

Breaking down an image into a set of wavelet coefficients of consistent size is a crucial step in the transformative process. This intricate decomposition enables the extraction of comprehensive information from the input image, including both global topological details and intricate local textural features. By encompassing these aspects, the transformation significantly enhances the neural network's ability to discern and understand the diverse characteristics present in the image. Additionally, this process contributes to the expansion of the receptive field for individual neurons within the architecture, ensuring a more holistic perception of the input data. As a result, the combination of global and local information, coupled with an enlarged receptive field, equips the neural network with a more nuanced understanding of the input, fostering improved feature representation and extraction (Daubechies, 1990). DWT employs multiple filter banks to partition the time and frequency components of the feature vector across different resolutions (Daubechies, 1990). In particular, we apply a 2D Discrete Wavelet Transform (DWT) using four convolutional Haar filters. These filters, or kernels, can be mathematically expressed as: $k_{LL} = [1 \quad 1; 1 \quad 1]^T$, $k_{LH} = [-1 \quad -1; 1 \quad 1]^T$, $k_{HL} = [-1 \quad 1; -1 \quad 1]^T$, and $k_{HH} = [1 \quad -1; -1 \quad 1]^T$, to decompose a abdominal US, I, into four sub-bands, i.e. I_{LL}, I_{LH}, I_{HL}, and I_{HH}.

$$\begin{cases} \mathbf{I}_{LL}(i,j) = \mathbf{I}(2i-1,2j-1) + \mathbf{I}(2i-1,2j) + \mathbf{I}(2i,2j-1) + \mathbf{I}(2i,2j) \\ \mathbf{I}_{LH}(i,j) = -\mathbf{I}(2i-1,2j-1) - \mathbf{I}(2i-1,2j) + \mathbf{I}(2i,2j-1) + \mathbf{I}(2i,2j) \\ \mathbf{I}_{HL}(i,j) = -\mathbf{I}(2i-1,2j-1) + \mathbf{I}(2i-1,2j) - \mathbf{I}(2i,2j-1) + \mathbf{I}(2i,2j) \\ \mathbf{I}_{HH}(i,j) = \mathbf{I}(2i-1,2j-1) - \mathbf{I}(2i-1,2j) - \mathbf{I}(2i,2j-1) + \mathbf{I}(2i,2j) \end{cases} \quad (1)$$

By decomposing the input features and feeding them to the network, the variety and richness of the input is increased. This results in a faster training process and quicker convergence. This technique is similar to dilated filtering operations, which divide the image into sub-images using DWT.

Our primary focus was on the self-attention mechanism used in the Vision Transformer to gain a deeper understanding of cross-covariance attention (Ali *et al.*, 2021). This mechanism allows the model to concentrate on important and relevant features in images while disregarding irrelevant information. The attention mechanism functions by computing a weighted sum of all the features obtained from given patch images. The Transformer model is then trained to learn the weights assigned to each extracted feature, which are used to estimate the attention coefficient.

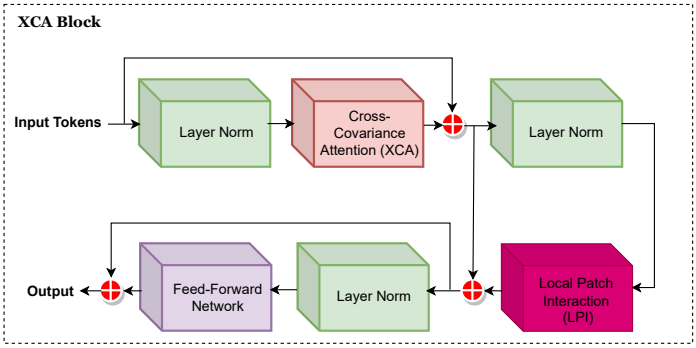

Fig. 2. *Illustration of the XCA layer.*

Assuming an input sequence representing US image patches with s patch embeddings (i.e., $z_1, z_2, z_3 \ldots\ldots z_s$), we can represent each entity using a feature embedding dimension of dim. The entire

sequence can be represented as a matrix in $\mathbf{Z} \in \mathbb{R}^{s \times dim}$, where s represents the number of tokens in the input sequence. For every patch embedding, three linear projections are used to acquire three vectors: Query (Q), Keys (K), and Values (V). To capture the global feature representation, the three learnable weight matrix representations of Queries $W^Q \in \mathbb{R}^{s \times dim_q}$, Keys $W^K \in \mathbb{R}^{s \times dim_k}$, and Values $W^V \in \mathbb{R}^{s \times dim_v}$ that computed from the input sequence can be explained as follows:

$$Q = W_Q \cdot Z, \quad K = W_K \cdot Z, \quad V = W_V \cdot Z \qquad (2)$$

Here, $W_Q \in \mathbb{R}^{dim \times dim_q}, W_K \in \mathbb{R}^{dim \times dim_k}$, and $W_V \in \mathbb{R}^{dim \times dim_v}$ belongs to the learnable parameters. Therefore, the self-attention can be computed by:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{dim_q}}\right)\mathbf{V} \qquad (3)$$

Where dim_q, dim_k, and dim_v correspond to the dimensions of Q, K, and V, respectively. Additionally, $Softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{dim_q}}\right)$ is the employed to produce the attention vectors. To address the gradient vanishing problem of the softmax function, the dot-products of queries and keys are divided by the square root of $\sqrt{dim_q}$. Furthermore, this allows for improving the training process of the model.

Fig. 2 shows the general description of the XCA layer. The cross-covariance attention computes the attention along the features or channels dimension rather than the token dimension, which can be expressed as follows:

$$XCA_{Attn}(Q,K,V) = V\mathscr{A}_{\mathrm{XCA}}(K,Q) \qquad (4)$$

Where $\mathscr{A}_{\mathrm{XCA}}(K,Q)$ is the $Softmax\left(\hat{K}^\top \hat{Q}/\tau\right)$ that generate the attention scores, and τ correspond to a learnable temperature that provides better model training. It is worth noting that the estimation of attention weights $\mathscr{A}$ relies on the cross-covariance matrix.

Local patch interaction

As XCA does not have a direct connection between tokens, it can restrict the model's strength to capture local associations between pixels in input images. Therefore, the foundation of the LPI layer lies in the combination of information between the tokens in the input sequence. The attention layers from the attention mechanism are usually employed to merge this information, as previously mentioned in self-attention-related literature. Nevertheless, the XCA attention layer extends this capability by enabling the integration of information between features or channels in the input sequence instead of just the tokens. This layer allows capturing local spatial features similar to CNNs and leads to better results. The LPI block depicted in Fig. 3, uses two depth-wise convolutions, which are separated by batch normalization and nonlinear Gaussian Error Linear Unit (GELU) activation function. These convolutional layers incorporated the kernel size of 3×3.

Feed forward network

The XCA block utilizes the point-wise FFN layer that contains the single hidden layer with four-dimensional hidden units. FFN permits interaction between all features when there are no feature relations in the LPI block.

The design utilized in this study is derived from the XCiT Transformer model, as described in the work by Ali et al. (Ali *et al.*, 2021). As seen in Fig. 2, every XCiT layer comprises three primary components: the core cross-covariance attention (XCA) operation, the local patch interaction (LPI) module, and a feed-forward network (FFN). LayerNorm is applied before each layer, and a residual connection is applied after each layer. The fundamental design elements of this architectural framework encompass the depth of the model, the dimensionality of the patch embedding denoted as d, and the utilization of a certain number of heads denoted as h in the context of Cross-Attention (XCA). The construction of proposed model involved the utilization of XCiT-L24, which was characterized by a model depth of 24, patch embeddings dimensions of 768, and 16 heads. The dimensions of the input image of the bus are 224×224, whereas the dimensions of each patch are 16×16.

PERFORMANCE MEASUREMENT

This article presents an evaluation of the proposed methods utilizing five evaluation metrics, namely accuracy, precision, recall, and F1-score. These metrics can be mathematically expressed as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (6)$$

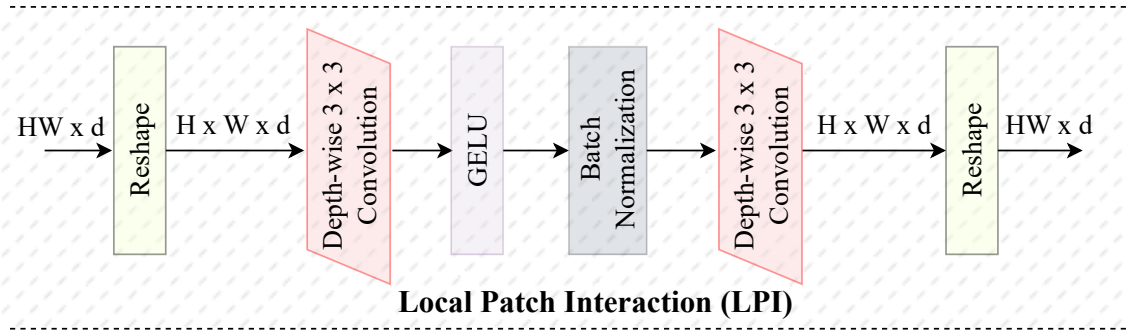

Fig. 3. *Illustration of local patch interaction block.*

$$\text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{F1-score} = \frac{TP}{TP+0.5(FP+FN)} \tag{8}$$

In the context of this study, TP represents the accuracy rate of correctly classifying malignant BUS images, while TN represents the accuracy rate of correctly classifying benign BUS images. FP denotes the rate of incorrectly identifying benign BUS images as malignant, and FN represents the rate of incorrectly classifying malignant BUS images as benign.

EXPERIMENTAL RESULTS AND DISCUSSION

IMPLEMENTATION DETAILS

In the present study, the original BUS images were subjected to rescaling, resulting in a resolution of 224×224 pixels. In order to increase the diversity of features, the data augmentation approach was employed, which involved applying a rotation of 30 degrees, a scaling probability of 0.5, and horizontal and vertical flipping with a chance of 0.5. The normalization of the breast tumor images was performed by calculating the mean and standard deviation. The model was optimized using an ADAM optimizer, with an initial learning rate of 0.0001. The classification models were trained for 50 epochs using a mini-batch size of four. It is noteworthy to mention that all of the trained models employed identical hyperparameter configurations. The cross-entropy loss function was utilized in order to enhance the optimization of the model. The BUS datasets were divided into three distinct subsets, namely training, validation, and testing, in the proportions of 70%, 10%, and 20% respectively. It is important to acknowledge that both ultrasound datasets underwent distinct processes of training, validation, and evaluation.

Computational Setup: The deep learning-based algorithms were trained and assessed using the PyTorch neural network library. The training and evaluation were conducted on a system equipped with an Intel Core-i9 CPU, 32GB of RAM, and a GeForce RTX 2080Ti GPU with 11GB of memory.

STATE-OF-THE-ART RESULTS COMPARISON

Table 1 compares the proposed breast tumor malignancy score prediction model to six different state-of-the-art classification approaches using UDIAT and Baheya ultrasound datasets. ConvNextv2 (Woo *et al.*, 2023), ResNet101 (Szegedy *et al.*, 2017), MobileNetV2 (Sandler *et al.*, 2018), ResNext101 (Xie *et al.*, 2017), EfficientNetV2 (Tan and Le, 2021), and XCiT were analyzed. In comparison to current approaches, the suggested model achieved the maximum classification accuracy of 96.98% for UDIAT and 95.10% for Baheya datasets. This enhanced performance is achieved by incorporating DWT features into the network input, resulting in a notable improvement of 1% compared to XCiT without the use of DWT. In the UDIAT dataset, ResNext101 received the third-highest scores, with accuracy, precision, recall, and F1-score metrics 8%, 9%, 8%, and 9% lower than proposed. However, MobileNetV2, and EfficientNetV2 perform similarly across all criteria. On Baheya, EfficientNetV2 obtained the third-best classification results. Finally, the proposed approach can efficiently extract features from noisy ultrasound and identify patterns as benign or cancerous. It optimizes the model using a limited set of BUS samples using cheaper computer procedures. These best methods of both datasets accurately eliminated image artifacts through self-attention mechanisms and identified the presence of neighboring tissue to precisely classify breast tumors.

The classification results of the proposed technique are presented in Table 2, alongside the latest state-of-the-art results on the UDIAT dataset. It is evident

Table 1. *Compared state-of-the-art CNN networks with the proposed model.*

Model	UDIAT					Baheya				
	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score	AUC
ConvNextv2	85.48	85.65	85.48	86.6	89.10	87.45	87.41	84.34	83.86	89.29
ResNet101	78.65	74.41	72.58	73.2	76.12	78.47	76.7	76.32	75.87	81.82
MobileNetV2	79.77	78.33	81.81	78.06	82.67	85.27	84.01	81.69	82.66	86.84
ResNext101	87.87	85.96	88.63	86.90	88.49	86.82	88.01	81.60	83.76	87.35
EfficientNetV2	77.02	77.95	80.54	77.46	80.9	88.74	86.8	88.11	87.05	89.4
XCiT (w/o DWT)	**95.87**	**93.54**	**95.48**	**94.22**	**97.9**	**93.40**	**92.14**	**93.57**	**92.82**	**95.66**
Proposed (With DWT)	**96.98**	**94.67**	**96.79**	**95.30**	**98.40**	**95.10**	**93.55**	**94.66**	**93.74**	**96.82**

from the results that the suggested mechanism exhibited superior performance compared to the current techniques mentioned in (Byra *et al.*, 2019) and (Byra and Andre, 2019), achieving an increase in accuracy of 13% and 21% respectively. The study conducted by Ning et al. (2020) utilized a multi-scale patch extraction approach, which yielded the second-highest outcomes of 90.90% accuracy and 93.90% AUC. The comparison of classification results between the proposed technique and current studies on the Baheya dataset is presented in Table 3. The suggested model has reached the maximum classification result of 95.10%, along with an AUC score of 0.968. Additionally, the model has shown an improvement of 2% in terms of AUC compared to the technique presented by (Moon *et al.*, 2020). It is important to highlight that the proposed model yields superior classification outcomes compared to other models. The suggested approach has demonstrated a significant enhancement in the performance of both BUS datasets, hence establishing a higher level of reliability and accuracy in the prediction of breast tumors.

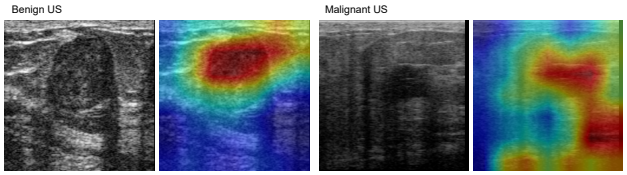

Fig. 4. *The proposed model is utilized to produce an illustration of Gradcam for breast tumor categorization. The red/yellow coloration delineates the specific area in which the network exhibits a heightened emphasis on tumor identification. Networks with blue highlights tend to capture feature representations of less significant background tissues.*

The visualization depicted in Fig. 4 illustrates the application of the proposed approach to both benign and cancerous samples. A higher intensity of red/yellow hues indicates that the network has successfully captured the most significant and pertinent characteristics linked to the tumor. Conversely, blue hues represent lesser priority areas, such as background pixels. The proposed methodology prioritized the identification of tumor by placing greater emphasis on the analysis of healthy tissues pixels inside a benign sample. The tumor exhibits neighboring tissue characterized by areas of reduced pixel intensity and alterations in pixel values. The suggested model effectively captures and delineates the region of interest. The suggested network filter effectively identified cancerous pixels and disregarded any accompanying artifacts in the malignant sample. The proposed model offers a notable benefit in terms of modifying spatial attention through the integration of local self-attention and global self-attention. This unique combination enables vision transformer to effectively collect a greater range of spatial data pertaining to the breast tumor. The approaches that were assessed shown a much superior ability to differentiate between benign and malignant tumors.

In a clinical context, the reliance on precise forecasts is both critical and highly delicate. Certain models demonstrated exceptional performance in particular scenarios, whilst others shown an inability to accurately classify data. An attempt was made to address these difficulties through the utilization of the proposed model. Based on a comprehensive analysis of both quantitative and qualitative assessments, it can be concluded that the suggested technique offers a dependable and resilient method for tumor classification. To achieve the proposed objective, the model utilized DWT to separate the input's lower and higher frequency components. This approach emphasized spatial information and encouraged the model to learn effective feature representations. The lower frequency component contained details such as shadows, speckle noise, and illumination changes, while the higher frequency pertained to essential shapes, edges, margins, and fine details. This enabled the model to avoid imaging artifacts and significantly improve classification results, as demonstrated by comprehensive experimental results. The power of vision-Transformers was employed in our study due to their distinct ability to capture both local and long-range contextual information. This utilization was

Table 2. *Comparing the proposed method with existing works on UDIAT dataset. Dashed lines reflect results not published in the reference.*

Methods	Evaluation Metrics				
	Accuracy	Precision	Recall	F1-Score	AUC
Proposed	**96.98**	**94.67**	**96.79**	**95.30**	**98.40**
(Byra *et al.*, 2019)	84	–	85.10	–	89.30
(Byra and Andre, 2019)	76	–	78	–	81
(Ning *et al.*, 2020)	90.90	–	92.70	–	93.90

Table 3. *Comparing the proposed method with existing works on Baheya dataset.*

Method	Evaluation Metrics				
	Accuracy	Precision	Recall	F1-Score	AUC
Proposed	**95.10**	**93.55**	**94.66**	**93.74**	**96.82**
(Moon *et al.*, 2020)	90.77	72.50	96.67	82.86	94.89
(Das and Rana, 2021)	88.89	88	87	87	–
(Vigil *et al.*, 2022)	85.30	–	–	–	–

crucial for effectively categorizing the characteristics of breast tumors. It is important to acknowledge that the suggested methodology is not restricted to the estimation of malignancy scores for breast cancers using BUS images. Ultrasound and other medical imaging modalities, including mammography, MRI, CT, among others, have the capability to evaluate the malignancy of tumors in different anatomical locations, such as the liver, thyroid, brain, and prostate.

CONCLUSION

In this paper, we presented a wavelet based vision Transformer network used to predict breast tumor malignancy in ultrasound images. Since CNNs and vision-Transformers function differently, we used DWT in network input to extract tumor feature variability to characterize BUS tumors. Using a proposed network produces adequate results. Two datasets independently tested the proposed model, which outperformed previous techniques. With the UDIAT and Baheya datasets, it achieved AUC values of 0.984 and 0.968, respectively. The suggested approach will be used to determine the malignancy score for ultrasound images of kidney and cardiac.

ACKNOWLEDGMENT

Basic - clinical fusion innovation program of Xi'an Jiaotong University Health Science Center (No. YXJLRH2022052); Clinical research project of First Affiliated Hospital of Xi'an Jiaotong University (No. XJTU1AF-CRF-2022-026); National Natural Science Foundation of China (No.81872390); National Natural Science Foundation of China (No.82003047)

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Abdel-Nasser M, Melendez J, Moreno A, Omer OA, Puig D (2017). Breast tumor classification in ultrasound images using texture analysis and super-resolution methods. Engineering Applications of Artificial Intelligence 59:84–92.

Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020). Dataset of breast ultrasound images. Data in brief 28:104863.

Ali A, Touvron H, Caron M, Bojanowski P, Douze M, Joulin A, Laptev I, Neverova N, Synnaeve G, Verbeek J, *et al.* (2021). Xcit: Cross-covariance image transformers. Advances in neural information processing systems 34:20014–27.

Byra M (2021). Breast mass classification with transfer learning based on scaling of deep representations. Biomedical Signal Processing and Control 69:102828.

Byra M, Andre M (2019). Breast mass classification in ultrasound based on kendall's shape manifold. arXiv preprint arXiv190511159 .

Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M (2019). Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Medical physics 46:746–55.

Chowdhury A, Razzaque RR, Muhtadi S, Shafiullah A, Abir EUI, Garra BS, Alam SK (2022). Ultrasound classification of breast masses using

a comprehensive nakagami imaging and machine learning framework. Ultrasonics 124:106744.

Das A, Rana S (2021). Exploring residual networks for breast cancer detection from ultrasound images. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE.

Daubechies I (1990). The wavelet transform, time-frequency localization and signal analysis. IEEE transactions on information theory 36:961–1005.

Duffy SW, Tabár L, Chen HH, Holmqvist M, Yen MF, Abdsalah S, Epstein B, Frodis E, Ljungberg E, Hedborg-Melander C, *et al.* (2002). The impact of organized mammography service screening on breast carcinoma mortality in seven swedish counties: A collaborative evaluation. Cancer Interdisciplinary International Journal of the American Cancer Society 95:458–69.

Fan Z, Gong P, Tang S, Lee CU, Zhang X, Song P, Chen S, Li H (2023). Joint localization and classification of breast masses on ultrasound images using an auxiliary attention-based framework. Medical image analysis 90:102960.

Ge S, Ye Q, Xie W, Sun D, Zhang H, Zhou X, Yuan K (2023). Ai-assisted method for efficiently generating breast ultrasound screening reports. Current Medical Imaging 19:149–57.

Gheflati B, Rivaz H (2022). Vision transformers for classification of breast ultrasound images. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE.

Hassanien MA, Singh VK, Puig D, Abdel-Nasser M (2022). Predicting breast tumor malignancy using deep convnext radiomics and quality-based score pooling in ultrasound sequences. Diagnostics 12:1053.

Kalafi EY, Jodeiri A, Setarehdan SK, Lin NW, Rahmat K, Taib NA, Ganggayah MD, Dhillon SK (2021). Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. Diagnostics 11:1859.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017). A survey on deep learning in medical image analysis. Medical image analysis 42:60–88.

Luo Y, Huang Q, Li X (2022). Segmentation information with attention integration for classification of breast tumor in ultrasound image. Pattern Recognition 124:108427.

Mo Y, Han C, Liu Y, Liu M, Shi Z, Lin J, Zhao B, Huang C, Qiu B, Cui Y, *et al.* (2023). Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images. IEEE Transactions on Medical Imaging .

Moon WK, Lee YW, Ke HH, Lee SH, Huang CS, Chang RF (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Computer methods and programs in biomedicine 190:105361.

Nemat H, Fehri H, Ahmadinejad N, Frangi AF, Gooya A (2018). Classification of breast lesions in ultrasonography using sparse logistic regression and morphology-based texture features. Medical physics 45:4112–24.

Ning Z, Tu C, Xiao Q, Luo J, Zhang Y (2020). Multi-scale gradational-order fusion framework for breast lesions classification using ultrasound images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition.

Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H (2023). Transformers in medical imaging: A survey. Medical Image Analysis :102802.

Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence.

Tan M, Le Q (2021). Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. PMLR.

Vigil N, Barry M, Amini A, Akhloufi M, Maldague XP, Ma L, Ren L, Yousefi B (2022). Dual-intended deep learning model for breast cancer diagnosis in ultrasound imaging. Cancers 14:2663.

Wang L (2017). Early diagnosis of breast cancer. Sensors 17:1572.

Wei M, Du Y, Wu X, Su Q, Zhu J, Zheng L, Lv G, Zhuang J (2020). A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images. Computational and Mathematical Methods in Medicine 2020.

Woo S, Debnath S, Hu R, Chen X, Liu Z, Kweon IS, Xie S (2023). Convnext v2: Co-designing and scaling convnets with masked autoencoders.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Xie S, Girshick R, Dollár P, Tu Z, He K (2017). Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition.

Yap MH, Pons G, Martí J, Ganau S, Sentis M, Zwiggelaar R, Davison AK, Marti R (2017). Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics 22:1218–26.

Zourhri M, Hamida S, Akouz N, Cherradi B, Nhaila H, El Khaili M (2023). Deep learning technique for classification of breast cancer using ultrasound images. In: 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). IEEE.