# Breast Ultrasound Tumor Classification Using a Hybrid Multitask CNN-Transformer Network

Bryar Shareef, Min Xian[(✉)], Aleksandar Vakanski, and Haotian Wang

Department of Computer Science, University of Idaho, Idaho Falls, ID 83402, USA
`mxain@uidaho.edu`

**Abstract.** Capturing global contextual information plays a critical role in breast ultrasound (BUS) image classification. Although convolutional neural networks (CNNs) have demonstrated reliable performance in tumor classification, they have inherent limitations for modeling global and long-range dependencies due to the localized nature of convolution operations. Vision Transformers have an improved capability of capturing global contextual information but may distort the local image patterns due to the tokenization operations. In this study, we proposed a hybrid multitask deep neural network called Hybrid-MT-ESTAN, designed to perform BUS tumor classification and segmentation using a hybrid architecture composed of CNNs and Swin Transformer components. The proposed approach was compared to nine BUS classification methods and evaluated using seven quantitative metrics on a dataset of 3,320 BUS images. The results indicate that Hybrid-MT-ESTAN achieved the highest accuracy, sensitivity, and F1 score of 82.7%, 86.4%, and 86.0%, respectively.

**Keywords:** Breast Ultrasound · Classification · Multitask Learning · Hybrid CNN-Transformer

## 1 Introduction

Breast cancer is the leading cause of cancer-related fatalities among women. Currently, it holds the highest incidence rate of cancer among women in the U.S., and in 2022 it accounted for 31% of all newly diagnosed cancer cases [1]. Due to the high incidence rate, early breast cancer detection is essential for reducing mortality rates and expanding treatment options. BUS imaging is an effective screening option because it is cost-effective, nonradioactive, and noninvasive. However, BUS image analysis is also challenging due to the large variations in tumor shape and appearance, speckle noise, low contrast, weak boundaries, and occurrence of artifacts.

In the past decade, deep learning-based approaches achieved remarkable advancements in BUS tumor classification [2,3]. The progress has been driven by the capability of CNN-based models to learn hierarchies of structured image representations as semantics. To extract deep context features, CNNs apply a series of convolutional and downsampling layers, frequently organized into blocks

with residual connections. Nevertheless, one disadvantage of such architectural choice is that the feature representations in the deeper layers become increasingly abstract, leading to a loss of spatial and contextual information. The intrinsic locality of convolutional operations hinders the ability of CNNs to model long-range dependencies while preserving spatial information in images effectively.

Vision Transformer (ViT) [5] and its variants recently demonstrated superior performance in image classification tasks. These models convert input images into smaller patches and utilize the self-attention mechanism to model the relationships between the patches. Self-attention enables ViTs to capture long-range dependencies and model complex relationships between different regions of the image. However, the effectiveness of ViT-based approaches heavily relies on access to large datasets for learning meaningful representations of input images. This is primarily because the architectural design of ViTs does not rely on the same inductive biases in feature extraction which allow CNNs to learn spatially invariant features.

Accordingly, numerous prior studies introduced modifications to the original ViT network specifically designed for BUS image classification [13,14,23]. In addition, several works proposed network architectures that combined Transformers and CNNs [4,15,16]. For instance, Mo et al. [15] proposed a hybrid CNN-Transformer incorporating BUS anatomical priors. Qu et al. [16] employed squeeze and excitation blocks to enhance the feature extraction capacity in a hybrid CNN-based VGG16 network and ViT. Similarly, Iqbal et al. [4] designed two hybrid CNN-Transformer networks intended either for classification or segmentation of multi-modal breast cancer images. Despite the promising results of such hybrid approaches, effectively capturing the local patterns and global long-range dependencies in BUS images remains challenging [4,5,24].

Multitask learning leverages shared information across related tasks by jointly training the model. It constrains models to learn representations that are relevant to all tasks rather than learning task-specific details. Moreover, multitask learning acts as a regularizer by introducing inductive bias and prevents overfitting [25] (particularly with ViTs), and with that, can mitigate the challenges posed by small BUS dataset sizes. In [3], the authors demonstrated that multitask learning outperforms single-task learning approaches for BUS classification.

In this study, we introduce a hybrid multitask approach, Hybrid-MT-ESTAN, which encompasses tumor classification as a primary task and tumor segmentation as a secondary task. Hybrid-MT-ESTAN combines the advantages of CNNs and Transformers in a framework incorporating anatomical tissue information in BUS images. Specifically, we designed a novel attention block named Anatomy-Aware Attention (AAA), which modifies the attention block of Swin Transformer by considering the breast anatomy. The anatomy of the human breast is categorized into four primary layers: the skin, premammary (subcutaneous fat), mammary, and retromammary layers, where each layer has a distinct texture and generates different echo patterns. The primary layers in BUS images are arranged in a vertical stack, with similar echo patterns appearing horizontally across the images. The kernels in the introduced AAA attention blocks are

organized in rows and columns to capture the anatomical structure of the breast tissue. In the published literature, the closest approach to ours is the work by Iqbal et al. [4], in which the authors used hybrid single-task CNN-Transformer networks for either classification or segmentation of BUS images. Conversely, Hybrid-MT-ESTAN employs a multitask approach and introduces novel architectural design. The main contributions of this work are summarized as:

- The proposed architecture effectively integrates the advantages of CNNs for extracting hierarchical and local patterns in BUS images and Swin Transformers for leveraging long-range dependencies.
- The designed Anatomy-Aware Attention (AAA) block improves the learning of contextual information based on the anatomy of the breast.
- The multitask learning approach leverages the shared representations across the classification and segmentation tasks to improve the model performance.
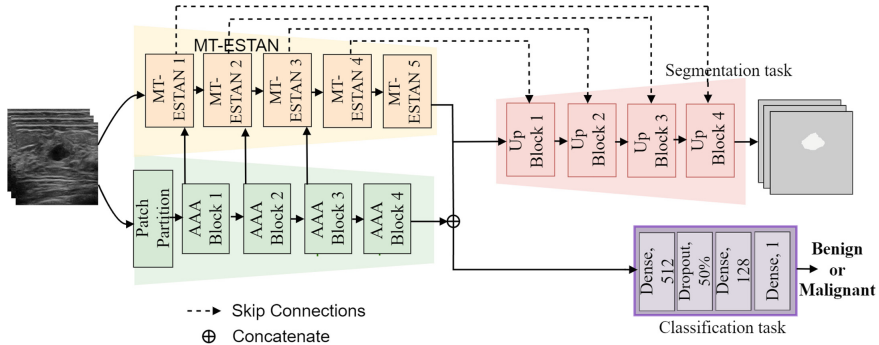


**Fig. 1.** Hybrid-MT-ESTAN consists of MT-ESTAN and AAA encoders, a segmentation branch, and a classification branch.

## 2   Proposed Method

### 2.1   Hybrid-MT-ESTAN

The architecture of Hybrid-MT-ESTAN is shown in Fig. 1, and consists of: (1) the MT-ESTAN encoder [3], and a Swin Transformer-based encoder with Anatomy-Aware Attention (AAA) blocks, (2) a decoder branch for the segmentation task, and (3) a branch with fully-connected layers for the classification task. MT-ESTAN [3] is a CNN-based multitask learning network that simultaneously performs BUS classification and segmentation. The encoder sub-network of MT-ESTAN is ESTAN [17], which employs row-column-wise kernels to learn and fuse context information in BUS images at different context scales (see Fig. 2). Specifically, each MT-ESTAN block is composed of two parallel branches consisting of four square convolutional kernels and two consecutive row-column-wise kernels. These specialized convolutional kernels effectively extract contextual information of small tumors in BUS images. Refer to [17,22], and [3] for the implementation details of ESTAN and MT-ESTAN. The source codes of these works are available at http://busbench.midalab.net.
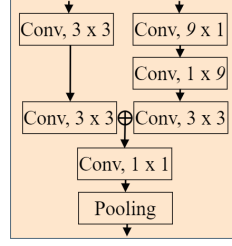
**Fig. 2.** MT-ESTAN blocks include parallel convolutional branches with different kernel size, followed by $1 \times 1$ convolution and a pooling layer.

## 2.2 Anatomy-Aware Attention (AAA) Block

Swin Transformer [18] is a hierarchical transformer-based approach that uses shifted windows to model global context information. Swin Transformer partitions an input image into non-overlapping patches of size $4 \times 4$, where each patch is treated as a "token". A linear layer receives the patches and projects them into an arbitrary dimension. Each Swin Transformer block consists of a LayerNorm layer (LN) layer, a multi-head self-attention module (MSA), and a multi-layer perceptron (MLP) with GELU activation. To model long-range dependencies, the original Swin Transformer relies on shifted windows, where the window-based multi-head self-attention (W-MSA) and shifted window-based multi-head self-attention (SW-MSA) modules are employed in each consecutive Swin block. The Swin block is formulated as follows.

$$\hat{f}^l = \text{W-MSA}(\text{LN}(f^{l-1})) + f^{l-1} \tag{1}$$

$$f^l = \text{MLP}(\text{LN}(\hat{f}^l)) + \hat{f}^l \tag{2}$$

$$\hat{f}^{l+1} = \text{SW-MSA}(\text{LN}(f^l)) + f^l \tag{3}$$

$$f^{l+1} = \text{MLP}(\text{LN}(\hat{f}^{l+1})) + \hat{f}^{l+1} \tag{4}$$

where $f^l$ and $\hat{f}^l$ are the output features of the MLP module and the (S)W-MSA module for block $l$, respectively; in the proposed Anatomy-Aware Attention (AAA) block, we redesigned the Swin blocks to enhance their ability to model both global and local features by adding an attention block based on the breast anatomy (see Fig. 3). The additional layers are defined by

$$y^i = M(f^{l+1}) \tag{5}$$

$$B^i = U(\text{MAX-P}(y^i) + \text{AVG-P}(y^i)) \tag{6}$$

$$O^i = y^i \cdot (\sigma(A(B))) \tag{7}$$

Concretely, we first reconstruct the $i$-th feature map ($y^i$) by merging ($M$) all patches, and afterward, we applied average pooling (AVG-P) and max pooling

(MAX-P) layers with size (2, 2). The outputs of (AVG-P) and (MAX-P) layers are concatenated and UP-SAMPLED ($U$) with size (2, 2) and stride (2, 2). ROW-COLUMN-WISE kernels ($A$) with size (9 , 1) and (1 , 9) are then employed to adapt to the anatomy of the breast, and finally a sigmoid function ($\sigma$) is applied to the output of (A) multiplied by the input feature map ($y^i$).
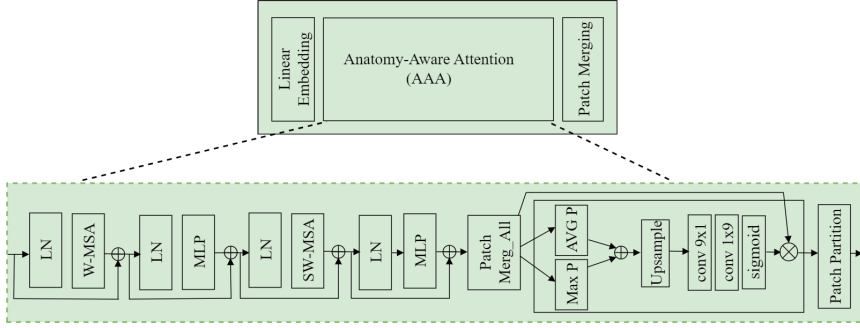


**Fig. 3.** Anatomy-Aware Attention (AAA) block.

### 2.3    Segmentation and Classification Branches/Tasks

The segmentation branch in Fig. 1 outputs dense mask predictions of BUS tumors. It consists of four Up Blocks, each with three convolutional layers and one upsampling layer (with size (2, 2) and stride (2, 2)). The settings of the convolutional layers are adopted from [3]. In addition, the blocks receive four skip connections from the MT-ESTAN encoder, i.e., there is a skip connection from each MT-ESTAN block 1 to 4. The classification branch consists of three dense layers, a dropout layer (50%), and the final dense layer that predicts the tumor class into benign or malignant.

### 2.4    Loss Function

We applied a multitask loss function ($L_{mt}$) that aggregates two terms: a focal loss $L_{Focal}$ for the classification task and dice loss $L_{Dice}$ for the segmentation task. Therefore, the composite loss function is $L_{mt} = w_1 \cdot L_{Focal} + L_{Dice}$, where the weight coefficient $w_1$ is set to apply greater importance to the classification task as the primary task. Since in medical image diagnosis achieving high sensitivity places emphasis on the detection of malignant lesions, we employed the focal loss for the classification task to trade off between sensitivity and specificity. Because malignant tumors are more challenging to detect due to greater differences in margin, shape, and appearance in BUS images, focal loss forces the model to focus more on difficult predictions. Specifically, focal loss adds a factor $(1 - p_i)^\gamma$ to the cross-entropy loss where $\gamma$ is a focusing parameter, resulting in $L_{Focal} = -1/N \sum_{i=1}^{N} [(\alpha \cdot t_i \cdot (1-p_i)^\gamma \cdot log(p_i) + (1-\alpha) \cdot p_i \cdot log(1-p_i)]$. In the formulation, $\alpha$

is a weighting coefficient, $N$ denotes the number of image samples, $t_i$ is the target label of the $i^{th}$ training sample, and $p_i$ denotes the prediction. The segmentation loss is calculated using the commonly-employed Dice loss ($L_{Dice}$) function.

## 3   Experimental Results

### 3.1   Datasets

We evaluated the performance of Hybrid-MT-ESTAN using four public datasets, HMSS [9], BUSI [10], BUSIS [20], and Dataset B [6]. We combined all four datasets to build a large and diverse dataset with a total of 3,320 B-mode BUS images, of which 1,664 contain benign tumors and 1,656 have malignant tumors. Table 1 shows the detailed information for each dataset. HMSS dataset does not provide the segmentation ground-truth masks, and for this study we arranged with a group of experienced radiologists to prepare the masks for HMSS. Refer to the original publications of the datasets for more details.

**Table 1.** Breast ultrasound (BUS) datasets. 'b' denotes benign tumor and 'm' is malignant tumor.

| BUS dataset | No. of images | Distribution | Source |
|---|---|---|---|
| HMSS | 1,948 | b:812, m:1136 | Netherlands |
| BUSI | 647 | b:437, m:210 | Egypt |
| BUSIS | 562 | b:306, m:256 | China |
| Dataset B | 163 | b:109, m:54 | Spain |
| Total | 3,320 | b: 1,664, m: 1,656 | |

### 3.2   Evaluation Metrics

For performance evaluation of the classification task, we used the following metrics: accuracy (Acc), sensitivity (Sens), specificity (Spec), F1 score, Area Under the Curve of Receiver Operating Characteristic (AUC), false positive rate (FPR), and false negative rate (FNR). To evaluate the segmentation performance, we used dice similarity coefficient (DSC) and Jaccard index (JI).

### 3.3   Implementation Details

The proposed approach was implemented with Keras and TensorFlow libraries. All experiments were performed on a machine with NVIDIA Quadro RTX 8000 GPUs and two Intel Xeon Silver 4210R CPUs (2.40GHz) with 512 GB of RAM. All BUS images in the dataset were zero-padded and reshaped to form square images. To avoid data leakage and bias, we selected the train, test, and validation sets based on the cases, i.e., the images from one case (patient) were

**Table 2.** Performance metrics of the compared methods for BUS image classification and segmentation.

| Methods | Classification | | | | | | | Segmentation | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc↑ | Sens.↑ | Spec.↑ | F1↑ | Auc↑ | FNR↓ | FPR↓ | DSC↑ | JI↑ |
| SHA-MTL [8] | 69.6 | 48.1 | **90.8** | 0.58 | 69.5 | 51.9 | **9.2** | 72.2 | 60.7 |
| MobileNet [19] | 71.0 | 82.0 | 61.0 | 0.74 | 71.5 | 18.0 | 39.0 | - | - |
| VGGA-ViT [16] | 73.6 | 61.8 | 79.8 | 0.61 | 70.8 | 38.2 | 20.2 | 74.9 | 64.9 |
| DenseNet121 [7] | 73.0 | 74.0 | 71.0 | 0.73 | 72.5 | 26.0 | 29.0 | - | - |
| EMT-Net [12] | 74.1 | 79.4 | 69.1 | 0.75 | 74.3 | 20.6 | 30.9 | 76.7 | 67.0 |
| ViT [5] | 72.1 | 74.1 | 69.3 | 0.73 | 71.7 | 25.9 | 30.7 | - | - |
| Chowdery [10] | 77.4 | 77.3 | 77.3 | 0.77 | 77.3 | 22.7 | 22.7 | 77.0 | 67.9 |
| Swin Transformer | 77.4 | 72.6 | 82.5 | 0.74 | 77.6 | 27.4 | 17.5 | - | - |
| MT-ESTAN | 78.6 | 83.7 | 72.6 | 0.83 | 78.2 | 16.3 | 27.4 | 78.2 | 69.3 |
| Ours | **82.8** | **86.4** | 79.2 | **0.86** | **82.8** | **13.6** | 20.8 | **84.1** | **75.7** |

Note: A dash '-' in the Segmentation column indicates that the model uses single-task learning.

assigned to only one of the training, validation, and test sets. Furthermore, we employed horizontal flip, height shift (20%), width shift (20%), and rotation (20°C) for data augmentation. The proposed approach utilizes the building blocks of ResNet50 and Swin-Transformer-V2, pretrained on ImageNet dataset. Namely, MT-ESTAN uses pretrained ResNet50 as a base model for the five encoder blocks (the implementation details of MT-ESTAN can be found in [3]). The encoder with AAA blocks uses the SwinTransformer_V2_Base_256 pretrained model as a backbone. For the composite loss function, we adopted a weight coefficient $w_1 = 3$, and in the focal loss $\alpha = 0.5$ and $\gamma = 2$. For model training we utilized Adam optimizer with a learning rate of $10^{-5}$ and mini batch size of 4 images.

### 3.4   Performance Evaluation and Comparative Analysis

We compared the performance of Hybrid-MT-ESTAN for BUS classification to nine deep learning approaches commonly used for medical image analysis. The compared models include CNN-based, ViT-based, and hybrid approaches. CNN-based networks are SHA-MTL [8], MobileNet [19], DenseNet121 [7], and EMT-Net [12]. ViT-based approaches include the original ViT [5], Chowdery [10], and Swin Transformer [18]. VGGA-ViT [16] is a hybrid CNN-Transformer network. The values of the performance metrics are shown in Table 2, indicating that the proposed Hybrid-MT-ESTAN outperformed all nine approaches by achieving the best accuracy, sensitivity, F1 score, and AUC with 82.8%, 86.4%, 86.0%, and 82.8%, respectively. Although SHA-MTL [8] obtained the highest specificity of 90.8% and FNR of 9.2%, the trade-off between sensitivity and specificity should be taken into consideration, as that approach had sensitivity of 48.1%.

**Table 3.** Ablation study for evaluating the components of Hybrid-MT-ESTAN.

| Methods | Classification | | | | | | | Segmentation | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc↑ | Sens.↑ | Spec.↑ | F1↑ | Auc↑ | FNR↓ | FPR↓ | DSC↑ | JI↑ |
| MT-ESTAN [10] | 78.6 | 83.7 | 72.6 | 0.83 | 78.2 | 16.3 | 27.4 | 78.2 | 69.3 |
| Swin Trans. | 77.4 | 72.6 | **82.5** | 0.74 | 77.6 | 27.4 | 17.5 | - | - |
| MT-ESTAN + Swin Trans. | 80.3 | 84.2 | 76.3 | 0.83 | 80.2 | 15.8 | 23.7 | 82.3 | 73.6 |
| Ours | **82.8** | **86.4** | 79.2 | **0.86** | **82.8** | **13.6** | 20.8 | **84.1** | **75.7** |

The preferred trade-off in medical image analysis typically is high sensitivity without significant degradation in specificity.

We evaluated the segmentation performance of Hybrid MT-ESTAN and compared the results to five multitask approaches, including SHA-MTL [8], EMT-Net [12], Chowdery [10], MT-ESTAN [3], and VGGA-ViT [16]. As shown in Table 2, the proposed Hybrid MT-ESTAN achieved the highest performance and increased DSC and JI by 5.9% and 6.4%, respectively compared to MT-ESTAN. Note that results of single-task models in Table 2 are not provided.

### 3.5 Effectiveness of the Anatomy-Aware Attention (AAA) Block

To verify the effectiveness of the Anatomy-Aware Attention (AAA) block, we conducted an ablation study that quantified the impact of the different components in Hybrid-MT-ESTAN on the classification and segmentation performance. Table 3 presents the values of the performance metrics for MT-ESTAN (pure CNN-based approach), Swin Transformer (pure Transformer network), a hybrid architecture of MT-ESTAN and Swin Transformer, and our proposed Hybrid-MT-ESTAN with AAA block. According to the results in Table 3, MT-ESTAN achieved better sensitivity and F1 score than Swin Transformer, with 83.7% and 83%, respectively. The hybrid architectures of MT-ESTAN with Swin Transformer improved the classification performance and has higher accuracy, sensitivity, F1 score, and AUC with 80.3%, 84.2%, 83%, and 80.2%, compared to MT-ESTAN and Swin Transformer individually. The proposed approach, Hybrid-MT-ESTAN with AAA block, further improved accuracy, sensitivity, F1 score, and AUC by 2.5%, 2.2%, 3%, and 2.6%, respectively, relative to the hybrid model without the AAA block.

To evaluate the segmentation performance, we compared the proposed approach with and without the AAA block and Swin Transformer. As shown in Table 3, MT-ESTAN combined with Swin Transformer improved DSC and JI by 4.1% and 4.3%, respectively compared to MT-ESTAN. Employing the proposed AAA block further improved DSC and JI by 1.8% and 2.1%, respectively.

## 4   Conclusion

In this paper, we introduced the Hybrid-MT-ESTAN, a multitask learning approach for BUS image analysis that alleviates the lack of global contextual infor-

mation in the low-level layers of CNN-based approaches. Hybrid-MT-ESTAN concurrently performs BUS tumor classification and segmentation, with a hybrid architecture that employs CNN-based and Swin Transformer layers. The proposed approach exploits multi-scale local patterns and global long-range dependencies provided by MT-ESTA and AAA Transformer blocks for learning feature representations, resulting in improved generalization. Experimental validation demonstrated significant performance improvement by Hybrid-MT-ESTAN in comparison to current state-of-the-art models for BUS classification.

# References

1. American Cancer Society, Cancer Facts & Figures, https://www.cancer.org, (2022)
2. Zhuang, Z., Yang, Z., Raj, A., Noel, J., Wei, C.: Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. Comput. Methods Programs Biomed. **208**, 106221 (2021)
3. Shareef, B., et al.: A benchmark for breast ultrasound image classification. Available at SSRN (2023). https://ssrn.com/abstract=4339660
4. Iqbal, A., Sharif, M., BTS-ST: Swin transformer network for segmentation and classification of multimodality breast cancer images. Knowl. Based Syst. **262**, 110393 (2023)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale, preprint arXiv:2010.11929 (2020)
6. Yap, M., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inf. **22**(4), 1218–1226 (2017)
7. Huang, G., Liu, Z., Mateen, L., Weinberger, K.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017)
8. Zhang, G., Zhao, K., Hong, Y., Qiu, X., Zhang, K., Wei, B.: SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification. Int. J. Comput. Assisted Radiol. Surg. **16**(10), 1719–1725 (2021). https://doi.org/10.1007/s11548-021-02445-7
9. Geertsma, T., Fujifilm.: Ultrasound cases, https://www.ultrasoundcases.info/ (2014)
10. Chowdary, J., Yogarajah, P., Chaurasia, P., Guruviah, V.: A multi-task learning framework for automated segmentation and classification of breast tumors from ultrasound images. Ultrason. Imaging **44**(1), 3–12 (2022)
11. Vakanski, A., Xian, M.: Evaluation of complexity measures for deep learning generalization in medical image analysis. In: 2021 IEEE 31st Int. Workshop on MLSP, pp. 1–6 (2021)
12. Shi, J., Vakanski, A., Xian, M., Ding, J., Ning, C.: EMT-NET: efficient multi-task network for computer-aided diagnosis of breast cancer. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022)

13. Gheflati, B., Rivaz, H.: Vision transformers for classification of breast ultrasound images, In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 480–483 (2022)
14. Hassanien, A., Singh, K., Puig, D., Abdel-Nasser, M.: Predicting breast tumor malignancy using deep ConvNeXt radiomics and quality-based score pooling in ultrasound sequences. Diagnostics **12**(5), 1053 (2022)
15. Mo, Y., et al.: Hover-trans: anatomy-aware hover-transformer for ROI-Free breast cancer diagnosis in ultrasound images IEEE Trans. Med. Imaging **42**, 1696–1706 (2023)
16. Qu, X., et al.: A VGG attention vision transformer network for benign and malignant classification of breast ultrasound images. Med. Phys. **49**(9), 5787–5798 (2022)
17. Shareef, B., Vakanski, A., Freer, P., Min Xian: ESTAN: enhanced small tumor-aware network for breast ultrasound image segmentation. Healthcare **10**(11), 2262 (2022)
18. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
19. Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T.: MobileNets: efficient convolutional neural networks for mobile vision applications. preprint arXiv:1704.04861 (2017)
20. Zhang, Y., et al.: BUSIS: a benchmark for breast ultrasound image segmentation. Healthcare **10**(4), 729 (2022)
21. Yap, M., et al.: Breast ultrasound lesions recognition: end-to-end deep learning approaches. J. Med. Imaging **6**(1), 011007. SPIE (2018)
22. Shareef, B., Xian, M., Vakanski, A.: Stan: small tumor-aware network for breast ultrasound image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2020)
23. Ayana, G., Choe, S.: BUViTNet: breast ultrasound detection via vision transformers, Diagnostics **12**(11), 2654 (2022). https://doi.org/10.3390/diagnostics12112654
24. Tang, S., et al.: Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. Comput. Bio. Med. **157**, 106723 (2023)
25. Sebastian, R.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)