Breast Tumor Classification Using Dynamic Ultrasound Sequence Pooling and Deep Transformer Features

Mohamed A Hassanien[1], Vivek Kumar Singh[2], Mohamed Abdel-Nasser[3], Domenec Puig[4]
Department of Computer Engineering and Mathematics, University Rovira i Virgili, Tarragona, Spain 43007[1,4]
Barts Cancer Institute (BCI), Queen Mary University of London, London, UK[2]
Electrical Engineering Department-Faculty of Engineering, Aswan University, Aswan, Egypt[3]

***Abstract*—Breast ultrasound (BUS) imaging is widely utilized for detecting breast cancer, one of the most life-threatening cancers affecting women. Computer-aided diagnosis (CAD) systems can assist radiologists in diagnosing breast cancer; however, the performance of these systems can be degrade by speckle noise, artifacts, and low contrast in BUS images. In this paper, we propose a novel method for breast tumor classification based on the dynamic pooling of BUS sequences. Specifically, we introduce a weighted dynamic pooling approach that models the temporal evolution of breast tissues in BUS sequences, thereby reducing the impact of noise and artifacts. The dynamic pooling weights are determined using image quality metrics such as blurriness and brightness. The pooled BUS sequence is then input into an efficient hybrid vision transformer-CNN network, which is trained to classify breast tumors as benign or malignant. Extensive experiments and comparisons on BUS sequences demonstrate the effectiveness of the proposed method, achieving an accuracy of 93.78%, and outperforming existing methods. The proposed method has the potential to enhance breast cancer diagnosis and contribute to lowering the mortality rate.**

Keywords*—*Breast ultrasound; breast cancer; CAD systems; deep learning; vision transformer

I. Introduction

Breast cancer in women is one of the most life-threatening cancers worldwide [1], [2]. Early detection significantly reduces the mortality rate, and mammography, an X-ray imaging technique of the breast, remains the gold standard for population-based breast cancer screening. Despite its effectiveness in detecting breast cancer, mammography has limitations, including low sensitivity and high false-positive rates, where normal cases are incorrectly classified as cancerous. To address these limitations, alternative imaging technologies such as magnetic resonance imaging (MRI), 3D tomosynthesis, and ultrasound are often used [3].

Breast ultrasound (BUS) imaging has been effectively used in the detection and diagnosis of breast cancer, specially in the case of women having dense breast tissue or with cases who are at high risk of developing breast cancer [4], [5]. The main merits of BUS imaging are non-invasive and non-ionizing technology, widely available and cost-effective solution, and capable of producing real-time images, which can enhance breast cancer detection sensitivity.

In the last two decades, various computer-aided detection (CAD) systems have been developed for breast cancer detection. In particular, artificial intelligence (AI) based CAD systems have helped in detecting breast cancer early by assisting radiologists in interpreting medical images, including BUS images [6]. AI-powered CAD systems can analyze images quickly and accurately, detecting subtle abnormalities and highlighting region of interest (ROIs), thereby boosting the sensitivity and specificity of breast cancer detection. It should be noted that BUS imaging has some limitations, notably poor contrast, speckle noise, and shadowing artifacts, which can degrade image quality and complicate interpretation (see Fig. 1). These issues make it challenging to differentiate between various tissues and structures. Additionally, BUS is operator-dependent, with image quality varying based on the skill and experience of the sonographer. This highlights the need for effective image processing, noise mitigation techniques, and robust AI-based image classification models, to improve the performance of these CAD systems in breast cancer detection.

In recent years, deep learning has significantly enhanced the automated analysis of BUS images over the past decade by extracting powerful representations from them. This has led to the development of several deep-learning aid (DLA) tools for detecting breast cancer and distinguishing between benign and malignant tumors. Recently, several DLA tools has been proposed, for instance, Ellis et al. of [7] explored deep learning as a classification tool for detecting cancerous ultrasound breast images, aiming to develop a simple, mobile-based classifier. With ResNet50, the CAD system achieved an accuracy of approximately 64% with minimal images, and up to 78% when pretrained. The authors of [8] introduced a novel few-shot learning approach for classifying ultrasound breast cancer images, leveraging the power of meta-learning techniques. Specifically, the authors employed prototypical networks and model-agnostic meta-learning (MAML) algorithms to enable our model to learn from limited data and adapt to new, unseen breast cancer images. Lanjewar et al. [9] integrated three widely used pretrained Convolutional Neural Network (CNN) models, namely, MobileNetV2, ResNet-50, and VGG16 with a long short term memory (LSTM) to extract features from BUS images. The authors used the synthetic minority over-sampling with Tomek (SMOTETomek) method in order to balance the number of extracted features. With the VGG16 model, they achieved an F1 score of 99.0%, Kappa coefficient of 98.9%, and an area under the curve (AUC) of 1.0.

The majority of existing studies have focused on classifying breast tumors using only one ultrasound image per tumor, whereas some studies, such as [10], [11], have utilized BUS sequences for detecting breast cancer malignancy. In this paper, we consider the quality of BUS images when designing the classification model. In particular, we propose a novel approach for breast tumor classification utilizing dynamic pooling of BUS image sequences. Specifically, this new method captures the temporal evolution of breast tissues in BUS sequences, effectively mitigating the influence of noise and artifacts. The dynamic pooling weights are computed based on image quality metrics, including blurriness and brightness. The resulting pooled BUS sequence is processed by the MobileViTv3 network, which is trained to classify breast tumors as either benign or malignant.

The remainder of this research is organized as follows: Section II reviews related work on breast lesion classification in ultrasound images. Section III details the proposed method. Section IV presents and discusses the experimental results. Finally, Section V concludes the study and provides suggestions for future work.

II. Related Work

It should be noted that most existing breast cancer CAD systems are trained to receive one ultrasound image (OUI) to determine whether it is benign or malignant [12], [13], [14], [15]. For instance, He et al. [16] proposed a new method for breast cancer classification using a wavelet-based vision transformer network. By incorporating the discrete wavelet transform (DWT) into the network input, we enhance the neural network's receptive fields, enabling the capture of significant features in the frequency domain. The proposed model effectively captures intricate characteristics of breast tissue, allowing for accurate breast cancer classification with high precision and efficiency. We evaluated the model using two breast tumor ultrasound datasets, comprising 780 cases from Baheya hospital in Egypt and 267 patients from the UDIAT Diagnostic Centre of Sabadell in Spain. The results show that the proposed transformer network achieves outstanding performance in breast cancer classification, with an AUC scores of 0.984 and 0.968 on both datasets.

Some recent studies showed that BUS sequences may give better detection results. For instance, the authors of [11] proposed a four-stage CAD system: super-resolution calculation, ROI extraction, feature extraction, and classification. The authors used five manually designed features, derived from various image analysis techniques, including GLCM, LBP, HOG, phase congruency-based LBP, and pattern lacunarity spectrum, to classify breast tumors into malignant and benign categories from a BUS image. However, this conventional approach has several limitations, including being computationally time-consuming, less resilient, and requiring specific feature choices and preprocessing activities.

To handle this issues mentioned above, recent studies employed deep learning networks for feature extraction. For instance, Yang et al. [17] presented a temporal sequence dual-branch network (TSDBN) breast cancer classification based on BUS and contrast-enhanced ultrasound (CEUS) sequences. It has two branches: one for BUS sequences and the other for CEUS sequences. In the branch of the BUS sequences, the ResNeXt-18 is employed. In the other branch, temporal sequence regression and a shuffle temporal sequence mechanisms are employed to enhance the temporal features of CEUS sequnces. They used a private dataset to evaluate their method. The dataset has 268 samples: 146 malignant and 122 benign. For each case, the BUS and CEUS sequences were recorded. TSDBN achieved an accuracy of 92.2%. One of the main limitation of this method is that it requires US and CEUS sequences for the same cases, which may not be available. Also, it does not consider the effect of noise and artifacts in US and CEUS sequences in the classification results. The study of [18] used 3D ResNet-50 to classify breast lesions in BUS sequences and 2D ResNet-50 to classify the same lesions in static images, finding that the BUS sequences lead to a higher AUC value of 0.969.

Han et al. [19] presented a ResViT model that combines residual neural networks with vision transformer to extract features from CEUS sequences, and employed a temporal segment network (TSN) to aggregate the spatio-temporal features of all frames in the input sequences. They achieved an accuracy of 78.79% with a private CEUS sequence dataset. Zhang et al. [20] proposed a segment-attention generator (SAG) module that can help deep learning classification models to focus on BUS sequence segments that have clear appearances for classifying breast lesions. The study of [10] introduced a deep-learning-based radiomics approach utilizing BUS sequences, comprising three key components. The ConvNeXt network, a deep CNN trained in the vision transformer style, is employed for radiomic feature extraction. An efficient pooling mechanism is also proposed to combine the malignancy scores of each breast US sequence frame, based on image-quality statistics. Finally, visual interpretations are provided to facilitate understanding. However, this methods achieved acceptable results, there is a big room for further enhancing the classification accuracy.

As mentioned earlier, a common limitation of existing studies is that they neglect the temporal information and image quality of BUS videos when developing classification models. Moreover, they rely on a single BUS image to develop their methods. However, the noisy nature of BUS images, the similarity between normal and abnormal tissues, and the degradation of image quality due to dense breast fat and glandular tissue, which attenuate ultrasonic waves, make accurate diagnosis challenging. These issues pose a significant challenge to building a robust BUS image classification model. The proposed method consider that the temporal information embedded in the BUS sequences and mitigate the effect of the noise utilizing the the weighted dynamic pooling technique.

III. Material and Method

A. Dataset

The proposed CAD system was developed and evaluated using a database consisting of 31 malignant and 28 benign BUS sequences, with each sequence corresponding to a single patient. The BUS sequences created by the Engineering Department of Cambridge University[1]. This dataset is a subset of a larger clinical database of ultrasonic radiofrequency strain imaging data, which was created by the Engineering Department at Cambridge University. The dataset includes 3911

[1] http://mi.eng.cam.ac.uk/research/projects/elasprj/

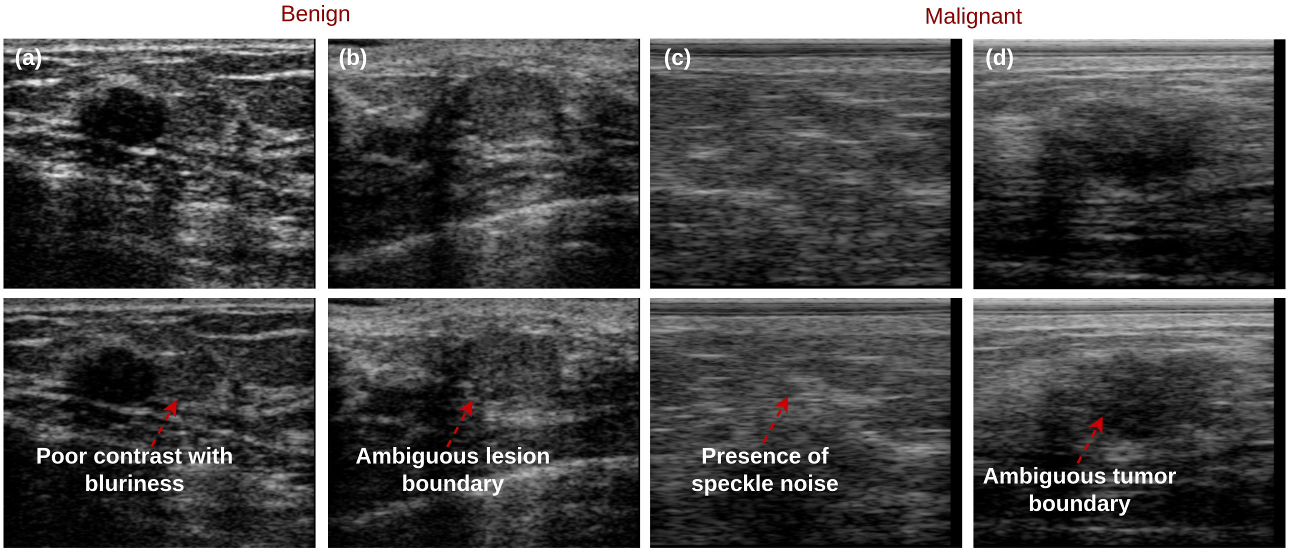
Benign
Malignant
(a)
(b)
(c)
(d)
Poor contrast with bluriness
Ambiguous lesion boundary
Presence of speckle noise
Ambiguous tumor boundary

Fig. 1. Examples of BUS images having malignant tumors and benign lesions with various artifacts and challenges like poor contrast, ambiguous lesion or tumor boundaries, and speckle noise.

images containing benign tumors and 5245 images having malignant tumors.

B. Proposed Method

Fig. 2 presents the key components of the proposed method: 1) generating dynamic BUS image from the input BUS sequence, 2) applying transfer learning on MobileViTv3 [21] to extract local and global features from BUS images to differentiate between benign and malignant tumors, and 3) employing different visual interpretation methods to explain the decisions of the classification model.

1) Dynamic BUS sequence pooling: Let $S = [s_1, s_2, \ldots, s_M]$ is an input BUS sequence, where s_i

$$q_t = \frac{1}{W} \sum_{i=t}^{t+W} s_i. \tag{1}$$

where W is a time window.

TVA can be expressed as follows:

$$q_t = \frac{\frac{1}{t} \sum_{i=1}^{t} s_i}{\| \frac{1}{t} \sum_{i=1}^{t} s_i \|}. \tag{2}$$

After obtaining the smoothed version of BUS images, a rank-pooling method can be employed to learn the relative ranks of the BUS images in the input sequence, for instance q_n comes after q_{n-1}, q_{n-1} comes after q_{n-2}, and so on. This relative ranks can expressed as follows:

$$q_n \succ q_{n-1} \succ q_{n-2} \cdots q_1 \tag{3}$$

The rank-pooling technique is used to learn pairwise linear functions $\rho(q_t; \alpha)$, where $\alpha \in R^D$. The ranking score of q_t is computed as $\rho(q_t; \alpha) = \alpha^T . q_t$.

The parameters α of $\rho(q_t; \alpha)$ are optimized using the following objective function [22]:

$$\arg\min_{\alpha} \quad \frac{1}{2} \|\alpha\|^2 + \delta \sum_{\forall i,j, q_i \succ q_j} \theta_{ij}, \quad s.t. \quad \alpha^T (q_{t_i} - q_{t_j}) \geq 1 - \theta_{ij}, \quad \theta_{ij} \geq 0, \tag{4}$$

In this expression, δ stands for the regularization parameter and θ stands for the tolerance margin. The term $\{\alpha^T (q_{t_i} - q_{t_j}) \geq 1 - \theta_{ij}\}$ represents the the constraint of the objective function $\forall t_i, t_j \quad q_{t_i} \succ q_{t_j} \iff \alpha^T . q_{t_i} \succ \alpha^T . q_{t_j}$.

2) Image quality-aware dynamic BUS image generation: The quality of the BUS images can considered when generating the dynamic BUS image from the input BUS sequence by modifying Eq. 1 and Eq. 2 as follows:

$$q_t = \frac{1}{W} \sum_{i=t}^{t+W} \omega_i . s_i. \tag{5}$$

$$q_t = \frac{\frac{1}{t} \sum_{i=1}^{t} \omega_i . s_i}{\| \frac{1}{t} \sum_{i=1}^{t} \omega_i . s_i \|}. \tag{6}$$

where ω_i represents the quality of the th BUS image in the input sequence. The value of w_i may be 0 or 1, where a value of 1 denotes that the BUS image quality exceeds the thresholds of the BUS image quality metrics.

In this study, two efficient general-purpose image quality assessment metrics are used to estimate the quality of BUS images, namely the brightness and blurriness [23], [24], [25].

BUS image blurriness metric: Here, we employ the image blurriness measure presented in [24], where a Gaussian filter

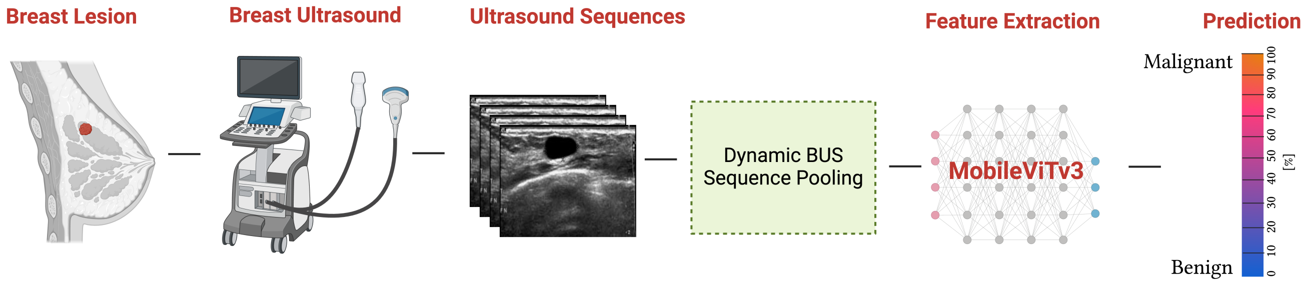

Fig. 2. Overview of the proposed method which consists of three consecutive steps: dynamic BUS sequence pooling, feature extracting using MobileViTv3 and malignancy classification.

is used to suppress the noise from the input image. Let I is a BUS image, the Gaussian filter can be written as:

$$f(a,b) = \frac{1}{(2\pi\sigma^2)} e^{-\frac{\left(a^2+b^2\right)}{2\sigma^2}}, \tag{7}$$

Here, σ is the standard deviation of the Gaussian distribution, and a and b are the coordinates of I.

After suppressing the noise, the variance of Laplacian operator is computed and used as a blurriness score. The 2D Laplacian operator can be expressed as follows:

$$\nabla^2 I(a,b) = \frac{\partial^2 I}{\partial a^2} + \frac{\partial^2 I}{\partial b^2}, \tag{8}$$

BUS images with a blurriness score lower than a threshold are considered as blurry images. Following [24], the blurriness threshold is set to the average blurriness value of benign and malignant BUS images from the training dataset.

BUS image brightness/darkness metric: In this study, we employ the brightness estimation algorithm proposed by Bezryadin et al. [25] as a BUS image quality metric. Following the study of [10], we selected the range from 10 to 30 for the brightness score.

3) Feature extraction: The main powerful approaches to extract features from images are CNNs and ViTs. In the context of breast tumors in ultrasound images, previous studies such as [9], [7] focused in the use of CNN models, [26], [27] used vision transformers, while [28] combined the decisions of different CNN and vision transformers. ViTs produce features representing global information in the images, due to their self-attention mechanism. CNNs extract local features in the images. Several hybrid models have emerged, integrating the strengths of both CNNs and ViTs into a single architecture. By combining the self-attention mechanism of ViTs, which excels at capturing long-range dependencies, with the local kernels of CNNs, which are adept at extracting local information, these models aim to achieve superior performance on various vision tasks. In order to extract robust descriptors from BUS images, in this paper we employ one of the most effective deep learning model that combines CNNs and ViTs, namely, MobileViTv3 [21].

Fig. 3 shows the block diagram of the MobileViTv3 that contains three blocks: local representation (LR) block, global representation (GR) block, and fusion block. The LR block (CNN components) consists of two layers: a 3×3 depthwise convolution layer and a 1×1 convolution layer. The GR block (ViT components) includes N linear transformations (self-attention). The fusion block uses 1×1 convolution layer to fuse the local and global features.

This study involves adapting and training various self-attention based deep vision transformer architectures to extract robust features, which can classify breast cancers as benign or malignant and predict the malignancy score of each input ultrasound image. By leveraging the transfer learning theory, the pre-trained vision transformer network and its parameters can be fine-tuned and applied to the target breast ultrasound dataset, enabling effective knowledge transfer. A support vector machine (SVM) classifier with a radial basis function (RBF) is used for classification.

4) Visual interpretation: To produce visual interpretations (explanations) for the proposed breast tumor classification model, this study employs the Grad-CAM (Gradient-weighted Class Activation Mapping) [29] and Local Interpretable Model-agnostic Explanations (LIME) techniques [30]. Let's denote the input image as I, the class of interest (benign or malignant) as c, and the output probability of the class as $P(c|I)$. The goal of Grad-CAM is to generate a heatmap $L^c_{Grad-CAM}$ that highlights the important regions of each BUS image that contribute to the prediction of the proposed model.

LIME is a model-agnostic method, which works by generating a dataset of similar instances around a specific instance for which we want to understand the model's prediction. Let's denote the original machine learning model as f, and the instance for which we want to explain the prediction as $\mathbf{x}$. LIME generates a dataset of m perturbed instances around $\mathbf{x}$, denoted as $\mathbf{x}'$, by randomly sampling from a distribution $\pi(\mathbf{x}'|\mathbf{x})$. The perturbed instances $\mathbf{x}'$ are then used to generate a new dataset $\mathcal{D} = (\mathbf{x}'_1, f(\mathbf{x}'_1)), \ldots, (\mathbf{x}'_m, f(\mathbf{x}'_m))$. Next, LIME trains an interpretable model g (e.g. a linear model) on the dataset $\mathcal{D}$ to approximate the behavior of the original model f locally around $\mathbf{x}$.

C. Evaluation metrics

To assess the performance of the proposed method for breast tumor classification in BUS sequences, we employ four well-known evaluation metrics, namely, the accuracy, precision, recall, and F1-score. The mathematical expression

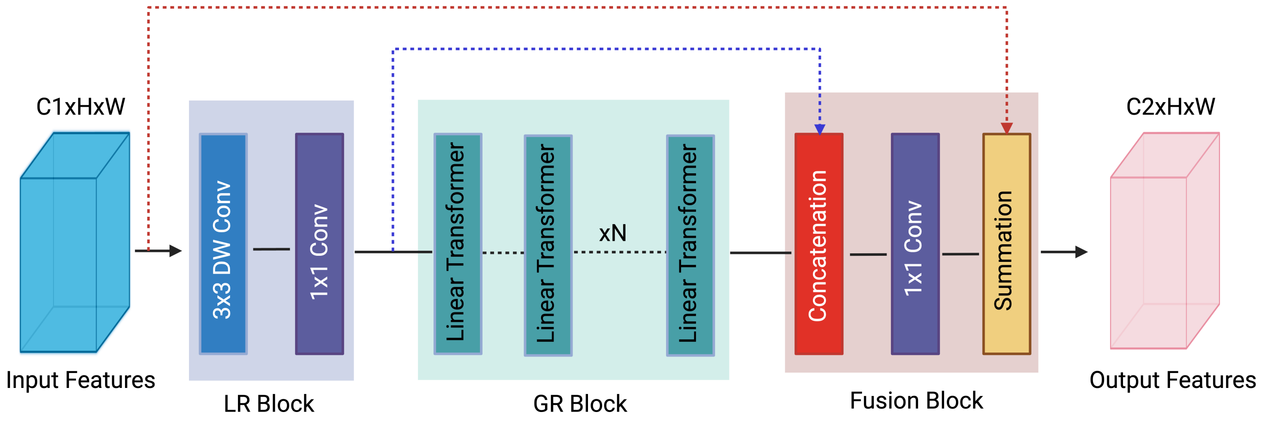

Fig. 3. Overview of the block diagram of the MobileViTv3 that consists of three interconnected blocks: LR, GR, and fusion.

of each evaluation metric is given below:

$$\text{Accuracy} = \frac{TP+TN}{P+N} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{11}$$

$$\text{F1-score} = \frac{TP}{TP+0.5(FP+FN)} \tag{12}$$

In these expressions, TP and TN denote the number of BUS sequences having malignant and benign tumors that have been successfully detected by the proposed method, respectively. Conversely, FN stands for the number of malignant tumors wrongly identified by the proposed method as benign tumors. FP stands for the number of benign tumors wrongly identified by the proposed method as malignant tumors.

IV. Experimental Results

A. Training Details

As the BUS images varied in size, all were resized to 224×224 pixels. The AdamW optimizer was employed with an initial learning rate of 0.001, a weight decay of 0.01, and a cosine learning rate scheduler, using binary cross-entropy loss to optimize the model. The training process was conducted with a batch size of two images over 50 epochs. To augment the training data, the input BUS sequences were split into overlapping sub-sequences with a window size of 20, generating multiple dynamic BUS images from each sub-sequence. Additional data augmentation techniques were applied, including 90-degree image rotation, 0.2 image scaling, horizontal flipping (with a probability of 0.5), median filter blurring, and contrast-limited adaptive histogram equalization. All models were developed in Python using the PyTorch framework and trained on an NVIDIA GeForce GTX 1070Ti GPU with 8 GB of RAM.

B. Results

Table I compares various backbone feature extractors, including CNNs, vision transformers, and hybrid models. MobileViTv3-S emerges as the top-performing model, significantly outperforming others with an accuracy of 89.33%, which is more than 1.2% higher than the second-best model, ConvNeXt V2. This demonstrates the strength of hybrid architectures like MobileViT, which combine the local feature-capturing ability of CNNs with the global context awareness of transformers. Despite its depth, ResNet-150 performs good but does not achieved similar results as MobileViTv3-S, recommending that deeper CNNs do not necessarily yield better performance in this task.

The transformer-based models, DEiT and BEiT v2, show satisfactory performance, indicating that transformers may require more fine-tuning for optimal results. XCiT performs the weakest, further highlighting the limitations of transformers without additional optimization. In contrast, the MobileViTv3 family, particularly MobileViTv3-S, shows the advantage of hybrid architectures, offering the best balance between efficiency and accuracy. Even the smaller versions, MobileViTv3-XS and XXS, perform well, making them suitable for resource-constrained situation while still providing competitive performance.

Table II presents the affect of various smoothing techniques on the performance of the proposed method for classifying breast ultrasound images into benign and malignant. Firstly, without applying any smoothing, the model achieved satisfactory results in classifying between the two types of lesions. When the MA smoothing technique was applied, there was a slight improvement in the model's ability to accurately classify the ultrasound images.

However, the most important advancement was observed when the TVA smoothing technique employed. TVA leads to a significant improvement in the classification performance, enhancing the accuracy, precision, recall, and F1-score by approximately 2%. This suggests that TVA allows the model to more effectively capture and utilize the subtle variations in ultrasound imaging data, which is critical for differentiating between benign and malignant lesions. TVA's ability to smooth

TABLE I. Results of the Proposed Method with Various Backbone Feature Extractors without Smoothing

Method	Accuracy	Precision	Recall	F1-score
ResNet-152 [31]	85.91	85.18	84.66	84.91
MobileViTv3-S [21]	89.33	88.90	88.79	88.84
MobileViTv3-XS [21]	86.55	84.61	84.02	84.31
MobileViTv3-XXS [21]	83.76	81.97	80.35	81.15
ConvNeXt V2 [32]	88.10	88.22	86.05	87.12
DEiT [33]	86.75	85.82	84.95	85.38
BEiT v2 [34]	85.46	84.24	83.74	83.98
XCiT [35]	83.67	82.54	81.78	82.15

the data while preserving key features seems to help the model focus on more relevant regions, thus improving overall diagnostic accuracy and reducing the risk of misclassification, which is crucial in breast cancer detection.

TABLE II. Results of the Proposed Method with Different Smoothing Methods

Method	Accuracy	Precision	Recall	F1-score
W/o smoothing	89.33	88.90	88.79	88.84
MA	89.56	88.94	88.86	88.89
TVA	**91.58**	**90.76**	**90.11**	**90.43**

Table III compares the performance of the proposed method with and without the use of quality weights, evaluating key metrics. The inclusion of quality weights clearly improves performance across all metrics. When quality weights are not applied, the method achieves an accuracy of 91.58%, with a precision of 90.76%, recall of 90.11%, and F1-score of 90.43%. These are strong results, indicating that the model can effectively make predictions, but there is room for improvement in its ability to generalize and balance precision and recall.

TABLE III. Results of the Proposed Method with and without Quality Weights

Method	Accuracy	Precision	Recall	F1-score
w/o quality weights	91.58	90.76	90.11	90.43
with quality weights	**93.78**	**93.65**	**92.94**	**93.29**

When quality weights are introduced, the performance improves substantially, with accuracy increasing to 93.78%, a significant boost over the baseline. Similarly, precision rises to 93.65%, recall to 92.94%, and the F1-score to 93.29%. This improvement can be attributed to the model's ability to assign higher importance to more informative lesion or tumor related features during training, resulting in more refined feature extraction and better classification outcomes. The use of quality weights enhances the model's ability to focus on higher-quality data, leading to better overall predictions and higher consistency in its results. Fig. 4 shows the area under the receiver operating characteristic (AUROC) scores of 0.94 and 0.97 for the model without and with quality weights, respectively.

Fig. 5 shows the explainability of the proposed model using GradCam and LIME. The red refers to a higher probability of the presence of lesion or tumor, while the blue represents a lower probability of the existence of background region. Based on visual inspection, the model correctly identified both benign lesions [Fig. 5(a), (b)] by focusing on hypoechoic regions.

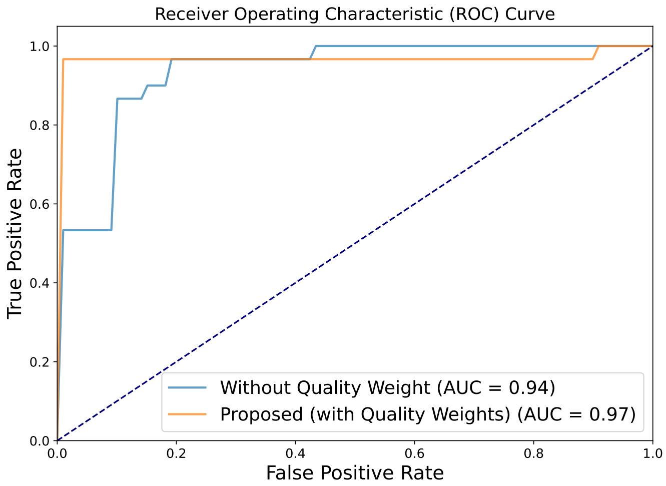

Fig. 4. The AUROC curves of the proposed method with and without quality weights.

Additionally, small tumors [Fig. 5(c), (d)] with ambiguous boundaries were accurately classified as malignant, with the model highlighting critical regions while ignoring background pixels.

C. Comparisons with Related Methods and Discussion

Table IV compares the performance of the proposed method with an existing method from the literature, specifically the method from [10]. The proposed method achieves the highest performance across all metrics, with an accuracy of 93.78%, precision of 93.65%, recall of 92.94%, and F1-score of 93.29%. These results demonstrate a clear improvement over the existing method, which, while still effective, yields slightly lower accuracy (91.66%) and F1-score (92.33%).

The improved performance of the proposed method can be attributed to its ability to better capture important features and balance precision and recall. The higher F1-score indicates that the proposed method handles the trade-off between precision and recall more effectively, resulting in more accurate and reliable predictions. In comparison, while [10] method performs well, it falls short in terms of overall accuracy and F1-score, suggesting that the proposed method offers a more refined approach to the problem.

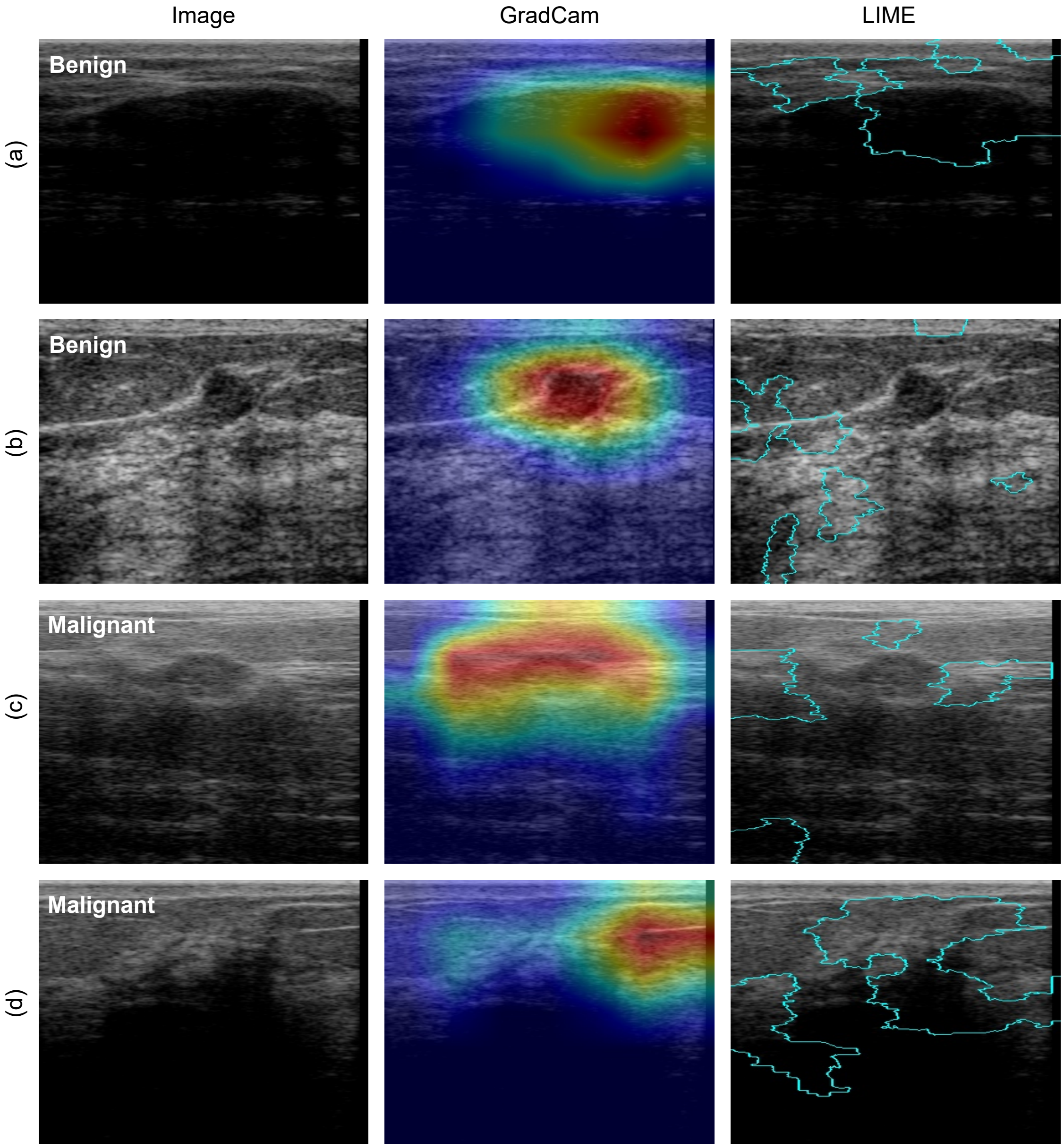

Fig. 5. Explanation of the proposed model using Grad-CAM [29] and LIME [30] methods. (a,b) benign cases, and (c,d) malignant cases.

TABLE IV. State-of-the-art Results Comparison

Method	Accuracy	Precision	Recall	F1-score
Proposed	**93.78**	**93.65**	**92.94**	**93.29**
[10]	91.66	93.05	92.69	92.33

V. Conclusion and Future Work

This paper presents a novel approach to breast tumor classification using dynamic pooling of BUS sequences, combining the strengths of both CNN and transformer architectures. By incorporating weighted dynamic pooling based on image quality metrics, such as blurriness and brightness, our method effectively mitigates the impact of noise and artifacts commonly found in BUS images. Comprehensive experiments demonstrate that our approach, particularly when using MobileViTv3-S, significantly outperforms existing methods, achieving an accuracy of 93.78%. The inclusion of quality weights further enhances classification performance, highlighting the importance of prioritizing high-quality image frames. Not only does our model achieve higher accuracy, but it also provides better interpretability through Grad-CAM visualizations, facilitating the understanding of tumor characteristics. The results suggest that our approach can offer a robust, reliable, and interpretable solution for breast cancer detection in clinical settings.

One limitation of this study is the reliance on a single BUS video sequence dataset to evaluate the efficacy of the proposed method. Additionally, the small sample size of the dataset presents another limitation.

Future work will focus on integrating additional ultrasound modalities (e.g. BUS and CEUS) with the proposed method to further enhance classification accuracy. Additionally, larger and more diverse datasets will be collected to improve the robustness and performance of the developed classification models.

Acknowledgment

This research was partly supported by Spanish Government through Project TED2021-130081B-C21, Project PDC2022-133383-I00, and Project PID2019-105789RB-I00.

References

[1] J. D. B. Fuentes, E. Morgan, A. de Luna Aguilar, A. Mafra, R. Shah, F. Giusti, J. Vignat, A. Znaor, C. Musetti, C.-H. Yip *et al.*, “Global stage distribution of breast cancer at diagnosis: a systematic review and meta-analysis,” *JAMA oncology*, 2024.

[2] R. L. Siegel, A. N. Giaquinto, and A. Jemal, “Cancer statistics, 2024.” *CA: a cancer journal for clinicians*, vol. 74, no. 1, 2024.

[3] M. I. Tsarouchi, A. Hoxhaj, and R. M. Mann, “New approaches and recommendations for risk-adapted breast cancer screening,” *Journal of Magnetic Resonance Imaging*, vol. 58, no. 4, pp. 987–1010, 2023.

[4] C. Shan, T. Tan, J. Han, and D. Huang, “Ultrasound tissue classification: a review,” *Artificial Intelligence Review*, vol. 54, no. 4, pp. 3055–3088, 2021.

[5] K. Malherbe and D. Tafti, “Breast ultrasound,” *StatPearls*, 2024.

[6] M. Bahl, J. M. Chang, L. A. Mullen, and W. A. Berg, “Artificial intelligence for breast ultrasound: Ajr expert panel narrative review,” *American Journal of Roentgenology*, 2024.

[7] J. Ellis, K. Appiah, E. Amankwaa-Frempong, and S. C. Kwok, “Classification of 2d ultrasound breast cancer images with deep learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5167–5173.

[8] G. Işık and İ. Paçal, “Few-shot classification of ultrasound breast cancer images using meta-learning algorithms,” *Neural Computing and Applications*, pp. 1–13, 2024.

[9] M. G. Lanjewar, K. G. Panchbhai, and L. B. Patle, “Fusion of transfer learning models with lstm for detection of breast cancer using ultrasound images,” *Computers in Biology and Medicine*, vol. 169, p. 107914, 2024.

[10] M. A. Hassanien, V. K. Singh, D. Puig, and M. Abdel-Nasser, “Predicting breast tumor malignancy using deep convnext radiomics and quality-based score pooling in ultrasound sequences,” *Diagnostics*, vol. 12, no. 5, p. 1053, 2022.

[11] M. Abdel-Nasser, J. Melendez, A. Moreno, O. A. Omer, and D. Puig, “Breast tumor classification in ultrasound images using texture analysis and super-resolution methods,” *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 84–92, 2017.

[12] S. Sushanki, A. K. Bhandari, and A. K. Singh, “A review on computational methods for breast cancer detection in ultrasound images using multi-image modalities,” *Archives of Computational Methods in Engineering*, vol. 31, no. 3, pp. 1277–1296, 2024.

[13] A. Sahu, P. K. Das, and S. Meher, “An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images,” *Biomedical Signal Processing and Control*, vol. 87, p. 105377, 2024.

[14] K. S. Rao, P. V. Terlapu, D. Jayaram, K. K. Raju, G. K. Kumar, R. Pemula, V. Gopalachari, and S. Rakesh, “Intelligent ultrasound imaging for enhanced breast cancer diagnosis: Ensemble transfer learning strategies,” *IEEE Access*, 2024.

[15] M. Ragab, A. O. Khadidos, A. M. Alshareef, A. O. Khadidos, M. Altwijri, and N. Alhebaishi, “Optimal deep transfer learning driven computer-aided breast cancer classification using ultrasound images,” *Expert Systems*, vol. 41, no. 4, p. e13515, 2024.

[16] C. He, Y. Diao, X. Ma, S. Yu, X. He, G. Mao, X. Wei, Y. Zhang, and Y. Zhao, “A vision transformer network with wavelet-based features for breast ultrasound classification,” *Image Analysis and Stereology*, vol. 43, no. 2, pp. 185–194, 2024.

[17] Z. Yang, X. Gong, Y. Guo, and W. Liu, “A temporal sequence dual-branch network for classifying hybrid ultrasound data of breast cancer,” *Ieee Access*, vol. 8, pp. 82 688–82 699, 2020.

[18] G. Zhao, D. Kong, X. Xu, S. Hu, Z. Li, and J. Tian, “Deep learning-based classification of breast lesions using dynamic ultrasound video,” *European Journal of Radiology*, vol. 165, p. 110885, 2023.

[19] M. Han, D. Guo, J. Yuan, and C. Lu, “A spatio-temporal feature fusion network for intelligent analysis of breast cancer contrast-enhanced ultrasound video,” in *Proceedings of the 2024 4th International Conference on Bioinformatics and Intelligent Computing*, 2024, pp. 270–274.

[20] Y. Zhang, D. Kong, J. Li, T. Yang, F. Yao, and G. Yang, “Using segment-level attention to guide breast ultrasound video classification,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[21] S. N. Wadekar and A. Chaurasia, “Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features,” *arXiv preprint arXiv:2209.15159*, 2022.

[22] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and computing*, vol. 14, pp. 199–222, 2004.

[23] V. K. Singh, B. Kucukgoz, D. C. Murphy, X. Xiong, D. H. Steel, and B. Obara, “Benchmarking automated detection of the retinal external limiting membrane in a 3d spectral domain optical coherence tomography image dataset of full thickness macular holes,” *Computers in Biology and Medicine*, vol. 140, p. 105070, 2022.

[24] L. Francis and N. Sreenath, “Pre-processing techniques for detection of blurred images,” in *Proceedings of International Conference on Computational Intelligence and Data Engineering*, 2019, pp. 59–66.

[25] S. Bezryadin, P. Bourov, and D. Ilinih, “Brightness calculation in digital image processing,” in *International Symposium on Technologies for Digital Photo Fulfillment*, vol. 2007, no. 1, 2007, pp. 10–15.

[26] M. A. Hassanien, V. Kumar Singh, D. Puig, and M. Abdel-Nasser, “Transformer-based radiomics for predicting breast tumor malignancy score in ultrasonography,” in *Artificial Intelligence Research and Development*. IOS Press, 2022, pp. 298–307.

[27] B. Gheflati and H. Rivaz, “Vision transformers for classification of breast ultrasound images,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 480–483.

[28] V. K. Singh, E. M. Mohamed, and M. Abdel-Nasser, “Aggregating efficient transformer and cnn networks using learnable fuzzy measure for breast tumor malignancy prediction in ultrasound images,” *Neural Computing and Applications*, vol. 36, no. 11, pp. 5889–5905, 2024.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[34] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022.

[35] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems*, vol. 34, pp. 20 014–20 027, 2021.