Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220351

Transformer-Based Radiomics for Predicting Breast Tumor Malignancy Score in Ultrasonography

Mohamed A. HASSANIEN [a,1], Vivek KUMAR SINGH [c], Domenec PUIG [a] and Mohamed ABDEL-NASSER [a,b]

[a] *Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Spain*
[b] *Electrical Engineering Department; Aswan University, Aswan, Egypt*
[c] *Queen's University Belfast, United Kingdom*

Abstract. Breast cancer must be detected early to reduce the mortality rate. Ultrasound images can make it easier for the clinician to diagnose cases of dense breasts. This study presents a deep vision transformer-based approach for predicting breast cancer malignancy scores from ultrasound images. In particular, various state-of-the-art deep vision transformers such as BEiT, CaiT, Swin, XCiT, and ViS-Former are adapted and trained to extract robust radiomics to classify breast tumors in ultrasound images as benign or malignant. The best-performing model is used to predict the malignancy score of each input ultrasound image. Experimental results revealed that the proposed approach achieves promising results for the detection of malignant tumors of the breast on ultrasound images.

Keywords. Radiomics, Breast cancer, Ultrasound imaging, CAD systems, Vision transformers

1. Introduction

Mammography is the most commonly used imaging technique to detect the early stages of breast cancer. Breast ultrasonography, however, substitutes mammography in the case of dense breasts as it can not penetrate through the tissue. In pregnant women, breast ultrasound is a viable alternative to mammography to prevent the use of radiation that can harm the fetus [1]. Computer-aided diagnostic (CAD) solutions help clinicians free themselves from processing multiple breast images of a patient, thereby improving the quality of clinical diagnostics [2,3,4]. The leading steps of a CAD system for classifying breast tumors with ultrasound images include region of interest (ROI) detection, tumor segmentation, feature extraction and selection, and machine learning-based classification model development.

Nevertheless, the accuracy of such CAD systems may be restricted due to the low signal-to-noise ratio (SNR) and the presence of artifacts such as speckle noise and shadows in the breast ultrasound images. Inadequate contact between the probe and the skin surface

[1] Corresponding Author; E-mail: mohamed.abdelhameedhassanien@estudiants.urv.cat.

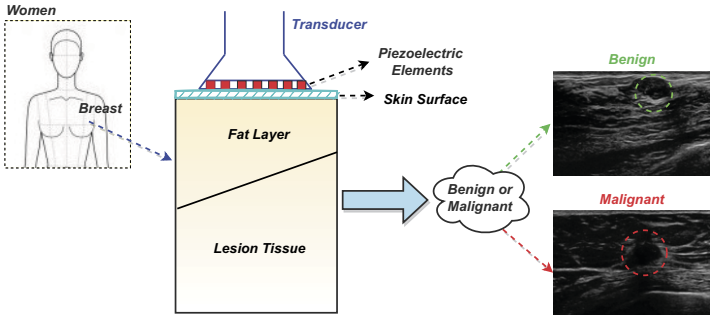


Figure 1. General description of breast tumor detection with breast ultrasound (BUS). Green and red dotted circles highlight the benign and malignant breast tumors respectively.

can cause a shadowing effect. These artifacts infer a clinical diagnosis, as it is difficult for a sonographer accurately acquire the appropriate breast tumor information.

Figure 1 shows a general description of breast tumor detection with breast ultrasound (BUS). Transducers with piezoelectric elements (128 or 192) generate sound waves reflected by the breast tissue to produce echoes. The ultrasound examiner or user must place the transducer correctly on the skin surface at the appropriate pressure to avoid the imaging artifact shadowing effects. The early layers of breast tissue contain subcutaneous fat and later depth approach to the lesion region. When an ultrasound scan is performed, the resulting B-mode image is examined by sonographers to see if the detected tumors are benign or malignant. Breast tumor is commonly recognized as a hypoechoic region than white breast tissue or light gray fat. The boundaries of most benign breast tumors are well defined and are round or oval in shape. In contrast, malignant tumor boundaries are often irregular and poorly defined, with lobules.

In the last two decades, many CAD systems have been proposed for breast cancer diagnosis. For instance, Shia et al. [5] used a deep residual network model to extract texture features of breast ultrasound images and then trained a support vector machine (SVM) to classify breast tumors as benign or malignant. The authors used an ultrasound dataset containing 302 benign and 241 malignant cases. The recommended method yielded a sensitivity score of 94.34% and a specificity score of 93.22%. In [6], Mishra et al. proposed a machine learning-based radiomics method classifying breast tumors with ultrasound. The authors extracted several hand-crafted features from the ultrasound images and later used recursive feature elimination to select the best set of those features. They also used synthetic minority oversampling techniques to address the problem of class imbalances. The employed texture features were Hu-moments based shape features, tumor shape features (area, convex area, eccentricity, solidity, EquivDiameter, extent, major axis length, minor axis length, orientation, and perimeter), histogram oriented gradients (HOG), and 13 grey-level co-occurrence matrix (GLCM). The authors employed an ultrasound dataset that included 437 benign, 210 malignant, and 133 normal cases in their experiments. They obtained results for accuracy and area under the curve (AUC) of 97.4% and 97%, respectively.

Zhuang et al. [7] applied the fuzzy enhancement, bilateral filtering, and image morphology operations on breast ultrasound images and the corresponding masks (binary image contains the segmented tumor). The authors concatenated various combinations of

the original and processed images, with each combination comprising three images (i.e., RGB channel fusion). The fused images were fed into pre-trained CNN models in order to extract feature vectors, which were subsequently merged using the adaptive spatial feature fusion approach. Finally, an artificial neural network (ANN) was used for the final classification. A dataset of 1328 breast ultrasound images was utilized to train and test this approach, which yielded an accuracy of 95.48%. Cui et al. [8] proposed two enhanced combined-tumoral region modules to gradually enhance the combined-tumoral features. Besides, the authors proposed a three-stream module for extracting and combining intratumoral, peritumoral, and combined tumor area features. The author used the channel attention module to adaptively combine the features of the three regions. The proposed method achieved a precision of 94.50% using the UDIAT dataset. Yu et al. [9] extracted discriminative regions from the input ultrasound images: the inner region, the marginal zone, and the posterior echo region of the lesion image. Then, they used an Inception-V3 pre-trained CNN model to extract texture features from the three regions and the whole image, followed by a principal components analysis (PCA) to reduce the dimensionality of the extracted features and an ensemble learning classifier for classifying the input image as benign or malignant. With a dataset of 479 cases, they achieved an accuracy of 85%.

Recently, several transformer-based methods have advanced exponentially for medical imaging tasks. Transformers have proven their capability to capture long-range dependencies and learn better relevant feature representations to be an alternative to convolutional neural networks. In the ultrasound domain, only limited research has been conducted to measure the effectiveness of transformer methods for classifying breast tumors in challenging conditions. This paper presents a deep learning-based radiomics approach for detecting breast tumor malignancy.Various self-attention based deep vision transformer architectures are adapted and trained to extract robust radiomics to classify breast cancers as benign or malignant, and predict the malignancy score of each input ultrasound image. Based on the transfer learning theory, the pretrained vision transformer network and parameters can be applied to this study's target breast ultrasound dataset by fine-tuning the selected vision transformer networks. Furthermore, we investigate the feasibility of incorporating the malignancy scores of the top-performing transformer model to enhance detection accuracy.

The rest of this paper is organized as follows. Section 2 details the proposed methodology. Section 3 contains experimental results and discussions.. We conclude our finding and suggest some future lines of research in Section 4.

2. Methodology

2.1. Breast cancer malignancy score prediction

Figure 2 presents the proposed radiomics approach for detecting breast tumor malignancy. A set of deep vision transformers are used to extract robust radiomics and classify breast tumors as benign or malignant. Various data augmentation techniques are used to increase the number of ultrasound images for training. The best model is identified based on different evaluation metrics. The extracted radiomics are used to compute the malignancy scores of breast cancer from the input ultrasound image.

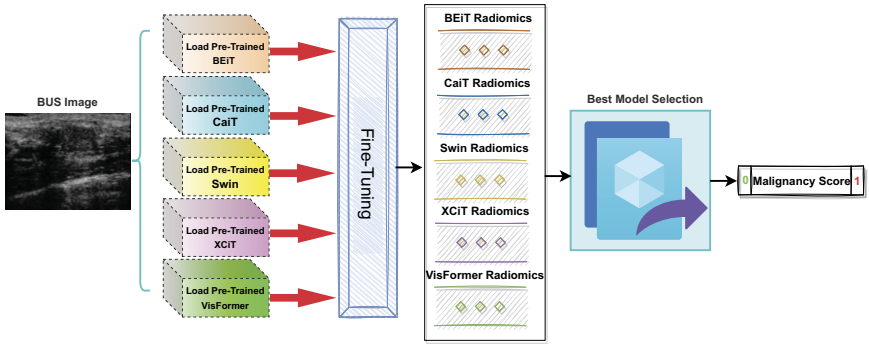


Figure 2. The pipeline of the proposed approach for breast tumor malignancy prediction in breast ultrasound images.

2.2. Input ultrasound images

The ultrasound images were taken from the UDIAT Diagnostic Centre of Sabadell (Spain) [10], [11]. It has a total of 163 breast ultrasound images extracted from 263 patients using a Siemens ACUSON Sequoia C512 system. It comprises two classes benign and malignant which have 110 and 53 breast ultrasound images, respectively. The annotation for each image of the lesion area was available.

2.3. Data augmentation

We applied several data augmentation techniques such as horizontal and vertical flipping with the probability of 0.5, scaling with 0.5, contrast limited adaptive histogram equalization (CLAHE), and rotation of 30 degrees which provides additional feature variability to the network. It is worth noting that all the trained transformer-based networks used the same data augmentation techniques.

2.4. Constructing transformer-based radiomics

In this study, five deep transformer models have been employed, namely, BEiT [12], CaiT [13], Swin [14], XCiT [15], and Visformer [16], for extracting deep learning-based radiomics for breast cancer malignancy prediction. Below, we briefly explain the architecture of each network.

- BEiT [12] is a self-supervised vision representation model from Image Transformers. In BEiT's pre-training, each Image has two views: image patches (16x16 pixels) and visual tokens (i.e., discrete tokens). Tokenization of the source image into visual tokens occurs first. The backbone Transformer is then fed with some randomly masked image patches. The pre-training goal is to use the corrupted image patches to recover the original visual tokens. BEiT's model parameters are fine-tuned on downstream tasks by appending task layers to the pre-trained encoder after it has been pre-trained. The implementation of the BEiT transformer is available at `https://github.com/microsoft/unilm/tree/master/beit`.

- CaiT [13] stands for class attention in image transformers. It is a way of enhancing the training of more profound architecture compared to other image transformer approaches. A LayerScale technique is employed in CaiT, where a learnable diagonal matrix is attached to the output of each residual block, in which some elements are initialized with values close to zero. The training dynamic is enhanced by attaching this basic layer after each residual block, allowing for the training of deeper image transformers. As a result, models whose performance does not saturate early with increased depth are produced. The implementation of the CaiT transformer is available at `https://github.com/facebookresearch/deit`.
- The Swin transformer [14] is a general-purpose transformer backbone that creates hierarchical feature maps with linear computing complexity concerning image size. Swin transformer creates a hierarchical representation by fusing neighboring patches (i.e., shifting windowing) in deeper transformer layers, starting with small patches. The shifted windowing approach enhances efficiency by limiting self-attention calculation to non-overlapping local windows while enabling cross-window connectivity. The hierarchical architecture of the Swin transformer facilitates the modeling of different scales. The implementation of the Swin transformer is available at `https://github.com/microsoft/Swin-Transformer`.
- XCiT (cross-covariance image transformer) [15] coalesces convolutional architecture scalability with classical transformer accuracy. It comprises a transposed variant of self-attention that interacts using the cross-covariance matrix between keys and queries rather than tokens. The resulting cross-covariance attention is linear in terms of token complexity and can swiftly analyze high-resolution images. Regardless of the number of tokens, XCiT attends to a predetermined number of channels. As a result, XCiT is far more resistant to variations in image resolution during testing and hence better suited to processing variable-size images The implementation of XCiT is available at `https://github.com/facebookresearch/xcit`.
- VisFormer (vision-friendly transformer) [16] improves visual identification by switching from a transformer-based model to a convolution-based one. VisFormer uses a gradual transition technique to bridge the gap between transformer-based and convolution-based models, revealing the features of the designs in both. It dissected the gap between these models and devised an eight-step transition process to connect DeiT-S and ResNet-50. The implementation of VisFormer is available at `https://github.com/danczs/Visformer`.

2.5. Implementation details

We resized the original BUS image resolution to the 224×224 pixels and calculated the mean and standard deviation to normalize the dataset. An Adam optimizer was used with an initial learning rate of 0.0001 and selected the default value of β_1 and β_2 to 0.9 and 0.999, respectively. All the transformer-based networks were trained at 40 epochs with the mini-batch of four samples and saved the best checkpoint of highest classification accuracy on validation to evaluate the performance on the independent test set. We used the cross-entropy loss function to minimize the error during network training. It should be noted that all the networks utilized the same hyperparameter setting to train and evaluate the classification performance. We divided the dataset into the three subsets of training,

validation, and test with ratios of 70%, 10%, and 20%, respectively. We trained and evaluated all the transformer-based models on PyTorch with NVIDIA GeForce GTX 1070Ti GPU of 8GB RAM.

2.6. Evaluation metrics

In this study, the performance of the proposed approach has been assessed using different evaluation metrics, namely accuracy, precision, recall, and F1-score. These metrics can be defined as follows:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \tag{1}$$

$$Precision = TP/(TP+FP) \tag{2}$$

$$Recall = TP/(TP+FN) \tag{3}$$

$$F1-score = TP/(TP+0.5(FP+FN)) \tag{4}$$

where TP represents the number of malignant cases correctly classified as malignant; TN represents the number of benign cases correctly classified as benign; FP represents the number of benign cases incorrectly classified as malignant, and FN represents the number of malignant cases incorrectly classified as benign.

3. Experimental Results and Discussion

Table 1 presents the breast ultrasound classification results of the deep radiomics for the five state-of-the-art transformer-based deep learning methods. It includes the BEiT [12], CaiT [13], Swin [14], XCiT [15], and VisFormer [16]. Figure 3, and 4 demonstrate the confusion matrix of each methods on both the test and validation sets. The test set contains a total of 22 and 11 samples for benign (0) and malignant (1) classes, respectively. In the test set, Swin and XCiT showed comparable performance and achieved an accuracy of 87.88%. The Swin transformer offers the benefits of a small BUS patch and exponentially increases its size by fusing to maintain scale-invariant properties. This mechanism helps capture lesions of various sizes in BUS images and provides robust feature representation to distinguish between benign and malignant tissue patterns. XCiT, on the other hand, provides an efficient self-attention mechanism for BUS features, acting through functional channels instead of tokens, improving classification performance. Both the Swin and XCiT models achieved a precision rate, recall, and F1 score of over 85%. However, due to the increased complexity (i.e., 300M parameter), BEiT has achieved an accuracy of 66.66% less accurately than existing methods. It failed to learn the pattern of benign and malignant classes that require more train samples to effectively achieve better results. CaiT provided the second-highest result, 6% lower than the Swin and XCiT methods. In addition, VisFormer has reached an accuracy of 75.76%.

We calculated each method's computational complexity by measuring the trainable parameters. As one can see, the BEiT method achieved the highest number of parameters (300 million) than existing methods that require additional breast lesion ultrasound samples to improve the classification performance. In turn, Swin, XCiT, VisFormer, and CaiT

Table 1. State-of-the-art Transformer based models comparison on BUS dataset. The best results are highlighted in bold.

Methods	Evaluation Metrics				
	Accuracy	Precision	Recall	F1-Score	Parameters (M)
BEiT-Radiomics	66.66	100.00	66.66	80.00	300
CaiT-Radiomics	81.82	81.82	90.00	85.71	46.5
XCiT-Radiomics	87.88	86.36	95.00	90.48	188.16
VisFormer-Radiomics	75.76	95.46	75.00	84.00	39.45
Swin-Radiomics	87.88	90.91	90.91	90.91	195

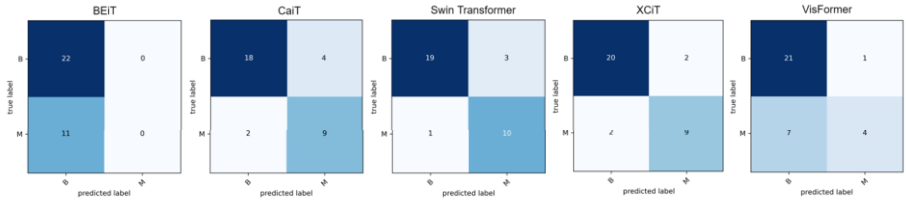


Figure 3. Confusion matrix on test set.

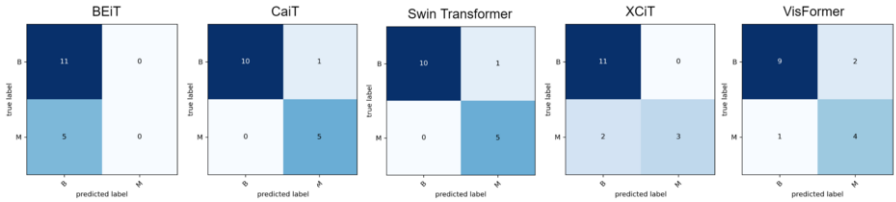
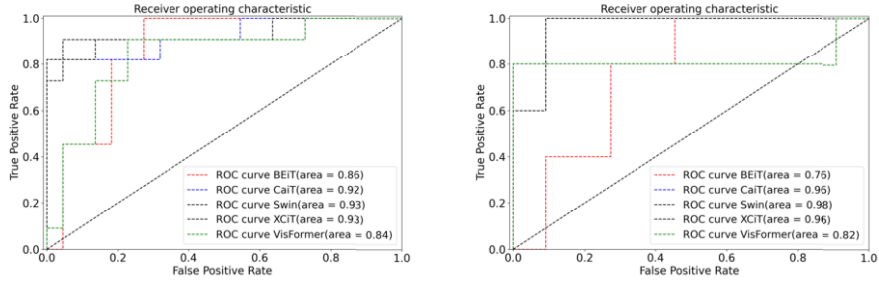


Figure 4. Confusion matrix on val set.

utilized the 194M, 188M, 39M, and 46M trainable parameters respectively. Conclusively, we have found that Swin transformer efficiently extract the radiomics features from BUS ultrasound and achieved state-of-the-art results compared to recently published works.



a (Testing) b (Validation)

Figure 5. ROC curves and AUC values of the radiomics of test, and validation.

Figure 5 shows the receiver operating characteristics (ROC) curves and AUC values of all the examined models. We computed the ROC curve for both evaluated testing and

validation sets. On the validation samples, the Swin transformer has achieved the highest AUC score of 0.98%. However, CaiT and XCiT obtained identical results of 0.96%. The BEiT has only yielded the AUC value of 0.76% which was the lowest of all the transformer-based compared methods. However, on the testing set, the Swin transformer and XCiT have obtained the equal highest AUC score of 0.93% than other existing deep models. At the same time, CaiT obtained the AUC values of 0.92%. The VisFormer yielded the lowest AUC score of 0.84% compared to other methods, while BEiT only attained the 0.86%.

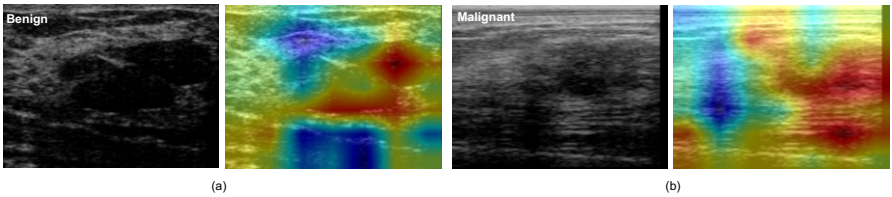


Figure 6. Illustration of Gradcam visualization generated with Swin transformer for two examples of breast ultrasound images. The higher the intensity of the red color, the more attention the model pays to the area of interest.

Figure 6 shows the Gradcam visualization of the best results obtained by the Swin transformer on an ultrasound image of the breast tissues. The core idea was to investigate how the model filters capture the targeted lesion region underlying several noises presented in ultrasound. We found that the model precisely captured the hypoechoic lesion areas by paying more attention to the dense structure of the model. The two different examples presented from benign and malignant classes have different textural patterns and imaging characteristics. For the benign cases, the lesion presented in ultrasound has increased neighboring artifacts that are surrounded by shadows and speckle noises. Due to the great feature representation of the Swin transformer, it is noticeable that it can efficiently highlight the relevant lesion features (red) and ignores the noisy background regions (blue). However, it shows similar characteristics for malignant samples were correctly identified the breast lesion pixels and focused less on neighboring artifacts.

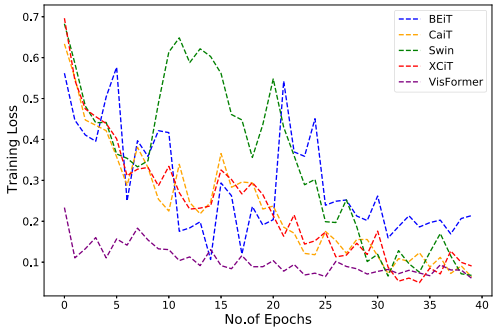


Figure 7. Convergence of each transformer-based methods.

We also investigated the convergence of individually trained deep models such as BEiT,

CaiT, Swin, XCiT and VisFormer. Figure 7 shows the loss convergence of all five existing methods. We observed that all the methods has trained well and training error were minimized at epoch 40.

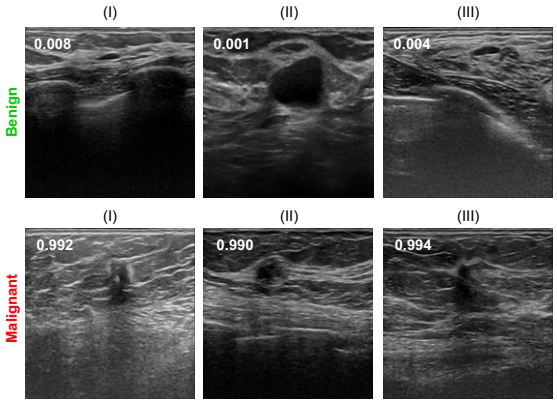


Figure 8. Illustration of malignancy scores (benign and malignant classes) achieved by the highest scores achieved by Swin Transformer. Three examples from each class, score are pasted on the image. High scores above 0.5 correspond to malignant otherwise benign.

Figure 8 shows the malignancy score examples from the test set obtained through the Swin transformer network. From the visual inspection, both benign and malignant class samples show distinct textural patterns. Malignant tumors have an ambiguous boundary, while benign-class tumors contain smooth structures with hypoechoic regions. The scores above 0.5 are malignant. The three malignant examples achieved very high scores of more than 0.99. In turn, the benign class samples obtained the lowest malignancy score. With efficient feature representation, the Swin transformer precisely classified the presented six samples into their classes.

4. Conclusions

A deep vision transformer-based strategy for predicting breast cancer malignancy scores from ultrasound pictures was proposed in this research. BEiT, CaiT, Swin, XCiT, and VisFormer are among the deep vision transformers that have been adopted and trained to extract robust radiomics for categorizing benign and malignant breast cancers in ultrasound pictures. For each input ultrasound image, the highest-performing model is used to forecast the malignancy score. The experimental results demonstrated that the Swin-based radiomics yielded the best classification results. Both the Swin and XCiT models achieved a precision rate, recall, and F1-score of over 90%. The future work will include analyzing other deep learning methods and investigating the use of different multi-model aggregation techniques for enhancing the classification results.

Acknowledgement

The Spanish Government partly supported this research through Project PID2019-105789RB-I00.

References

[1] Carol H Lee, D David Dershaw, Daniel Kopans, Phil Evans, Barbara Monsees, Debra Monticciolo, R James Brenner, Lawrence Bassett, Wendie Berg, Stephen Feig, et al. Breast cancer screening with imaging: recommendations from the society of breast imaging and the acr on the use of mammography, breast mri, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. *Journal of the American college of radiology*, 7(1):18–27, 2010.

[2] Vivek Kumar Singh, Mohamed Abdel-Nasser, Farhan Akram, Hatem A Rashwan, Md Mostafa Kamal Sarker, Nidhi Pandey, Santiago Romani, and Domenec Puig. Breast tumor segmentation in ultrasound images using contextual-information-aware deep adversarial learning framework. *Expert Systems with Applications*, 162:113870, 2020.

[3] Mohamed Abdel-Nasser, Jaime Melendez, Antonio Moreno, and Domenec Puig. The impact of pixel resolution, integration scale, preprocessing, and feature normalization on texture analysis for mass classification in mammograms. *International Journal of Optics*, 2016.

[4] M Abdel-Nasser, A Moreno, and D Puig. Temporal mammogram image registration using optimized curvilinear coordinates. *Computer Methods and Programs in Biomedicine*, 127:1–14, 2016.

[5] Wei-Chung Shia and Dar-Ren Chen. Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine. *Computerized Medical Imaging and Graphics*, 87:101829, 2021.

[6] Arnab K Mishra, Pinki Roy, Sivaji Bandyopadhyay, and Sujit K Das. Breast ultrasound tumour classification: A machine learning—radiomics based approach. *Expert Systems*, 38(7):e12713, 2021.

[7] Zhemin Zhuang, Zengbiao Yang, Alex Noel Joseph Raj, Chuliang Wei, Pengcheng Jin, and Shuxin Zhuang. Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. *Computer methods and programs in biomedicine*, 208:106221, 2021.

[8] Wenju Cui, Yunsong Peng, Gang Yuan, Weiwei Cao, Yuzhu Cao, Zhengda Lu, Xinye Ni, Zhuangzhi Yan, and Jian Zheng. Fmrnet: A fused network of multiple tumoral regions for breast tumor classification with ultrasound images. *Medical Physics*, 49(1):144–157, 2022.

[9] Hailong Yu, Hang Sun, Jing Li, Liying Shi, Nan Bao, Hong Li, Wei Qian, and Shi Zhou. Effective diagnostic model construction based on discriminative breast ultrasound image regions using deep feature extraction. *Medical Physics*, 48(6):2920–2928, 2021.

[10] Moi Hoon Yap, Gerard Pons, Joan Martí, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K Davison, and Robert Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4):1218–1226, 2017.

[11] Mohamed Abdel-Nasser, Domenec Puig, Antonio Moreno, Adel Saleh, Joan Marti, Luis Martin, and Anna Magarolas. Breast tissue characterization in x-ray and ultrasound images using fuzzy local directional patterns and support vector machines. In *VISAPP (1)*, pages 387–394, 2015.

[12] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[13] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[15] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34, 2021.

[16] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 589–598, 2021.