Multi-Task Breast Ultrasound Image Segmentation and Classification Using Convolutional Neural Network and Transformer

Joanna Loja*, Armando Mendez*, Kuan Huang†
Department of Computer Science and Technology
Kean University
Union, USA
{lojaj, mendearm, khuang}@kean.edu

***Abstract*—Breast ultrasound (BUS) imaging offers a non-invasive and radiation-free method to examine breast tissues. Automated BUS image segmentation and classification can help doctors identify lesions and possible abnormalities early, enabling healthcare professionals to detect breast cancer or other conditions in time for early intervention. In this research, we first conduct a comprehensive performance comparison between transformer networks and convolutional networks; secondly, we propose a novel approach by merging segmentation and classification networks, creating a multi-task network tailored explicitly for BUS image segmentation and classification; thirdly, we thoroughly investigate network performance and refine training parameters to prevent overfitting. Finally, we create a user-friendly GUI demo showing our classification and segmentation results. The results demonstrate that the ResNet-50 Multi-Task model exhibits the best overall performance for both segmentation and classification tasks.**

***Index Terms*—breast ultrasound imaging, deep learning, image segmentation, image classification, multi-task neural network**

I. Introduction

Tumors arise from the aggregation of abdominal cells, leading to the formation of lumps or growths. These tumors can be categorized as either benign or malignant. Benign tumors are non-cancerous, while malignant tumors have the potential to become cancerous [1]. Unfortunately, cancerous tumors rank among the primary causes of death globally [2]. Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer mortality among women worldwide in 2020 [3]. According to the report, 50 percent of women who have experienced cancer had not been detected of it until two years of screenings. With this, the timing of detecting and preventing tumors in patients is crucial as it will give patients enough time to care for the issue before it's too late [4]. Early detection and appropriate treatments can increase survival rates. To identify early signs of tumors, modern machine-learning algorithms can be employed. These algorithms, based on artificial intelligence (AI) techniques, analyze specific data and make predictions or decisions regarding the presence of tumors. These types of algorithms use very complex methods to make a decision that even sometimes the average person cannot see. Thus, we investigate deep neural networks and their applications in breast ultrasound (BUS) image segmentation and classification in this project. We aim to develop and evaluate various tumor classification and segmentation networks. After creating these networks, they will undergo rigorous testing to assess their performance. Each network will be tested, and their respective results will be compared to determine which framework achieves the fastest run time. The outcome of this research will be integrated into a user-friendly GUI demo. This tool will assist doctors in identifying and diagnosing breast tumors in ultrasound images efficiently. Several existing AI-based research methods are conducted for ultrasound image classification and segmentation.

Given the excellent performance of AI-based methods utilized in this research study, it is noteworthy that several previous studies have employed diverse classification methods as well. For instance, Tanaka *et al.* [5] ensembled a VGG-19 and a ResNet-152 for BUS image classification; this method would classify different sections of tumors to determine whether they were benign or malignant. In [6], a transfer learning method was developed for BUS image classification. A classic convolutional neural network extracted features, and then a Support Vector Machine (SVM) classifier was used to classify them into benign or malignant. In [7], Zhuang *et al*. developed a novel BUS image classification method with hand-crafted features and convolutional features. The classifier used in this research was SVM. The previous methods tried to use different network structures, different features, and ensemble learning methods to get high classification accuracy; however, they did not incorporate tumor shapes and locations when classifying tumors. In order to acquire the tumor shapes and locations, BUS image segmentation needs to be conducted.

In addition to employing various classification methods in our study, we have integrated distinct segmentation techniques. Interestingly, other researchers have also conducted previous studies exploring different segmentation methods. For

*Equal contribution.

†Correspondence to Kuan Huang. This work is partially supported by the Students Partnering with Faculty (SpF) 2023 Program of Kean University.

979-8-3503-0965-2/23/$31.00 ©2023 IEEE

instance, Huang *et al.* [8] developed a BUS image segmentation method using a fully convolutional network (FCN) and conditional random fields (CRFs) for post-processing. The input images were pre-processed by wavelet transformation. Another previous study was Amriri *et al.* [9]. In this study, researchers used a two-stage U-net architecture which was categorized into two parts: part one would be used for tumor detection, and part two would be used for tumor segmentation. They later proved in their study that part one of their tumor detection would help give more accurate results for part two of their tumor segmentation. Lastly, Moon *et al.* [10] proposed a deep learning approach for breast cancer detection. Three network structures were used: patch-based LeNet, U-Net, and transfer learning with a pre-trained fully convolutional network (FCN) with AlexNet.

In the computer vision community, many Multi-Task Learning (MTL) studies planned to combine BUS image segmentation and classification tasks in one network so that they can share information once the network starts training to receive more robust results. We understand that these MTL methods can be successfully applied to get positive results in our experiments since many researchers have successfully created models that use classification and segmentation methods. For instance, Zhang *et al.* [11] used an MTL framework that would use different types of mechanisms to give direction to each model, allowing it to focus more on other tumor regions to increase the level of accurate classification result. Another study using a different classification method was by Xu *et al.* [12]. Liao *et al.* [13] used a segmentation network and a feature combination method for breast tumor classification. These researchers incorporated an MTL framework with a module named context-oriented-self-attention. This module would use previous medical knowledge to direct the model to form circumstantial relationships to provide more substantial classification and segmentation results. While the methods employed for BUS image segmentation and classification have succeeded, they also have certain drawbacks. One significant challenge is the need for human intervention during the segmentation process. For instance, Xian *et al.* [14] discussed how radiologists' input can aid in segmenting complex BUS cases. However, applying semi-automatic approaches to larger datasets can increase labor and time costs.

Considering the evaluations of previous BUS image segmentation and classification methods, MTL methods hold promise for breast tumor classification. Including tumor shapes and locations as prior information can yield improved classification outcomes. Traditional convolutional neural networks only consider information within the kernel size, neglecting external details. In this context, the Transformer [15] emerges as an innovative operator capable of incorporating global information. This property makes it a valuable tool for enhancing BUS image analysis. Therefore, the objective of this project can be summarized as follows:

- We construct an MTL framework for simultaneous BUS image segmentation and classification.
- We use ResNet-50 and Tiny Swin Transformer (ST) as the backbones of our framework.
- To assess the effectiveness of the MTL framework and transformer, we perform experiments with six different network configurations for BUS image segmentation and classification. These configurations include ResNet-50 for classification only, U-Net with ResNet-50 for segmentation only, MTL network combining segmentation and classification with ResNet-50, ST-based network for classification only, U-Net with ST for segmentation only, and MTL network combining segmentation and classification with ST as the backbone.
- We develop a GUI demo that allows users to choose a image. The demo then displays the segmentation and classification results for the selected image.

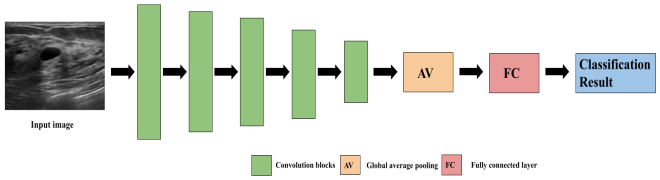

Fig. 1. Image classification network baseline.

II. Data

In this research, we utilize an online public dataset [16] comprising 780 images. Out of these images, we specifically employ 647 for our learning methods, with 210 malignant tumor images and 437 benign tumor images. The labels assigned to these images in our networks are 0 for benign and 1 for malignant. The original images are gray-scale images with dimensions around 400×500, but we resize them to 256×256 before inputting them into the networks. To conduct data analysis and obtain results, we employ Python and Google Colab. The networks are implemented using the PyTorch platform. For the experiments, we randomly select 80% of the images from our dataset for training purposes, while the remaining 20% are reserved for testing. Given the imbalanced number of BUS images in each category, we employ various image augmentation techniques to address this issue. Further details will be provided in the experiment section.

III. Methology

A. Classification Framework

In this research, we utilize the ResNet [17] classification framework as our baseline, which is depicted in Fig. 1. To begin, an input image with dimensions of 256×256 is fed into the system. This image then undergoes processing through five convolutional blocks indicated in green in Fig. 1. Subsequently, a convolutional feature map of size $8 \times 8 \times 2048$ is extracted from the processing. The extracted feature map then undergoes global average pooling. The classification task is accomplished through a fully connected layer that inputs the one-dimensional feature and generates the classification

results. After completing these steps, the classification results are obtained. To train this classification framework, the Softmax function is applied to the network's output, and the cross-entropy loss is computed for training the classification network. When using the ST to classify input images, the five convolutional blocks are replaced by ST blocks in [15].

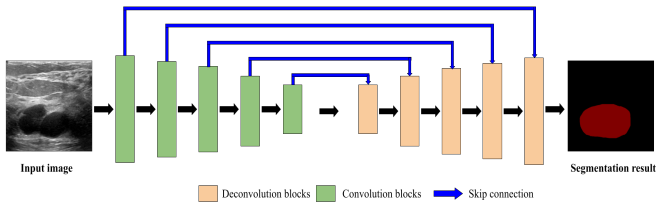

Fig. 2. Image segmentation network baseline (U-Net).

B. Segmentation Framework

The segmentation framework, illustrated in Fig. 2, employs the renowned U-Shape network structure [18] with ResNet-50 as the backbone for its five convolutional blocks. Similar to the classification network, an input image with dimensions 256×256 is initially fed into the system. Subsequently, the image passes through the five convolutional blocks (green labeled blocks), causing a decrease in size at each block. Once the image reaches the last convoluted block, it proceeds through 5 deconvolution blocks (orange labeled blocks), gradually enlarging the convolutional features until the last deconvolution block, which restores the feature map to the original image size. Skip connections (blue arrows) are integrated into the U-Shape network structure, providing a shortcut for low-level feature information from the encoder to the decoder. This mechanism enhances precise object localization during the segmentation process. Following the last convolutional block, a convolutional operation converts the feature map into a segmentation map with dimensions $256 \times 256 \times C$, where C represents the category number. In this research, there are two classes: background and tumor. Similar to the classification network, the segmentation map undergoes processing by the Softmax function, and a cross-entropy loss is employed for training the segmentation network. Notably, when the ST is used for image classification, the five convolutional blocks are replaced by ST blocks as described in [15].

C. Multi-Task Framework

To leverage the tumor location and shape information for improved classification, we propose integrating the classification and segmentation networks into a single Multi-Task Learning (MTL) network capable of simultaneously segmenting and classifying BUS images. The MTL network will share the convolutional blocks and go into two branches after the last convolutional block. Fig. 3 illustrates the framework of our MTL network. Initially, an input image is fed into the network and processed through the five convolutional blocks. Once the image completes the five blocks, it proceeds through two branches: segmentation and classification. First, the image

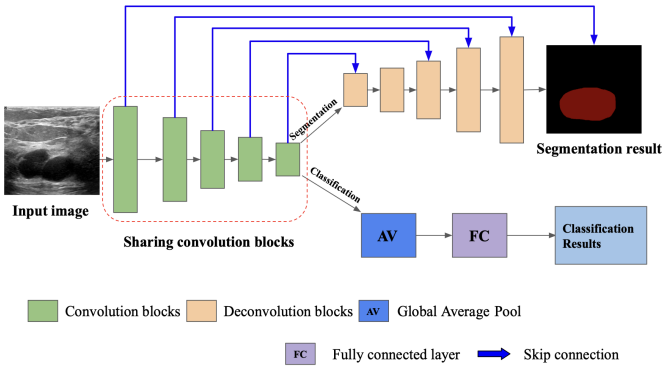

Fig. 3. The proposed Multi-Task framework.

undergoes segmentation, followed by the classification step. As a result, two separate sets of results are obtained from each method. Since the two tasks are combined, classification and segmentation loss functions must be considered. The loss function for the MTL is summarized as:

$$\mathcal{L}_{total} = l_{cls} + l_{seg} \tag{1}$$

where l_{cls} represents the cross entropy loss of classification task and l_{seg} represents the cross entropy loss of segmentation task. This loss function is essential in optimizing the MTL network for joint segmentation and classification tasks, and they enable us to effectively harness the benefits of tumor location and shape information in improving the classification performance. The MTL network that utilizes ResNet-50 as its backbone is referred to as the "ResNet-50 Multi-Task Network." On the other hand, the MTL network with ST as its backbone is named the "ST Multi-Task Network."

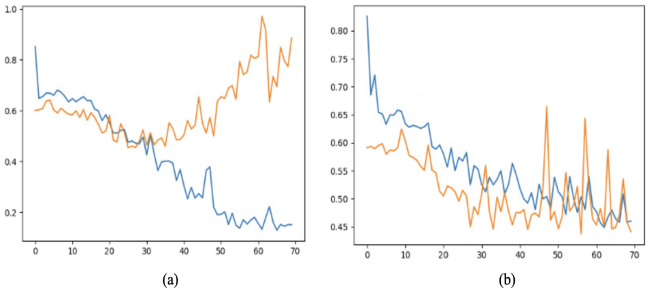

Fig. 4. Training loss curve (blue) and test loss curve (orange) for ST classification network: (a) without avoiding overfitting, (b) with avoiding overfitting.

IV. Experimental Results

A. Experiment Setting and Metrics

The proposed method is implemented using the PyTorch framework, which is a publicly available platform. All experiments are performed on free Google Colab Notebooks, utilizing an NVIDIA T4 GPU for computations. The Adam optimizer is adopted with an initial learning rate of 1e-4 and the momentum parameters β_1 of 0.9, β_2 of 0.99. During the

training of the networks, we encounter the issue of overfitting, as illustrated in Fig. 4 (a). The loss curve starts to increase after 30 epochs for the ST classification network. To mitigate overfitting, we employ several methods: 1) Data augmentation techniques are applied, including horizontal flipping, rotation, and shifting, to increase the diversity of the training dataset. 2) Weight decay of 1e-4 is introduced to regulate the magnitude of the weights during training. 3) The learning rate is reduced by 0.1 after every 20 epochs. After applying those methods, the overfitting is reduced (Fig. 4 (b)). To facilitate stable training, a batch size of 16 is used, and the networks are trained for a total of 70 epochs.

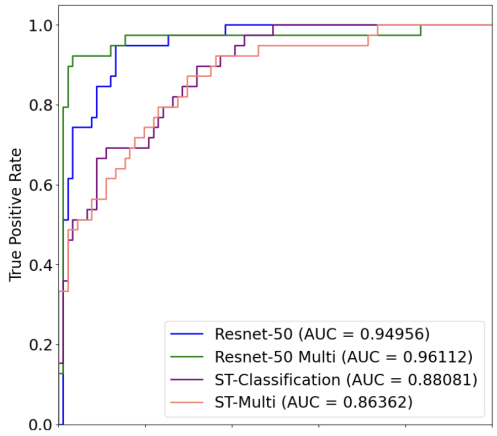

Fig. 5. ROC curves for different classification methods.

We use True Positive Ratio (TPR), False Positive Ratio (FPR), Classification Accuracy (ACC), Precision (PRE), F_1-score, and Receiver Operating Characteristic (ROC) curve to evaluate classification results. We use Intersection over Union (IoU) to evaulate segmentation results.

B. Classification Results

In this subsection, we will compare the classification results of four methods: ResNet-50 classification only, ST classification only, ResNet-50 Multi-Task, and ST Multi-Task networks. These methods' ROC curves and area under the curve (AUC) are presented in Fig. 5. From the figure, it is evident that the most successful method is the ResNet-50 Multi-Task network.

Table I displays five other classification metrics. The fold numbers are the best results. Comparing the second row (ResNet-50 Multi-Task) with the first row (ResNet-50), we observe that the ResNet-50 Multi-Task network outperforms the ResNet-50 in all five metrics. Likewise, comparing the fourth row (ST Multi-Task) with the third row (ST classification), we find that the ST Multi-Task network outperforms the ST classification network in terms of True Positive Rate (TPR), Accuracy (ACC), and F_1 score.

Overall, the ResNet-50 Multi-Task network delivers the best classification results. The experimental conclusion is that the MTL network can achieve superior classification performance, while the Swin Transformer falls short of achieving comparable results to ResNet in this context.

TABLE I
SUMMARY OF CLASSIFICATION RESULTS

Method	TPR	FPR	ACC	PRE	F1 score
ResNet-50	0.8205	0.0659	0.9000	0.8421	0.8311
ResNet-50 Multi-Task	**0.8974**	**0.0329**	**0.9461**	**0.9210**	**0.9090**
ST-Classification	0.5641	0.0879	0.8076	0.7333	0.6376
ST-Multi-Task	0.7435	0.1318	0.8307	0.7073	0.7250

C. Segmentation Results

In addition to the classification results, we compared the segmentation performance using four different methods. Fig. 6 illustrates two examples of original BUS images, their corresponding original labels, and the segmentation results obtained by the four methods. Column (a) displays the original BUS images, column (b) shows the label images, column (c) presents the segmentation results using the U-Net with ResNet-50, column (d) shows the segmentation results using the ResNet-50 Multi-Task method, column (e) demonstrates the segmentation results using the U-Net with ST method, and finally, column (f) exhibits the segmentation results using the ST Multi-Task method. Based on these visual comparisons, it is evident that the U-Net with ResNet-50 method performs the most successfully regarding segmentation.

TABLE II
SUMMARY OF SEGMENTATION RESULTS

Method	Tumor IoU	Background IoU
U-Net with ResNet-50	**0.6388**	**0.9526**
ResNet-50 Multi-Task	0.6104	0.9490
U-Net with ST	0.5280	0.9411
ST-Multi-Task	0.5118	0.9415

Table II provides the evaluated methods' Intersection over Union (IoU) metrics. Comparing the second row (ResNet-50 Multi-Task) with the first row (U-Net with ResNet-50), we observe that the ResNet-50 Multi-Task network slightly reduces the tumor IoU, but the difference is not substantial. Similarly, comparing the fourth row (ST Multi-Task) with the third row (ST with U-Net), we find that the ST Multi-Task method also does not outperform the U-Net with ST regarding segmentation. Based on these results, it is evident that the U-Net with ResNet-50 achieves the best segmentation results among the tested methods. However, the MTL networks can significantly improve the classification results by slightly reducing segmentation performance. In conclusion, the ResNet-50 Multi-Task network achieves the best classification results and comparable segmentation results.

D. GUI Demo

We have created a user-friendly GUI using Python that allows users to utilize our trained ResNet-50 Multi-Task model. Fig. 7 illustrates the functionality of the GUI. Users can select images by clicking on the image selection button. Once an image is chosen, the GUI displays it along with its segmentation result and classification outcome. The classification result includes the predicted class and the probability score for each class. This interactive GUI provides a convenient and intuitive

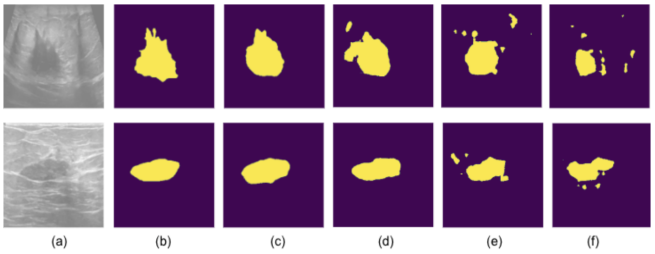

Fig. 6. Segmentation results for different segmentation methods: (a) original images (b) labels (c) U-Net with ResNet-50 (d) ResNet-50 Multi-Task (e) U-Net with ST (f) ST Multi-Task.

way for users to access and visualize the segmentation and classification results using our ResNet-50 Multi-Task model.

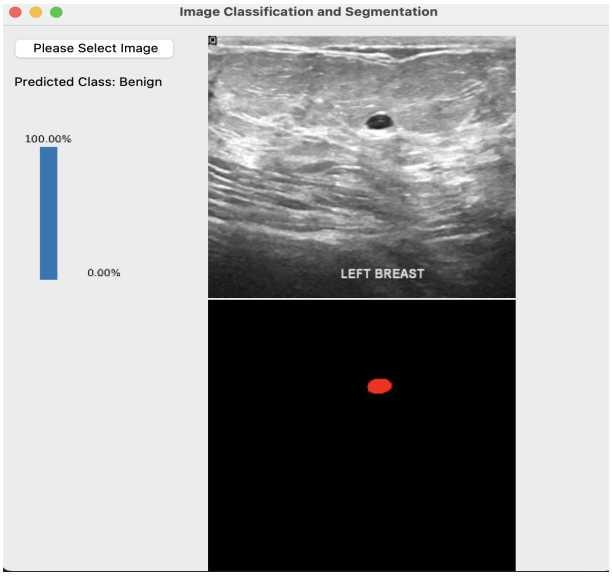

Fig. 7. GUI Demo Developed by Python

V. Conclusion

This study involves an experimental comparison between the Multi-Task Learning (MTL) network and single classification and segmentation networks. Additionally, a traditional convolutional neural network and a transformer-based network are compared. The goal is to identify the most effective method for segmenting and classifying BUS images from the public dataset BUSI. We conducted experiments using six different networks to evaluate their performance throughout the research. The results demonstrate that the ResNet-50 MultiTask method outperforms the dataset's other single-task segmentation, single-task classification methods, and the Swin Transformer method. Despite the promising outcomes, some drawbacks are also identified. Specifically, the Swin Transformer performs less successfully than ResNet-50. The transformer-based method is easier to get overfitting and needs more training epoch numbers. Consequently, future studies may focus on further experimentation and optimization of the Swin Transformer to improve its efficiency and performance. Furthermore, we have developed an interactive GUI that allows users to load images and models, providing them with the ability to visualize segmentation and classification results.

References

[1] N. I. of Health *et al.*, "Nci dictionary of cancer terms-national cancer institute," *Website: https://www. cancer. gov/publi catio ns/dicti onari es/cancer-terms. Accessed March*, vol. 18, 2019.

[2] R. Liang, Z. Liu, X. Piao, M. Zuo, J. Zhang, Z. Liu, Y. Li, and Y. Lin, "Research progress on gp73 in malignant tumors," *OncoTargets and therapy*, pp. 7417–7421, 2018.

[3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[4] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi, "Prediction of breast cancer using machine learning approaches," *Journal of Biomedical Physics & Engineering*, vol. 12, no. 3, p. 297, 2022.

[5] H. Tanaka, S.-W. Chiu, T. Watanabe, S. Kaoku, and T. Yamaguchi, "Computer-aided diagnosis system for breast ultrasound images using deep learning," *Physics in Medicine & Biology*, vol. 64, no. 23, p. 235013, 2019.

[6] W.-C. Shia and D.-R. Chen, "Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101829, 2021.

[7] Z. Zhuang, Z. Yang, S. Zhuang, A. N. Joseph Raj, Y. Yuan, and R. Nersisson, "Multi-features-based automated breast tumor diagnosis using ultrasound image and support vector machine," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–12, 2021.

[8] K. Huang, H.-D. Cheng, Y. Zhang, B. Zhang, P. Xing, and C. Ning, "Medical knowledge constrained semantic breast ultrasound image segmentation," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1193–1198.

[9] M. Amiri, R. Brooks, and H. Rivaz, "Fine-tuning u-net for ultrasound image segmentation: different layers, different outcomes," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2510–2518, 2020.

[10] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[11] G. Zhang, K. Zhao, Y. Hong, X. Qiu, K. Zhang, and B. Wei, "Sha-mtl: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, pp. 1719–1725, 2021.

[12] M. Xu, K. Huang, and X. Qi, "A regional-attentive multi-task learning framework for breast ultrasound image segmentation and classification," *IEEE Access*, vol. 11, pp. 5377–5392, 2023.

[13] W.-X. Liao, P. He, J. Hao, X.-Y. Wang, R.-L. Yang, D. An, and L.-G. Cui, "Automatic identification of breast ultrasound image based on supervised block-based region segmentation algorithm and features combination migration deep learning model," *IEEE journal of biomedical and health informatics*, vol. 24, no. 4, pp. 984–993, 2019.

[14] M. Xian, Y. Zhang, H.-D. Cheng, F. Xu, B. Zhang, and J. Ding, "Automatic breast ultrasound image segmentation: A survey," *Pattern Recognition*, vol. 79, pp. 340–355, 2018.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

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

[17] 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.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.