International Journal of Multiphysics
Volume 18, No. 2, 2024
ISSN: 1750-9548

Breast Ultrasound Image BI-RADS Classification Based on Vision Transformer

Yanbo Wei[1], Junbo Ye[2], Xiaofeng Li[3], Yuanyuan Zhao[4], Yanwei Wang[2*]

[1] School of Intelligent Engineering, Harbin Institute of Petroleum, Harbin 150027, China

[2] School of Mechanical Engineer, Heilongjiang University of Science & Technology, Harbin 150022, China

[3] Department of Information Engineering, Heilongjiang International University, Harbin 150025, China

[4] Heilongjiang Provincial Hospital, Harbin 150001, China

*Corresponding author.

Abstract

The most common malignancy among women is breast cancer. Medical ultrasound images are a common tool for detecting breast cancer. In medical ultrasound image classification, Convolutional neural networks(CNNs) have demonstrated great success. However, in most studies on convolutional neural networks categorizes breast tumors into benign and malignant types. Additionally, as convolutional neural networks have a limited receptive field, they are unable to acquire global information. In order to resolve this issue, we explored the feasibility of using Vision Transformer (ViT) in breast ultrasound image BI-RADS classification tasks through transfer learning. We collected publicly available breast ultrasound datasets and enhanced the quality of ultrasound images using the CLAHE algorithm. Through a transfer learning strategy, we trained the ViT model. Using an independent test set, we compared the classification results of ViT with CNNs serving as the baseline model. Breast cancer were categorized based on the BI-RADS criteria, and the results were evaluated using precision, accuracy, and F1 score. According to the experimental analysis results,ViT's transfer learning model produced 94.57% accuracy, 94.11% precision, and 94.29% F1 scores in the classification of breast ultrasound images, respectively, in the breast ultrasound image's BI-RADS classification. The classification performance of the ViT model outperformed the CNN models, including DenseNet201, Xception, MobileNet, and GoogLeNet. The study showed that ViT can be effectively utilized in classifying breast ultrasound images according to the BI-RADS system. The ViT model performed well in breast cancer classification and showed potential as an alternative to CNN.

Keywords: Ultrasound Imaging, Breast cancer, Vision Transformer, Convolutional Neural Network (CNN), Image Classification.

1. Introduction

Women's health and well-being are seriously threatened by breast cancer, which has overtaken lung cancer as the most common cancer in the world[1]. The safe, fast, and accurate diagnosis of breast diseases has received increasing attention in recent years. Currently, conventional screening methods for breast cancer include ultrasound examination, X-ray mammography, and MRI examination, among which ultrasound detection has become the preferred screening method as a result of its noninvasive, painless, low price, and highly accurate characteristics. The ultrasound diagnosis of breast nodules mainly relies on the physician's operating methods and clinical experience, and different levels of experience among ultrasound physicians may lead to different diagnostic results for the same ultrasound image. To standardize the terminology used to describe and evaluate breast tumor characteristics, the American College of Radiology[2] created the BI-RADS to provide detailed grading for malignant breast tumors, so diagnosis of breast disease can be improved significantly. The BI-RADS classification system divides breast lesions into 7 levels based on the nature of the lesion. Levels 0-2 indicate the absence of any malignant lesions, while the malignant probability of level 3 lesions ranges from 0 to 2%. Level 4 lesions are further divided into three subcategories: 4a-4c, with a malignant probability ranging from 2% -10%, 10% -50%, and 50% -95%, respectively. Level 5 lesions indicate a malignant probability of ≥95%, while level 6 lesions are basically diagnosed as malignant. However, the final diagnosis of breast nodules still mainly relies on the pathological results of nodal biopsy, which may cause significant psychological and economic pressure for some patients with benign nodules. Therefore, studying how to classify imaging information based on the BI-RADS classification system is very valuable and necessary. In recent years, deep learning has been a hot topic in image processing research, and the processing of medical images by CNNs has dominated the field from 2012 to 2020. CNNs comprises pooling layers, multiple convolutional layers, and fully connected layers at the bottleneck, and extract shallow and deep image features from shallow to deep depending on the depth of the network. For medical image classification tasks, ResNet[3], Inception-v3[4], EfficientNet[5], and other neural network models are commonly used, and the usual approach is to use pre-trained weights in large datasets like ImageNet for transfer learning. Even though these models have got hold of good results, their inherent inductive bias attributes limit further improvement in the performance of the model by only focusing on local features of the image. To address this issue, attention mechanisms have been introduced into CNN models, such as Axial-DeepLab, which expands the scope of the attention mechanism and reduces computational complexity by applying axial attention separately on height and width. EAC-net proposes an efficient channel attention module that effectively balances the contradiction between model performance and complexity. However, these studies cannot fundamentally overcome the inherent limitations of CNNs. Transformer's application to natural language processing was first proposed by Vaswani et al.[6], and has become the foundation of massive language models, for instance GPT. The modeling ability of long-distance association and the attention to global features of input information have enabled Transformer to achieve great success in the natural language processing field. Research staff have been attempting to apply Transformer to computer vision since 2020. The ViT structure was first proposed by Google, which applied Transformer to computer vision while maintaining its original structure as much as possible, and achieving experimental results comparable to the most advanced CNN models. Since then, many excellent visual Transformers, such as Facebook's DeiT and Microsoft Asia Research Institute's Swim Transformer[7], have emerged in computer vision field. Soon, there are a wide variety of specific tasks that can be solved using these models. This article will explore how to apply ViT Transformer to the breast ultrasound tumor image classification, analyze the experimental results, and compare them with classic CNN classification algorithms and the latest image classification algorithms.

2. Related Work

Recent years have seen CNNs become the standard for classifying medical images, on large-scale datasets, researchers using pre-trained CNN models and transferring their weights to breast ultrasound image classification models. Hatamizadeh et al.[8] focused on the application of transfer learning, VGG-16 and InceptionV3 for detecting lesions in breast ultrasound images. Al-Dhabyani et al.[9] applied VGG16, Inception, NASNet and ResNet to breast ultrasound image classification, with results showing that NASNet outperformed other CNNs. Although CNNs has achieved success in image processing, their limited local receptive fields restrict their performance, while ViT models can extract global information well. When trained on large-scale datasets, the Vision Transformer network outperforms CNN models in image classification. In various visual tasks, many researchers have studied the model of Vision Transformer because of its excellent performance. Carion et al. presented a new architecture in the field of object detection, using Transformer encoder-decoder and a set-based global loss algorithm, on the challenging COCO dataset, demonstrated results comparable to those achieved with the dominant R-CNN method. Karimi et al. proposed a new segmentation model specifically for 3D medical image for image segmentation tasks, achieving better segmentation accuracy than existing CNNs on multiple datasets. Hatamizadeh et al.[8] put forward a original U-Net Transformer architecture by combining U-Net and Transformer. U-Net Transformer combines U-Net advantage in terms of image segmentation and Transformer's ability to extract global information, encodes and decodes images using Transformer and CNN-based architectures. It consistently performs well in brain tumor and spleen segmentation tasks. In image classification tasks, Gheflati et al.[10]applied the ViT model to breast ultrasound images, achieving equivalent or even better results than the most advanced CNN models by fine-tuning the pre-trained weights.

We classified breast ultrasound images using ViT model. However, in contrast to previous work, our classification results do not just distinguish between benign, malignant, and normal cases, but rather correspond to the six categories in the BI-RADS classification standard.

3. Methodology

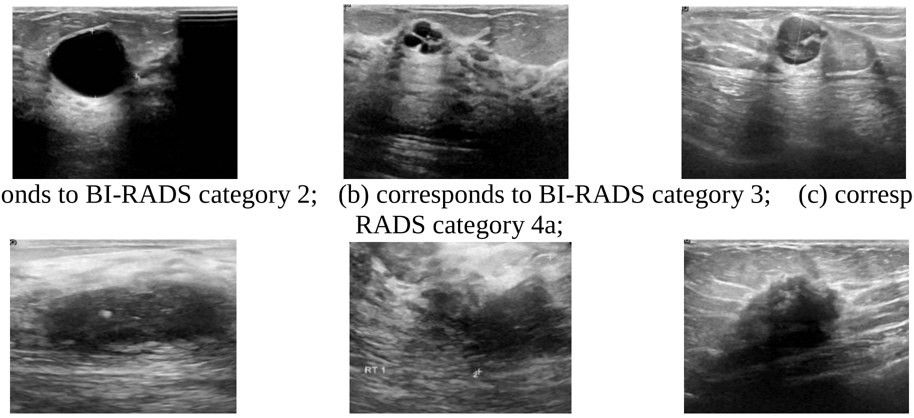
We referred to the ViT model in Ref. 11, loaded its pre-trained weights and followed the training method in reference to fine-tune the model on the breast ultrasound image dataset, and finally obtained a ViT model suitable for BI-RADS breast ultrasound grading.

3.1 Data

The datasets used in this study include four publicly available datasets, three of which are combined into a training set and a test set in an 8:2 ratio, while the remaining dataset is used for independent validation. The first dataset is called BUSI[11], collected using LOGIQ E9 and LOGIQ E9 Agile US scanners at the Baheya Hospital, Cairo, Egypt. There are 780 breast ultrasound images in BUSI from 600 women, averaging 500 by 500 pixels. A total of 210 benign tumors, 437 malignant tumors, and 133 normal images are included in the dataset. The second dataset is named UDIAT[12], in the UDIAT dataset, there are 163 images, 110 of which are benign and 53 of which are malignant. This data were obtained from the UDIAT Diagnostic Centre of the Parc Tauli Corporation, Sabadell, Spain, using Siemens ACUSON scanner. The third is called OASBUD[13], collected from patients at the Maria Sklodowska-Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland. OASBUD contains 100 images from the United States, including 48 benign and 52 malignant breast masses. The last dataset is S1 Data[14], which is used for independent validation analysis. Based on these publicly available datasets, we collaborated with expert sonographers to remove irrelevant data and perform breast ultrasound images into six categories according to the BI-RADS classification standard. Subsequently, we augmented the breast ultrasound data by applying techniques such as rotation and mirroring. The dataset statistics for each category are shown in Table 1. And Figure 1 shows examples of breast ultrasound images classified by BI-RADS category.

Table 1 Number of cases and images for different BI-RADS categories in the dataset.

Types	2	3	4a	4b	4c	5
Quantity	1100	1300	730	500	420	310



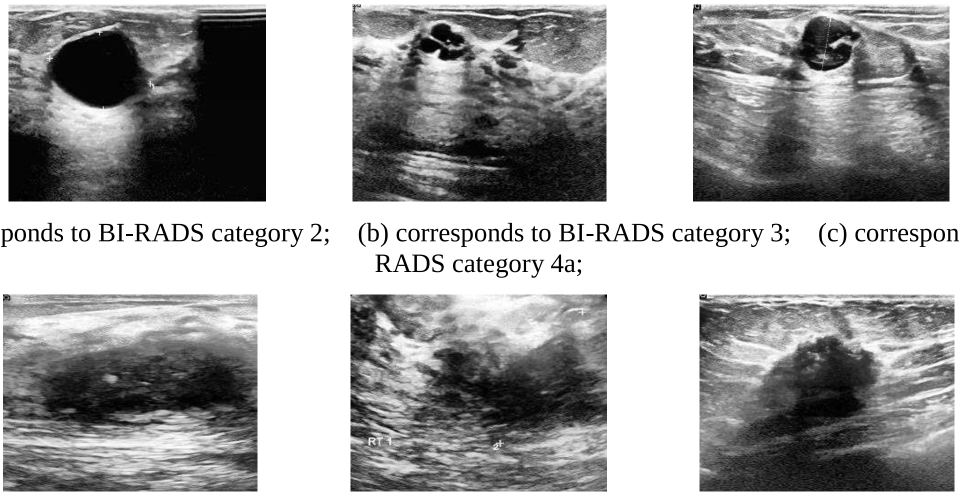
(a) corresponds to BI-RADS category 2; (b) corresponds to BI-RADS category 3; (c) corresponds to BI-RADS category 4a;

(d) corresponds to BI-RADS category 4b; (e) corresponds to BI-RADS category 4c; (f) corresponds to BI-RADS category 5.

Figure 1 Example of breast US images with six different BI-RADS classifications

3.2 Data Enhancement

To enhance the features of breast ultrasound images, we used data augmentation techniques. Adaptive Histogram Equalization (AHE) is a commonly used image enhancement method. The AHE algorithm divides the input image into several sub-blocks and applies histogram equalization to each sub-block individually, which corresponds to different parts of the image. This process redistributes the intensity values of the image. However, while enhancing the image, the AHE algorithm also amplifies the image noise. Contrast Limited Adaptive Histogram Equalization (CLAHE) is an upgrade version of AHE that mitigates noise amplification. CLAHE adds a contrast limitation by clipping the local histogram of the image using a predetermined threshold. This reduces the extent of contrast enhancement and minimizes unwanted noise amplification, thereby alleviating excessive enhancement in certain areas of the image. Therefore, we employed the CLAHE algorithm for data augmentation, and the processed images are shown in figure 2.



(a)corresponds to BI-RADS category 2; (b) corresponds to BI-RADS category 3; (c) corresponds to BI-RADS category 4a;

(d) corresponds to BI-RADS category 4b; (e) corresponds to BI-RADS category 4c; (f) corresponds to BI-RADS category 5.

Figure 2 Example of enhanced breast ultrasound images by CLAHE

3.3 Experimental environment and evaluation metrics

In this experimental study, using Pycharm and Python as the IDE and programming language respectively. The PyTorch deep learning framework was used for the experiments, which were carried out on a computer equipped with an Intel(R) Xeon(R) Gold 6226R 2.90GHz CPU, 64GB DDR4 RAM, and an NVIDIA GeForce RTX3080 GPU with 10GB of RAM. The operating system used was 64-bit Windows. Medical image classification research typically uses performance metrics to evaluate classification results, including accuracy (A), precision (P), F1 score. The calculation formulas are provided below:

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{1}$$

$$P = \frac{T_p}{T_p + F_p} \tag{2}$$

$$F_1 = 2 \times \frac{P \times \frac{T_p}{T_p + F_n}}{P + \frac{T_p}{T_p + F_n}} \tag{3}$$

In the formula, true positives are represented by Tp, true negatives by Tn, false positives are represented by Fp, and false negatives are represented by Fn. As this experiment involves a six-classification task with imbalanced samples, weighted averaging is adopted for the consideration of scores, precision,and accuracy. The calculation formulas for the three performance metrics after weighted averaging are provided below:

$$A = \sum_{k}^{m}\left(A_k \times \frac{N_k}{P}\right) \tag{4}$$

$$P = \sum_{k}^{m}\left(P_k \times \frac{N_k}{P}\right) \tag{5}$$

$$F_1 = \sum_{k}^{m}\left(F_{1k} \times \frac{N_k}{P}\right) \tag{6}$$

Where N_k represents sample size in each class, and P represents the total sample size.

3.4 ViT Architecture

The ViT model architecture used in this study is similar to the ViT model mentioned in Ref. 11, with the difference being the conversion of the MLP head to a linear classifier. Figure 3 shows the network framework of the VIT model, the input image is partitioned into i partstches, the 1D patch embedding sequence is subsequently input into the transformer encoder, wherein self-attention modules are leveraged to compute the relation-based weighted sum of the outputs originating from each hidden layer. This approach facilitates the Transformers' learn capacity to the global dependencies inherent within the input image. The core idea behind the ViT model is to divide the image into smaller image patches and capture their relationships through the Transformer encoder. This attention-based approach enables the model to simultaneously consider both local features and global contextual information of the image, enhancing its understanding of the image content. In this model, the self-attention mechanism plays a critical role by aggregating information through computing the weighted sum of related image patches. By automatically learning the correlations between image patches, the model can better capture the semantic connections among different regions of the image.

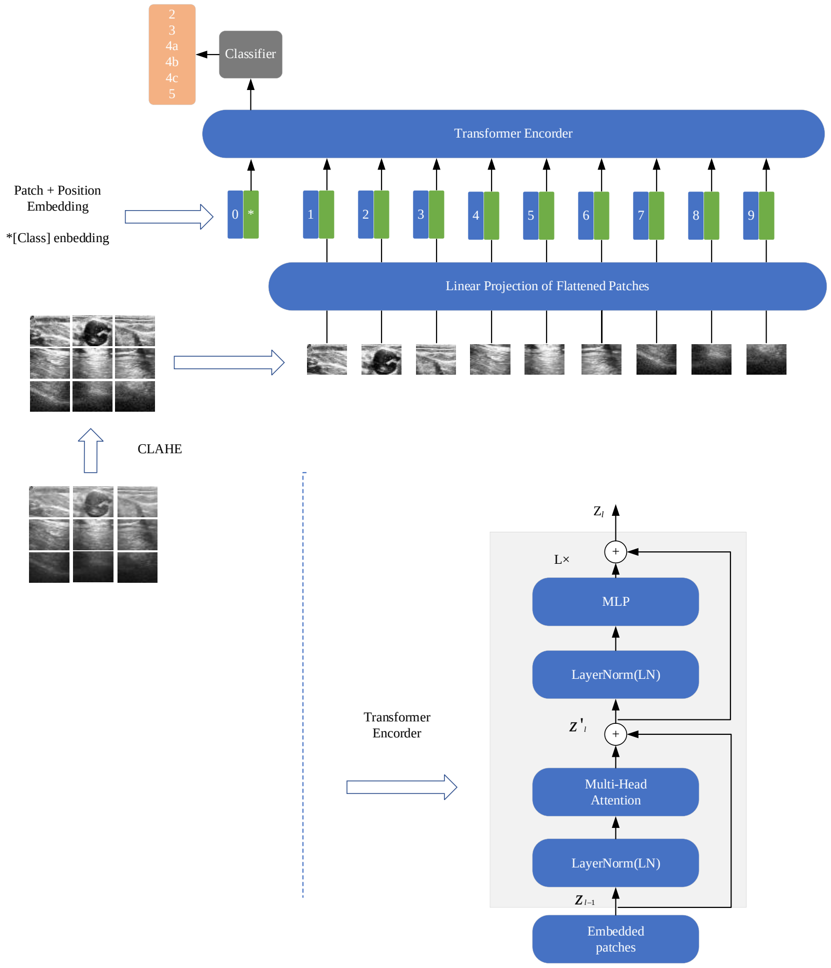


Figure 3 Framework for Classifying Breast Ultrasound Images on the VIT Network

3.5 Fine-tuning details

There are currently many ViT models of different scales and pre-trained weights on large datasets, but the question of how to choose pre-trained weights for transfer learning remains a problem. There are many options concerning model selection and configurations for fine-tuning weights according to the new dataset. In study, an 8:2 ratio was used to split the dataset into training and testing sets. To ensure the fairness of the experiment, all models were trained on this dataset. The final evaluation metric utilized the mean of 5-fold cross-entropy validation. For ViT models, We utilized pre-trained weights based on the ImageNet dataset with different patch sizes, used the cross-entropy loss function to assess the difference between the model's true values and predicted values, and used the stochastic gradient descent (SGD) algorithm to reduce the loss function and update the model parameters. Utilizing the SGD optimizer with a momentum of 0.9, the training process involved 150 epochs, a batch size of 64, and an initial learning rate set at 0.0001. For CNN models, a classifier with a softmax activation function was used to alter the output layer. The optimizer used was Adam, and the model was trained for 150 epochs.

4. Results

After testing categories 2, 3, 4a, 4b, 4c, and 5, we found that CNNs and ViT models are effective for BI-RADS classification. As far as we know, tahere is a lack of research on the classification of BI-RADS 2, 3, 4a, 4b, 4c, and 5 using the ViT approach. Therefore, in order to verify the validity of our proposed method, we compared the fine-tuned ViT model in this study with several models that are currently recognized to have good classification performance. These methods include DenseNet201[15], Xception[16], MobileNet[17] and GoogLeNet[18]. The experimental results are presented in Table 2. In the ViT model testing process, we employed a transfer learning strategy by incorporating pre-trained weights based on the ImageNet dataset.

Table 2 Test accuracy, precision and F1 score of different algorithms (%)

Algorithm	A	P	F1
DenseNet201	93.05	93.11	93.25
Xception	91.81	92.02	92.09
MobileNet	91.90	91.84	91.93
GoogLeNet	94.11	94.12	94.22
ViT-B/32	93.21	93.47	92.86
VIT-B/16	94.57	94.29	94.15

The data results in Table 2 show that the ViT-B/16 model achieved better test results, reaching 94.57% accuracy, 94.29% accuracy, and 94.15% F1 score, compared to the other classification algorithms. Moreover, it can be observed that, for breast ultrasound images, the results of the ViT model with a patch size of 16 are better than those with a patch size of 32. This is because a smaller patch size can provide higher feature resolution. In breast ultrasound images, some important details may be encoded within smaller regions, thus using smaller patches can better capture these crucial features. If there are more patches, the size of each patch will decrease, potentially resulting in the loss of important details. In summary, this experimental result demonstrates the feasibility and superiority of the application of ViT model in medical breast ultrasound image classification.

To further analyze the generalization of the model, we tested the classification model using the public dataset S1 Data Since this datasets only have benign and malignant labels, we conducted a binary classification experiment for breast ultrasound image benignity. On account of the small amount of dataset, we also preprocessed the data, so we augmented the data by rotating 90 degrees and mirroring, and the raw data and the augmented data are shown in Table 3.

Table 3 augmentation results of S1 Data

Dataset	benign	malignant	augmentation		sum
			benign	malignant	
S1_Data	200	248	400	496	892

Similarly, DenseNet201, Xception, MobileNet and GoogLeNet were used as comparison experiments, and Table 4 shows the accuracy, precision, and F1 score of the experiments.

Table 4 Validation accuracy, precision and F1 score results of several algorithms (%)

Algorithm	A	P	F1
DenseNet201	93.27	93.86	93.17
Xception	92.11	92.16	92.12
MobileNet	92.43	92.48	92.51
GoogLeNet	94.24	93.71	94.22
VIT-B/16	94.57	94.11	94.29

The results in Table 4 show that VIT-B/16 model’s accuracy, precision, and F1 score on S1_Data are 94.57%, 94.11%, and 94.29%, respectively, Compared to convolutional neural network (CNN)-based methods such as DenseNet201, Xception, MobileNet, and GoogLeNet, VIT-B/16 achieved superior results. VIT-B/16 outperformed GoogLeNet by 0.33% in terms of accuracy, demonstrating comparable performance. Moreover, it exhibited an improvement of 1.12% in F1 score, thus showing a certain advantage in F1 score. Furthermore, VIT-B/16 demonstrated a precision improvement of 0.25% compared to DenseNet201.These findings demonstrate the generalizability of the fine-tuned ViT model on breast ultrasound images and its suitability for clinical applications.

5. Conclusion

In this research, we applied the Transformer-based architecture to the BI-RADS classification task of breast ultrasound images. Considering that the training of ViT models requires a good deal of dataset, we adopted the transfer learning strategy of pre-training the ViT model to further adapt it to the BI-RADS classification task of breast ultrasound images. We compared the classification performance of the fine-tuned ViT model with CNNs model, demonstrating the enormous potential of the pre-trained ViT model based on the transfer learning strategy for breast ultrasound image classification. This study proves the possibility of replacing the CNN model with the Transformer architecture in breast ultrasound image analysis tasks, providing a new research direction for future breast ultrasound image analysis.

This study also has some limitations: This paper’s original intention is to use the pre-trained ViT model for BI-RADS classification of breast ultrasound images, but the clinical data available for collection is still limited. To verify the generalization performance of the fine-tuned ViT model, we used a dataset with benign and malignant classification labels, and did not use data with BI-RADS grading labels for validation. In future studies, we will continue to optimize the algorithm to improve classification accuracy and train on more diverse datasets.

Acknowledgment

This work was supported by Natural Science Foundation of Heilongjiang Province of China under grant number LH2021F039.

References

[1] J. Ferlay, M. Colombet, I. Soerjomataram, D. M. Parkin, M. Piñeros et al., "Cancer statistics for the year 2020: An overview," International Journal of Cancer 149(4), 778-789 (2021).

[2] L. Liberman, and J. H. Menell, "Breast imaging reporting and data system (BI-RADS)," Radiologic Clinics 40(3), 409-430 (2002).

[3] K. He, X. Zhang, S. Renet, J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition 770-778 (2016).

[4] C. Szegedy, V. Vincent, I. Sergey, S. Jonathon, W. Zbigniew et al., "Rethinking the inception architecture for computer vision," Proceedings of the IEEE conference on computer vision and pattern recognition 2818-2826 (2016).

[5] M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," International conference on machine learning 6105-6114 (2019).

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones et al., "Attention is all you need," Advances in neural information processing systems 30((2017).

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei et al., "Swin transformer: Hierarchical vision transformer using shifted windows," Proceedings of the IEEE/CVF international conference on computer vision 10012-10022 (2021).

[8] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko et al., "Unetr: Transformers for 3d medical image segmentation," Proceedings of the IEEE/CVF winter conference on applications of computer vision 574-584 (2022).

[9] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy et al., "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images," International Journal of Advanced Computer Science and Applications 10(5), 1-11 (2019).

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

[11] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy et al., "Dataset of breast ultrasound images," Data in brief 28(104863 (2020).

[12] Yap M H, Pons G, Marti J, Ganau S, Sentis M et al. “Automated breast ultrasound lesions detection using convolutional neural networks,” IEEE journal of biomedical and health informatics 22(4), 1218-1226 (2017).

[13] Piotrzkowska‐Wróblewska, Hanna, Dobruch-Sobczak Katarzyna, Byra Michał, Nowicki Andrzej et al. "Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions." Medical physics 44(11), 6105-6109 (2017).

[14] S. M. Badawy, A. E. AMohamed, A. A. Hefnawy, H. E. Zidan, M. T. GadAllah et al., "Automatic semantic segmentation of breast tumors in ultrasound images based on combining fuzzy logic and deep learning—A feasibility study," PLoS One 16(5), e0 251 899 (2021).

[15] G. Huang, Z. Liu, L. Maaten, K. Q. Weinberger, "Densely connected convolutional networks," Proceedings of the IEEE conference on computer vision and pattern recognition 4700-4708 (2017).

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proceedings of the IEEE conference on computer vision and pattern recognition 1251-1258 (2017).

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:.04861 (2017).

[18] C. Szegedy, W. Liu 0015, Y. Q. Jia, S. Pierre, S. E. Reed et al., "Going deeper with convolutions," Proceedings of the IEEE conference on computer vision and pattern recognition 1-9 (2015).