Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(8): 15244–15264.
DOI: 10.3934/mbe.2023682
Received: 29 May 2023
Revised: 29 June 2023
Accepted: 04 July 2023
Published: 20 July 2023

Research article

Improved breast ultrasound tumor classification using dual-input CNN with GAP-guided attention loss

Xiao Zou[1,*], Jintao Zhai[1], Shengyou Qian[1], Ang Li[1], Feng Tian[1], Xiaofei Cao[2] and Runmin Wang[2,*]

[1] School of Physics and Electronics, Hunan Normal University, Changsha 410081, China

[2] College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China

* **Correspondence:** Email: shawner@hunnu.edu.cn, runminwang@hunnu.edu.cn.

Abstract: Ultrasonography is a widely used medical imaging technique for detecting breast cancer. While manual diagnostic methods are subject to variability and time-consuming, computer-aided diagnostic (CAD) methods have proven to be more efficient. However, current CAD approaches neglect the impact of noise and artifacts on the accuracy of image analysis. To enhance the precision of breast ultrasound image analysis for identifying tissues, organs and lesions, we propose a novel approach for improved tumor classification through a dual-input model and global average pooling (GAP)-guided attention loss function. Our approach leverages a convolutional neural network with transformer architecture and modifies the single-input model for dual-input. This technique employs a fusion module and GAP operation-guided attention loss function simultaneously to supervise the extraction of effective features from the target region and mitigate the effect of information loss or redundancy on misclassification. Our proposed method has three key features: (i) ResNet and MobileViT are combined to enhance local and global information extraction. In addition, a dual-input channel is designed to include both attention images and original breast ultrasound images, mitigating the impact of noise and artifacts in ultrasound images. (ii) A fusion module and GAP operation-guided attention loss function are proposed to improve the fusion of dual-channel feature information, as well as supervise and constrain the weight of the attention mechanism on the fused focus region. (iii) Using the collected uterine fibroid ultrasound dataset to train ResNet18 and load the pre-trained weights, our experiments on the BUSI and BUSC public datasets demonstrate that the proposed method outperforms some state-of-the-art methods. The code will be publicly released at https://github.com/425877/Improved-Breast-Ultrasound-Tumor-Classification.

Keywords: ultrasound image; breast ultrasound tumor; convolutional neural network; feature fusion; classification

1. Introduction

Medical image classification is a critical task in accurately identifying lesions in targeted areas and distinguishing intricate lesion information that may transcend human perception, ultimately enhancing the reliability of medical diagnosis. Despite the potential benefits, accurate classification is often hindered by the limited variability in the morphological characteristics, location and size of benign and malignant tumors. Consequently, the precise and reliable classification of medical images in clinical settings remains a crucial and challenging objective [1].

Breast cancer is a widespread malignancy that holds the dubious distinction of being the most prevalent cancer globally, accounting for an incidence of 24.2% of all female cancers [2]. A large number of researchers have carried out research on the diagnosis of breast cancer [3–8]. Ultrasound technology is a non-invasive and reproducible diagnostic modality applied for breast cancer detection. To facilitate quantitative clinical analysis, it is imperative to achieve the accurate and effective automatic classification of pathological information. Ultrasound images provide comprehensive information regarding the soft tissue layers and pathologic findings of the breast, thereby rendering the classification of breast ultrasound images crucial for the determination of pathological information. However, a considerable amount of noise is present in ultrasound images, including characteristic speckles caused by the interference of acoustic signals during the imaging process [9, 10]. Moreover, the classification of pathological information poses a significant challenge due to the variability of breast soft tissues and the complexity of target shapes.

The problem of ultrasound image classification can be tackled using traditional methods and deep learning-based methods. Traditional methods, including support vector machines [11–13], decision trees [6, 14, 15], and random forests [4, 16, 17], have been applied for breast image classification. However, these methods are usually dependent on the supervision of medical professionals and highly sensitive to noise.

In recent times, the confluence of deep learning and medical image processing has been instrumental in addressing the challenge of intricate and time-consuming manual feature selection. By leveraging extensive data for training, computers can autonomously learn features. Convolutional neural networks (CNNs) have emerged as a popular deep learning technique due to their superior capacity for extracting local information and delivering superior classification outcomes. Consequently, CNN-based classification methods have gained substantial traction in subsequent research endeavors. For example, Das et al. [18] combined CNNs with image preprocessing to achieve the automatic classification of brain tumors. Hao et al. [19] proposed an active learning framework model for tumor classification based on transfer learning, using a pre-trained model to calculate the classification probability of each sample. Zhang et al. [20] improved the average pooling layer in the residual network (ResNet) to classify X-ray images. The self-attention mechanism of the transformer structure can help the network capture global contextual information and compensate for the shortcomings of traditional CNNs. Dai et al. [21] proposed Transmed, which combines the respective advantages of the CNN and transformer to extract both local and global information from medical images. Aladhadh [22] designed a data-enhanced transformer for the classification of skin cancer images, which expands the amount of data and improves the generalization ability of the network through operations such as flipping and scaling. Moreover, some recent papers [23–26], all based on transformers, further validate the feasibility of transformers by comparing them with CNN-based networks.

In the field of breast pathological information classification, previous studies have made significant progress, such as that by Spanhol et al. [27] who applied AlexNet to the BreaKHis dataset and achieved a recognition rate 6% higher than traditional machine learning algorithms. Lotter et al. [28] proposed a multi-scale CNN and optimized the learning strategy to improve the accuracy of breast x-ray image classification. The authors of [29] designed a hybrid model combining a CNN and long short-term memory network to enhance the network's long-range dependence capability. Mewada et al. [30] proposed a novel CNN structure for the classification of histopathological cancer images by combining spectral features obtained from a wavelet transform with the spatial features of a CNN. Despite the significant progress that has been made in recent years, the performance of neural networks in the field of computer vision, specifically in the analysis of ultrasound images, can still be hindered by speckle noise. Moreover, the limited input of real ultrasound images presents a constraint on the continued enhancement of network performance. In order to address these challenges, we propose a novel approach which involves modifying the single channel network input to a dual channel input. This modification serves to enhance the focus on the target region, while simultaneously reducing the negative impact of speckle noise on the preservation of pathological information. Inspired by [29], our work not only combines a CNN and transformer, but it also improves the overall loss of the network which assists the network in the extraction and fusion of features.

The main contributions of this research are as follows:

• We leverage a combination of ResNet and MobileViT architectures to enhance the model's capacity for extracting both local and global information. In addition, we integrate the original images and corresponding attention images into a dual-input channel, which effectively mitigates the impact of noise and artifacts.

• Our proposed dual-input feature fusion module is designed by incorporating a guided attention loss operation based on global average pooling (GAP). Furthermore, we utilize high-level feature information generated by GAP to optimize the attention weights.

• Before using ResNet18 as the backbone network in this study, we used the collected uterine fibroid ultrasound dataset and trained it for the segmentation task. When training our classification model, pretraining weights of ResNet18 were loaded to improve its ability to extract feature information from ultrasound images.

• We conducted experiments on the BUSI [31] and the BUSC [32] datasets; the results demonstrates that our method outperforms some state-of-the-art methods on the aforementioned datasets.

2. Related work

2.1. Multi-channel input

Obtaining reliable classification outcomes with raw breast ultrasound images alone is a challenging task due to various issues such as noise, artifacts and interference from surrounding regions of the lesion [33, 34]. Hence, researchers have shown great interest in multi-image or multi-model fusion for ultrasound image classification. In this regard, the authors of [35] have suggested feeding each breast ultrasound image together with its mask into the classification network to compensate for information loss caused by noise. In [36], three classical networks have been utilized to extract breast ultrasound image features independently and fuse the obtained features. Other related studies such as [37, 38] have also demonstrated the benefits of incorporating multiple input channels. In this work, we adopt

the dual channel input approach to train the model and enhance the network's capacity to learn edge and other critical information.

2.2. Transformers

CNNs [18–20, 28] have demonstrated their proficiency in extracting local information, but the accumulation of convolutional layers causes a loss of effective information and the number of parameters increases. The transformer architecture, initially proposed in the context of natural language processing (NLP), has produced satisfactory outcomes on various NLP tasks. Unlike CNNs, the transformer architecture can extract global features of images and outperform a CNN after being trained on a considerable volume of data. Nevertheless, the self-attention mechanism of the transformer often overlooks local feature details. As a remedy, the author of [21] proposed a hybrid approach involving the combination of a CNN and transformer for the extraction of local and global features from medical images, respectively. Several works [22–26, 33] have reported remarkable performance gains in network classification resulting from the introduction of the transformer architecture. In previous work we found that the transformer structure has a large number of parameters and is more time consuming to train. In contrast, Mobile ViT has fewer parameters and lower computational complexity. It employs lightweight convolution and self-attention mechanisms to reduce computational and storage costs while maintaining model performance. However, the amalgamation of the ResNet and MobileViT networks should not be a simple concatenation, as the connection might not discern whether the transformer structure contains critical information. To address this issue, we present a novel approach in this paper, wherein we propose combining the ResNet architecture, which is based on CNNs, with the lightweight MobileViT network. Additionally, we design a feature fusion module to serve as the bridging component between the two networks. The efficacy of these enhancements is demonstrated through subsequent experiments.

2.3. Loss function

The loss function constitutes a pivotal element in the training of neural networks, quantifying the disparity between the predicted output and the ground truth. Through backpropagation, the network's parameters are adjusted to minimize this loss, leading to convergence. However, the usage of excessively deep network layers during training can result in the loss of crucial information. Hence, in recent years, significant research efforts have been dedicated to the refinement of loss functions with the aim of achieving superior experimental outcomes. Specifically, in [39], the authors proposed an enhanced loss function for deep convolutional networks to improve the training effect and classification ability of the network. On a breast cancer classification task it obtained high accuracy. The authors of [40] introduced a new integrated loss function to improve the model discrepancy between classified lesions and their labels; their model achieved better performance in terms of breast cancer classification. Similar efforts have been made in [41–43] to enhance the loss function for medical image classification, with promising accuracy improvements. In our study, the usage of two distinct model architectures, ResNet and MobileViT, has led to the creation of over-deep network layers. To improve the fusion of dual input information, we propose the integration of a GAP operation-guided attention loss mechanism that facilitates the selection of essential features, thereby preventing any loss or redundancy of information. This proposal is motivated by similar endeavors that aim to enhance the fusion of information in neural

networks.

3. Approach

This section presents a comprehensive introduction of the classification network structure, followed by a detailed description of the two-channel input, the fusion module, and the improved loss function. Moreover, the essentiality of output visualization is analyzed.

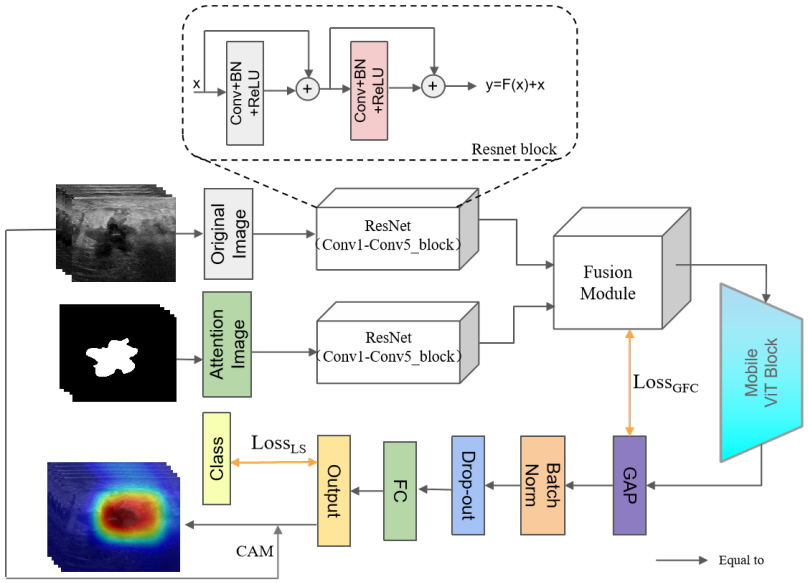

Figure 1. Structure of the overall network.

3.1. Structure of the overall network

In this study, a classification network has been designed for breast ultrasound tumor images by effectively combining the ResNet and MobileNet, based on previous works [44, 45] and [46, 47]. The network utilizes dual-channel input, feature fusion, and the GAP operation-guided attention loss. Figure 1 illustrates the network architecture, in which ResNet extracts features from the input images and attention images, producing feature extraction maps. These maps are then passed to the fusion module, which generates attention scores for feature extraction and integration of information from the original and attention images, as discussed in Section 3.2. The GAP operation-guided attention loss, as explained in Section 3.3, is used to supervise and adjust the weights generated by the attention mechanism. The MobileViT module is subsequently employed to obtain contextual data and incorporate global information. Finally, a simple classification header is applied to obtain the prediction category. It is worth noting that class activation mapping (CAM) is employed during training and testing to aid physicians in verifying the classification network's output categories' reliability while visualizing the output results, as elaborated in Section 3.4.

3.2. Fusion module

In the field of medical classification, the extraction and integration of multiple sources of information is essential to achieve accurate predictive results. While current approaches often concentrate on extracting information from individual images, the presence of noise and artifacts in ultrasound images presents a significant challenge to the extraction of relevant information. To address this issue, we propose the use of attentional images to guide the model in identifying and emphasizing important regions within the input. Our research also focuses on the effective fusion of dual-input feature information to further improve the accuracy of predictions. In addition, attention mechanisms have been widely utilized to extract effective features of images, as demonstrated in several recent studies [48–50]. Motivated by these works, we introduce a dual-channel feature fusion module (see Figure 2) that can extract and aggregate semantic information from different spatial domains of original images and attention images. Specifically, we utilize ResNet to extract the feature information of both the original image and the attention images, and we overlap them using 1×1 convolution for cross-channel interaction and information integration. We then perform channel dimensionality reduction, followed by GAP and the implementation of commonly used attention mechanisms to generate attention scores. These scores are multiplied for each channel of the input feature map to obtain a weight map, which is then sliced and processed before being passed to Mobile ViT Block.

In Mobile ViT Block, the self-attention mechanism is used to calculate the correlation between each location in the feature map and other locations to determine the importance of each location. This approach enables adaptive focus on regions with important feature information. The self-attention mechanism usually consists of multiple attention heads, each of which can independently learn and focus on a different feature representation to extract valid information from the feature map.

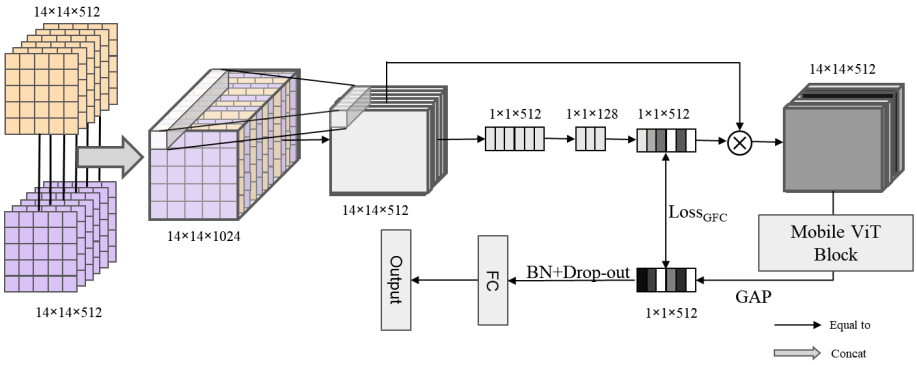

Figure 2. Structure of the fusion module.

3.3. Improved loss function

In medical image classification, the loss function plays a critical role in supervising the network and achieving fast convergence by measuring the error between the target and predicted results. However, in order to avoid information loss caused by over-deep layers of the network in this paper, and, considering the fusion of information in the fusion module, which may cause the redundancy and loss

of information, we propose to split the overall loss function into two parts (indicated by the yellow arrows in Figure 1); besides the loss function of the target and prediction results, the loss of the GAP operation-guided attention is added. This approach aims to enhance the feature extraction ability of the network, while also avoiding the loss of critical image information. In the subsequent experiments, we demonstrate the importance of the improved loss function.

In Figure 1, the weights of each channel α are obtained after the feature fusion module operation. After that, δ_{k} is obtained by the calculation of Mobile ViT Block and simple classification head GAP, where δ_{k} can reflect the contribution of each channel to the classification, and we use it to constrain the channel attention mechanism α. Here, we propose GAP operation-guided attention loss, GFC-Loss in short, to enhance the fusion module's ability to integrate information and prevent the loss of important information; it can be defined as follows:

$$L_{\mathrm{GFC}}=\frac{1}{2}(KL(\alpha\|\delta_{\mathrm{k}}) + KL(\delta_{\mathrm{k}}\|\alpha)) \tag{3.1}$$

where $KL(\mathrm{x}\|\mathrm{y})$ is the Kullback-Leibler divergence from x to y.

In addition, we utilize the classical cross-entropy loss (L_{CE}) to calculate the difference in probability distribution between the predicted outcome and the true label, and the total network-wide loss function can be defined as follows:

$$L = L_{\mathrm{CE}}\,(x, \widetilde{x}) + \lambda L_{\mathrm{GFC}} \tag{3.2}$$

where $\widetilde{x}$ is the final output category of the classification network; the specific implementation of L_{CE} is described in [51]. λ is the hyperparameter, which is set to 0.4 and described in detail in subsequent experiments.

3.4. Gradient-weighted CAM

In our study, we have incorporated the gradient-weighted CAM(Grad-CAM) visualization technique to generate a heat map that is presented alongside the output prediction categories of our proposed model. By leveraging this approach, we aim to overcome the limitations of conventional medical classification networks and effectively identify lesion regions, while mitigating potential misclassification. The generated heat map facilitates the visual interpretation of the model's discrimination between cancer categories, and enables medical professionals to accurately identify crucial areas within the input images. For further details, please refer to [52]. Briefly, the final output of the network is back-propagated, and the gradient information of each layer is collected and weighted to obtain the visualization result, whose weights are calculated by using the following formula.

$$\delta_{\mathrm{c}}^{\mathrm{k}} = \frac{1}{W \times H} \sum_{i}^{W} \sum_{j}^{H} \frac{\partial y^{k}}{\partial A_{i,j}^{k}} \tag{3.3}$$

In this formula, $A_{i,j}^{k}$ represents the data of feature layer A in channel k with coordinates at position i, j; y^{k} denotes the score predicted by the network for category c; H and W are the width and height of the image respectively.

4. Experiment

In this section, we present a comprehensive outline of the dataset employed to evaluate the network's performance, followed by a meticulous description of the implementation procedures. Furthermore, we provide an extensive analysis of the experimental results obtained from the evaluation set.

4.1. Datasets

• Uterine fibroid surveillance ultrasound datasets were collected from the Shenzhen Pro-Huiren Company. We used this dataset to train ResNet18 to complete the segmentation task and load its pre-trained weights to the classification task in this paper. The dataset consisted of 495 ultrasound surveillance images of clinical treatments for uterine fibroids. The target area of treatment is the tumor region in the ultrasound images, which was localized by the clinician.

• The breast ultrasound images dataset (BUSI) [31], which consists of medical images of breast cancer obtained through ultrasound scanning, was categorized into three distinct classes, namely normal, benign, and malignant. With an average image size of 500×500 pixels, this dataset was randomly divided into a training set and a test set using five-fold cross-validation at a ratio of 7:3.

• The Mendeley ultrasound dataset (BUSC) [32] includes 100 benign images and 150 malignant cancer images. The original resolution of the breast ultrasound images was 64×64 pixels, which was later converted to 128×128 pixels. The dataset contains the original images, the labels and the segmented areas of its tumors, all annotated by specialized physicians. In this study, we divided the training and test sets using five-fold cross-validation at a 7:3 ratio.

It is worth noting that the ablation experiments in this paper were all performed on the BUSI dataset. Through comparison with different methods, we further demonstrate the superiority of the proposed method in this paper by using the BUSC dataset and the BUSI dataset.

During the experiments, we employed *Accuracy*, *Precision*, *Recall* and the F_1 score as the evaluation metrics to assess the effectiveness of various classification networks. These metrics were calculated as follows:

$$\text{Accuracy} = \frac{T_P + T_N}{F_N + F_P + T_P + T_N} \tag{4.1}$$

$$\text{Precision} = \frac{T_P}{F_P + T_P} \tag{4.2}$$

$$\text{Recall} = \frac{T_P}{F_N + T_P} \tag{4.3}$$

$$\text{F}_1 = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Here, true positive (T_P) indicates that positive samples are classified as positive samples, true negative (T_N) indicates that negative samples are classified as negative samples, false positive (F_P) indicates that negative samples are classified as positive samples, and false negative (F_N) indicates that the samples are predicted as positive but classified as negative ones.

4.2. Implementation details

All experiments were conducted by using PyTorch on an NVIDIA GeForce RTX 3080 Ti GPU. To mitigate overfitting, we employed data augmentation techniques such as horizontal flipping, vertical flipping, random rotation and center cropping. Moreover, to enable a comparative analysis of the proposed attention mechanism with other methods, each image was resized to 448×448.

Backbone. ResNet18 was used as the backbone network for feature extraction. To enhance the network's learning of ultrasound image features, we pre-trained the ResNet using an ultrasound dataset of uterine fibroids. Specifically, a simple segmentation header was added to ResNet18 and it was trained for the segmentation task by using a dataset containing the original images and the target regions of the tumor delineated by a specialist physician. When the training stabilized, the ResNet pre-training weights, except the segmentation head, were loaded and used for the classification task in this study.

Training. For the training of the classification task, after loading the pre-training weights of ResNet18, the overall network was trained on the BUSI dataset using the loss function proposed in this paper. The training initial learning rate was set to 1×10^{-3} and the weights decayed to 5×10^{-2}. The stochastic gradient descent algorithm with 0.99 momentum was used.

4.3. Experimental results

4.3.1. Ablation study on dual channel input

We trained and evaluated seven representative classification models using a standardized dataset (i.e., the BUSI dataset) and training methodology. The term "dual" in Table 1 refers to utilizing two models of the same structure to extract feature information from an ultrasound image and its corresponding attention map separately (weights are not shared between models). The extracted information is subsequently fused using the feature fusion module (as in Figure 2), and the final output is produced after a basic classification layer.

To assess the impact of dual input channels on the classification performance, we conducted ablation experiments, which are described in Table 1. Compared to ResNet34, 50 and 101, ResNet18 was better trained on smaller datasets, demonstrating that it is not the case that the deeper the number of layers of the network, the better the results. The addition of dual channels effectively improved each classical network for all four metrics. Although the precision of MobileNet was slightly higher than that of ResNet18 after adding dual-channel input, whereas the rest of the metrics of ResNet18 were better than other networks. Our findings demonstrate that the dual channel input significantly enhances the network's ability to acquire information, resulting in improved classification performance. Specifically, the experimental results indicate that the dual channel input led to a 6.9% increase in *Accuracy*, 9.8% increase in *Precision*, 6.6% increase in *Recall* and 8.4% increase in F_1 score for ResNet18. The effectiveness of the dual channel input has been empirically validated.

*https://github.com/weiaicunzai/pytorch-cifar100
†https://github.com/xiaolai-sqlai/mobilenetv3
‡https://github.com/jaxony/ShuffleNet
§https://github.com/4uiiurz1/pytorch-res2net

Table 1. Effect of dual channel input on classification network performance.

Method	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
ResNet18* [53]	86.7	84.7	86.2	85.2
ResNet34* [53]	84.1	82.2	79.9	80.8
ResNet50* [53]	83.6	80.9	83.2	81.5
ResNet101* [53]	83.0	81.4	80.8	80.9
MobileNet† [54]	86.1	83.9	85.0	84.4
ShuffleNet‡ [55]	84.9	87.7	83.9	85.8
Res2Net§ [56]	81.3	76.7	79.6	78.1
ResNet18 + dual	**93.6**	94.5	**92.8**	**93.6**
ResNet34 + dual	91.3	94.2	87.1	89.9
ResNet50 + dual	92.0	91.7	93.3	91.8
ResNet101 + dual	90.9	94.1	89.4	90.7
MobileNet + dual	93.4	**95.1**	92.1	93.5
ShuffleNet + dual	92.7	94.3	92.1	93.2
Res2Net + dual	88.5	87.0	89.2	88.0

4.3.2. Ablation experiment for ResNet18 layers

It is easy to find from the dual-input ablation channel experiments that the performance of the network does not improve with increasing depth; as shown in Table 1, the four evaluation metrics obtained from the ResNet18 tested were higher than those for ResNet34. Therefore, in this paper, we present an ablation study on the selection of ResNet18 backbone network extraction layers for dual-channel input, the results of which are shown in Table 2. The experimental results show that the lowest metrics were selected for Conv1-Conv2_x, and each metric increased with the addition of layers Conv3_x, Conv4_x and Conv5_x. Notably, we achieved the highest *accuracy*, *Precision*, *Recall* and F_1 score of 86.7, 84.7, 86.2 and 85.2%, respectively, when employing the Conv1-Conv5_x structure as the feature extractor. Based on this finding, we utilized the Conv1-Conv5_x structure in all subsequent experiments, demonstrating the superiority of ResNet18 in dual-channel input analysis.

Table 2. Effect of the choice of layers on the performance of classification networks.

Layer_name	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
Conv1-Conv2_x	80.0	77.9	79.1	62.1
Conv1-Conv3_x	83.1	81.4	80.8	80.9
Conv1-Conv4_x	84.9	83.4	84.4	83.3
Conv1-Conv5_x	**86.7**	**84.7**	**86.2**	**85.2**

4.3.3. Ablation study of loss function λ

To avoid possible redundancy and loss of information in the fusion module, we propose to split the overall loss function into two parts; besides the loss function of the target and prediction results, the loss of the GAP operation-guided attention is added in this paper. In order to assess the efficacy of the GAP operation-guided attention loss function, we conducted a sensitivity analysis on the hyperparameter λ to investigate its effect on network performance. The impact of varying values of λ on the experimental results is illustrated in Figure 3, where λ is incremented by 0.1 from 0.1 to 0.9, with the dotted line denoting the training strategy employing the unimproved loss. Notably, the performance of attention loss guided by adding GAP operations was always better than baseline and *Accuracy* reached 98.3% when $\lambda = 0.4$, which is better than the other values chosen; also, its *Accuracy* improved by 0.6% relative to the pre-improvement loss function. The results demonstrate that the addition of the GAP operation-guided attention loss consistently outperformed the baseline and is thus a robust and effective improvement to the loss function.

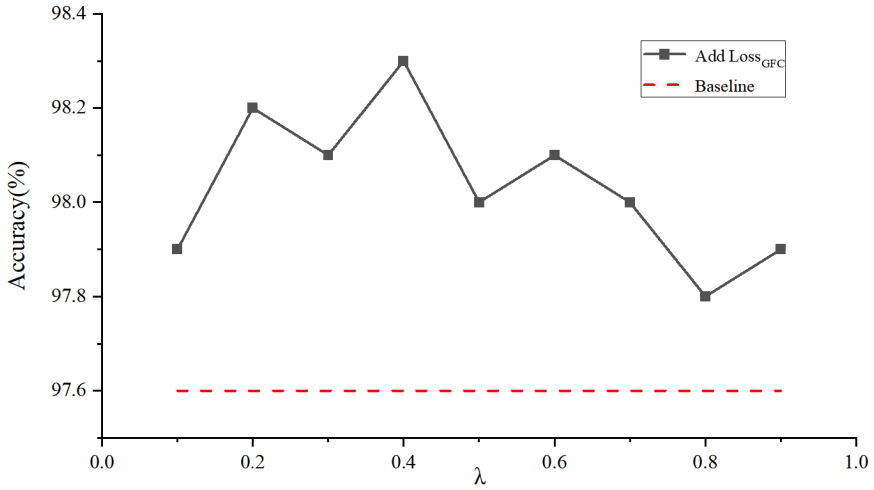

Figure 3. Effect of λ values on experimental results.

4.3.4. Ablation study of each module

In our experiments on the breast ultrasound dataset, we utilized a variety of refinement approaches, including the integration of Mobile ViT Block, dual input channels, and enhanced loss functions. A summary of our findings can be found in Table 3, with the "√" symbol indicating the adoption of these techniques in conjunction with the baseline ResNet. Since the improved loss function needs to be used with the dual-input scheme, it is not possible to add the improved loss function alone. Our results indicate that with the continuous addition of improvement schemes, the prediction results obtained by the network become more and more accurate. Compared to other improvement schemes, the performance of the network is significantly improved when the dual-channel input is added alone, further demonstrating the importance of the dual-channel input. Incorporating the improved loss function on the basis of the two-channel input, resulted in the *Accuracy*, *Precision*, *Recall* and F_1 score increasing by 1.6, 3.7, 1.2 and 2.6%, respectively, which verifies the rationality of the improved loss function. The amalgamation of all proposed techniques in this study has led to a noteworthy enhancement in the *Accuracy* of the baseline classifier to 98.3%. Additionally, we observed notable improvements in

Precision, *Recall*, and $F1$ score, which increased by 13.7, 11.9 and 13.1%, respectively, surpassing the baseline performance.

Table 3. Effect of each improvement scheme on experimental results.

	Improvement Scheme			Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
	MobileViT	Dual-input	LossGFC				
Baseline				86.7	84.7	86.2	85.2
	√			89.7	88.2	89.8	88.9
		√		93.6	94.5	92.8	93.5
	√	√		97.6	97.4	97.2	95.8
	√		√	90.3	89.2	90.3	89.7
		√	√	95.2	98.2	94	96.1
	√	√	√	**98.3**	**98.4**	**98.1**	**98.3**

4.3.5. Comparison with other methods

To validate the superior classification accuracy of the proposed algorithm, we present Table 4, which shows the results of employing four distinct evaluation metrics and compares the classification results of BUSI datasets obtained via different classifications through a five-fold cross-validation. The number before "±" in the table indicates the mean of the five-fold cross-validation, while the latter indicates the variance. It is noted that the experimental results obtained for the other networks in the table all use the same training strategy as the proposed method. From the table, it can be found that ResNet18, MobileNet, VGG16 and Res2Net obtained lower results. Swin-Transformer and ConvMixer, although further improved over the previous four methods, still have large gaps with the best results. The proposed method significantly outperformed the state-of-the-art methods in terms of all of the aforementioned evaluation metrics. Specifically, the averages of *Accuracy*, *Precision*, *Recall* and F_1; score improved by 4.0, 1.4, 4.9 and 3.1%, respectively.

Table 4. Comparison of classification results via different methods on the BUSI dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
ResNet18 [53]	86.4 ± 0.29	84.4 ± 0.31	86.2 ± 0.15	85.3 ± 0.20
MobileNet [54]	86.0 ± 0.28	83.9 ± 0.05	84.9 ± 0.08	84.4 ± 0.06
ShuffleNet [55]	84.9 ± 0.08	87.7 ± 0.05	83.9 ± 0.09	85.8 ± 0.04
Res2Net [56]	81.1 ± 0.11	76.6 ± 0.21	79.5 ± 0.12	78.0 ± 0.15
Swin-Transformer[¶] [57]	90.4 ± 0.15	89.7 ± 0.14	85.8 ± 0.18	87.7 ± 0.16
ConvMixer[‖] [58]	92.7 ± 0.08	95.6 ± 0.10	91.9 ± 0.08	93.7 ± 0.09
Conformer[**] [59]	94.1 ± 0.12	96.9 ± 0.24	93.3 ± 0.17	95.1 ± 0.20
Ours	**98.1 ± 0.08**	**98.3 ± 0.11**	**98.2 ± 0.13**	**98.2 ± 0.12**

Table 5 shows the performance of the proposed method compared to other state-of-the-art methods on the BUSC dataset. As we can see, our network has a considerable advantage over all other networks for all metrics. Compared to the traditional networks MobileNet and VGG16, ResNet18 obtained better

metrics on this dataset. However, the traditional networks all performed more poorly than the latest methods such as ConvMixer. The experiments prove that Conformer scored the highest on *Precision*, while the proposed method obtained satisfactory results on the other three metrics, with the averages of *Accuracy*, *Recall* and F_1 score improving by 0.7, 1.6 and 0.7%, respectively, relative to the advanced Conformer.

Table 5. Comparison of classification results via different methods on the BUSC dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
ResNet18 [53]	91.9 ± 0.21	91.6 ± 0.15	93.3 ± 0.17	92.4 ± 0.16
MobileNet [54]	90.1 ± 0.25	93.1 ± 0.12	88.5 ± 0.19	90.7 ± 0.15
ShuffleNet [55]	82.5 ± 0.27	82.3 ± 0.11	80.5 ± 0.18	81.4 ± 0.13
Res2Net [56]	84.3 ± 0.09	89.5 ± 0.11	80.2 ± 0.12	84.6 ± 0.11
Swin-Transformer[¶] [57]	94.5 ± 0.14	96.1 ± 0.14	93.4 ± 0.15	94.7 ± 0.14
ConvMixer[‖] [58]	95.9 ± 0.07	96.9 ± 0.09	95.1 ± 0.10	96.0 ± 0.10
Conformer[**] [59]	97.2 ± 0.08	**97.7 ± 0.05**	96.5 ± 0.09	97.1 ± 0.06
Ours	**97.9 ± 0.10**	97.5 ± 0.08	**98.1 ± 0.12**	**97.8 ± 0.10**

Due to the unbalanced categories of the BUSI and BUSC datasets, to further demonstrate the stability of the proposed method in this paper, the relationship between the true positive rate and the false positive rate for each category in the dataset was examined. In Figures 4 and 5, classes 0, 1 and 2 denote classes of benign, malignant and normal, respectively. The curves in the figure show that the proposed model, with area under the curve (AUC) values close to 1 for all categories, has a high true positive rate and a low false positive rate. It can identify each category accurately and effectively.

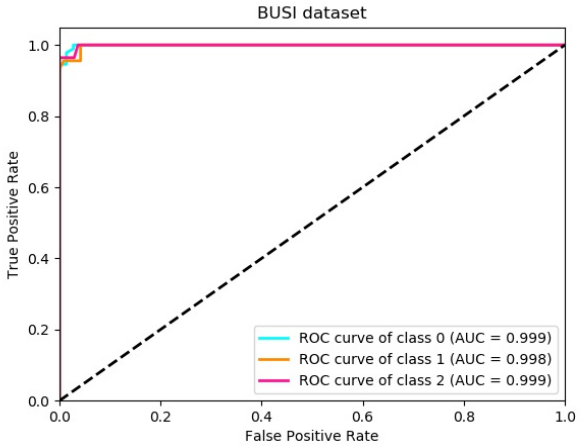

Figure 4. Receiver Operating Characteristic (ROC) Curve for BUSI.

As an additional proof of the effectiveness of our proposed method, we show the results of prediction and visualization for the BUSI and BUSC datasets in Figures 6 and 7, respectively. The results show that our method achieved satisfactory classification accuracy by effectively utilizing both edge and

[¶]https://github.com/microsoft/Swin-Transformer
[‖]https://github.com/locuslab/convmixer
[**]https://github.com/pengzhiliang/Conformer

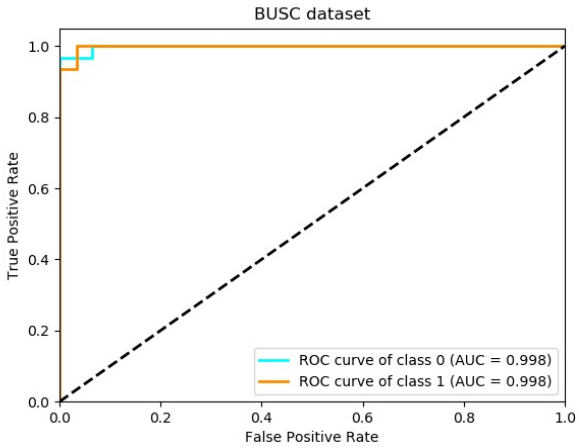

Figure 5. Receiver operating characteristic (ROC) curves for BUSC.

internal information of the target region, and then determining whether the target region is diseased or not. The visualization results can assist physicians in understanding the lesion area and avoiding deterioration due to incorrect predictions. However, it should be noted that the dual-input structure utilized in our approach requires more computational time than some of the aforementioned methods, despite its superior classification performance in breast cancer diagnosis.

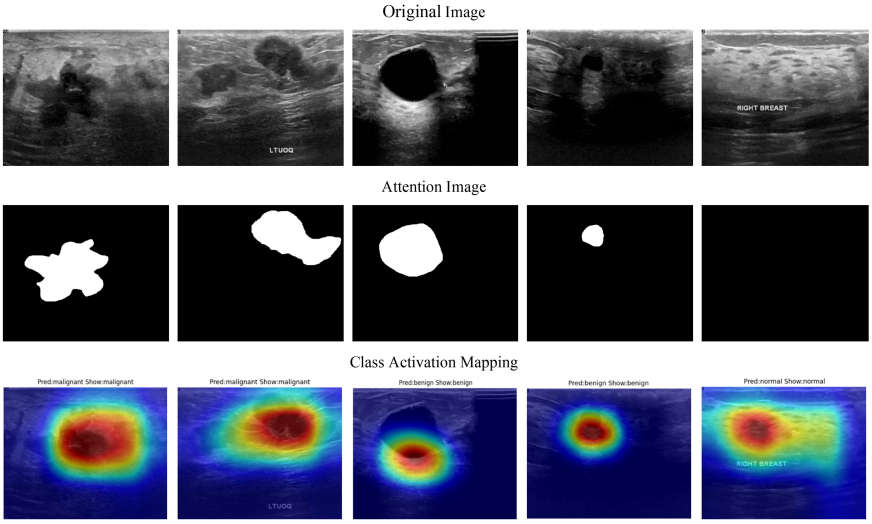

Figure 6. Visualization of the BUSI dataset results.

4.3.6. Weakness

As evidenced in the preceding experimental results, the proposed methodology exhibited remarkable proficiency in accurately predicting the class of breast tumors in the majority of cases. Nonetheless, as elucidated by the depiction of failure cases on the breast ultrasound image dataset, particularly in the presence of excessive noise interference, the neural network continues to manifest erroneous assessments (as exemplified in Figure 8). It is worth noting that this predicament constitutes a ubiquitous

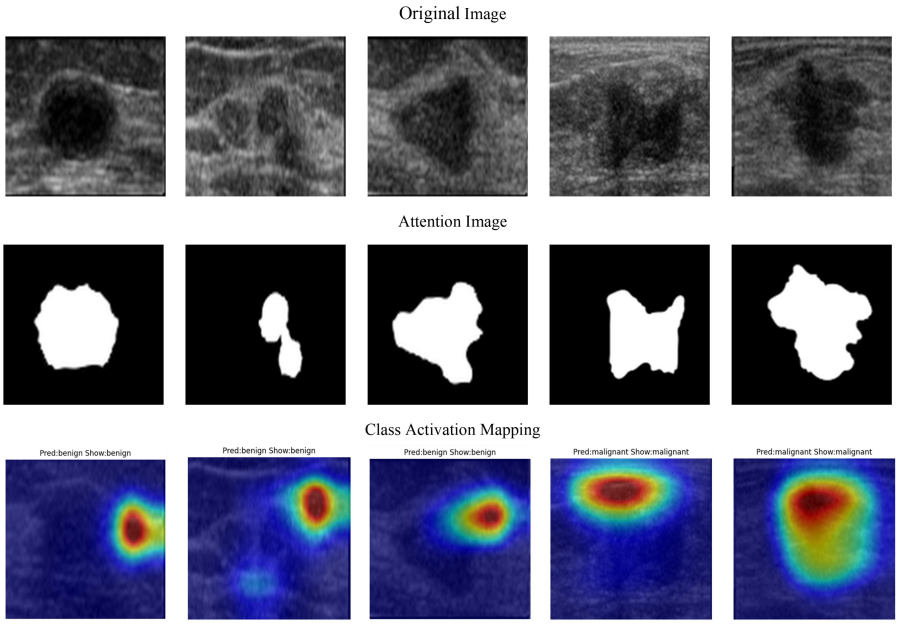

Figure 7. Visualization of the BUSC dataset results.

challenge for other contemporary state-of-the-art techniques. In future endeavors, we will continue our endeavor to enhance the network's resilience in response to noise.

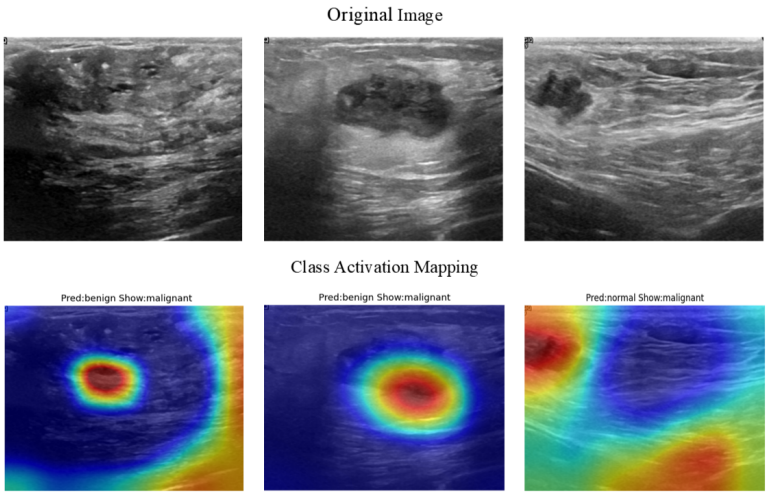

Figure 8. Illustration of the failure cases for the proposed method on the breast ultrasound image dataset.

5. Conclusions

In this paper, we propose a novel approach for breast ultrasound tumor image classification that utilizes a dual input feature fusion model. On the BUSI and the BUSC datasets, our proposed tech-

nique exhibited superior classification accuracy as compared to popular methodologies, and it offers several notable advantages: 1) The dual channel inputs employed in our methodology compensate for information loss stemming from noise and artifacts prevalent in ultrasound images. 2) The inclusion of a guided attention loss, in the form of a GAP operation during the feature fusion stage endows the network with enhanced feature learning capabilities, ultimately leading to improved classification accuracy. 3) The generated output results are visually presented, aiding medical professionals in diagnosis and the tailoring of personalized treatment plans. Notwithstanding the satisfactory classification outcomes, it is important to acknowledge that our methodology pertains to strongly supervised image classification, and is thus subject to certain limitations. In forthcoming work, we aim to synergistically combine weakly supervised learning techniques with our proposed approach, thereby augmenting its performance.

Use of AI tools declaration

The authors declare that they have not used artificial intelligence tools in the creation of this article.

Acknowledgement

The work was partly supported by the National Natural Science Foundation of China (Nos. 12274200, 61502164), the Natural Science Foundation of Hunan Province (No. 2020JJ4057), the Changsha Municipal Natural Science Foundation of China(No. kq2202239), the Key Research and Development Program of Changsha Science and Technology Bureau (No. kq2004050) and the Scientific Research Foundation of Education Department of Hunan Province of China (Nos. 21A0052, 22B0036).

Conflict of interest

The authors declare no conflict of interest.

References

1. N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, Deep neural networks improve radiologists' performance in breast cancer screening, *IEEE Trans. Med. Imaging*, **39** (2019), 1184–1194. https://doi.org/10.1109/TMI.2019.2945514

2. D. M. van der Kolk, G. H. de Bock, B. K. Leegte, M. Schaapveld, M. J. Mourits, J. de Vries, et al., Penetrance of breast cancer, ovarian cancer and contralateral breast cancer in BRCA1 and BRCA2 families: high cancer incidence at older age, *Breast Cancer Res. Treat.*, **124** (2010), 643–651. https://doi.org/10.1007/s10549-010-0805-3

3. Q. Xia, Y. Cheng, J. Hu, J. Huang, Y. Yu, H. Xie, et al., Differential diagnosis of breast cancer assisted by s-detect artificial intelligence system, *Math. Biosci. Eng.*, **18** (2021), 3680–3689. https://doi.org/10.3934/mbe.2021184

4. S. Williamson, K. Vijayakumar, V. J. Kadam, Predicting breast cancer biopsy outcomes from bi-rads findings using random forests with chi-square and mi features, *Multimedia Tools Appl.*, **81** (2022), 36869–36889. https://doi.org/10.1007/s11042-021-11114-5
5. D. J. Gavaghan, J. P. Whiteley, S. J. Chapman, J. M. Brady, P. Pathmanathan, Predicting tumor location by modeling the deformation of the breast, *IEEE Trans. Biomed. Eng.*, **55** (2008), 2471–2480. https://doi.org/10.1109/TBME.2008.925714
6. M. M. Ghiasi, S. Zendehboudi, Application of decision tree-based ensemble learning in the classification of breast cancer, *Comput. Biol. Med.*, **128** (2021), 104089. https://doi.org/10.1016/j.compbiomed.2020.104089
7. S. Liu, J. Zeng, H. Gong, H. Yang, J. Zhai, Y. Cao, et al., Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach, *Comput. Biol. Med.*, **92** (2018), 168–175. https://doi.org/10.1016/j.compbiomed.2017.11.014
8. Y. Dong, J. Wan, L. Si, Y. Meng, Y. Dong, S. Liu, et al., Deriving polarimetry feature parameters to characterize microstructural features in histological sections of breast tissues, *IEEE Trans. Biomed. Eng.*, **68** (2020), 881–892. https://doi.org/10.1109/TBME.2020.3019755
9. I. Elyasi, M. A. Pourmina, M. S. Moin, Speckle reduction in breast cancer ultrasound images by using homogeneity modified bayes shrink, *Measurement*, **91** (2016), 55–65. https://doi.org/10.1016/j.measurement.2016.05.025
10. H. H. Xu, Y. C. Gong, X. Y. Xia, D. Li, Z. Z. Yan, J. Shi, et al., Gabor-based anisotropic diffusion with lattice boltzmann method for medical ultrasound despeckling., *Math. Biosci. Eng.*, **16** (2019), 7546–7561. https://doi.org/10.3934/mbe.2019379
11. J. Levman, T. Leung, P. Causer, D. Plewes, A. L. Martel, Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines, *IEEE Trans. Biomed. Eng.*, **27** (2008), 688–696. https://doi.org/10.1109/TMI.2008.916959
12. A. Ed-daoudy, K. Maalmi, Breast cancer classification with reduced feature set using association rules and support vector machine, *Network Modeling Analysis in Health Informatics and Bioinformatics*, **9** (2020), 1–10. https://doi.org/10.1007/s13721-020-00237-8
13. R. Ranjbarzadeh, S. Dorosti, S. J. Ghoushchi, A. Caputo, E. B. Tirkolaee, S. S. Ali, et al., Breast tumor localization and segmentation using machine learning techniques: Overview of datasets, findings, and methods, *Comput. Biol. Med.*, (2022), 106443. https://doi.org/10.1016/j.compbiomed.2022.106443
14. P. Sathiyanarayanan, S. Pavithra, M. S. Saranya, M. Makeswari, Identification of breast cancer using the decision tree algorithm, in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, (2019), 1–6. https://doi.org/10.1109/ICSCAN.2019.8878757
15. J. X. Tian, J. Zhang, Breast cancer diagnosis using feature extraction and boosted c5. 0 decision tree algorithm with penalty factor, *Math. Biosci. Eng.*, **19** (2022), 2193–205. https://doi.org/10.3934/mbe.2022102
16. S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, Y. Jin, An improved random forest-based rule extraction method for breast cancer diagnosis, *Appl. Soft Comput.*, **86** (2020), 105941. https://doi.org/10.1016/j.asoc.2019.105941

17. T. Octaviani, d. Z. Rustam, Random forest for breast cancer prediction, in *AIP Conference Proceedings*, AIP Publishing LLC, **2168** (2019), 020050. https://doi.org/10.1063/1.5132477

18. S. Das, O. R. R. Aranya, N. N. Labiba, Brain tumor classification using convolutional neural network, in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, IEEE, (2019), 1–5. https://doi.org/10.1007/978-981-10-9035-6_33

19. R. Hao, K. Namdar, L. Liu, F. Khalvati, A transfer learning–based active learning framework for brain tumor classification, *Front. Artif. Intell.*, **4** (2021), 635766. https://doi.org/10.3389/frai.2021.635766

20. Q. Zhang, C. Bai, Z. Liu, L. T. Yang, H. Yu, J. Zhao, et al., A gpu-based residual network for medical image classification in smart medicine, *Inf. Sci.*, **536** (2020), 91–100. https://doi.org/10.1016/j.ins.2020.05.013

21. Y. Dai, Y. Gao, F. Liu, Transmed: Transformers advance multi-modal medical image classification, *Diagnostics*, **11** (2021), 1384. https://doi.org/10.3390/diagnostics11081384

22. S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, M. Islam, An effective skin cancer classification mechanism via medical vision transformer, *Sensors*, **22** (2022), 4008. https://doi.org/10.3390/s22114008

23. S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, et al., Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, Springer, (2021), 45–54. https://doi.org/10.1007/978-3-030-87237-3_5

24. F. Almalik, M. Yaqub, K. Nandakumar, Self-ensembling vision transformer (sevit) for robust medical image classification, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, Springer, (2022), 376–386. https://doi.org/10.1007/978-3-031-16437-8_36

25. Y. Wu, S. Qi, Y. Sun, S. Xia, Y. Yao, W. Qian, A vision transformer for emphysema classification using ct images, *Phys. Med. Biol.*, **66** (2021), 245016. https://doi.org/10.1088/1361-6560/ac3dc8

26. B. Hou, G. Kaissis, R. M. Summers, B. Kainz, Ratchet: Medical transformer for chest x-ray diagnosis and reporting, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, Springer, (2021), 293–303. https://doi.org/10.1007/978-3-030-87234-2_28

27. F. A. Spanhol, L. S. Oliveira, C. Petitjean, L. Heutte, Breast cancer histopathological image classification using convolutional neural networks, in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, (2016), 2560–2567. https://doi.org/10.1109/IJCNN.2016.7727519

28. W. Lotter, G. Sorensen, D. Cox, A multi-scale cnn and curriculum learning strategy for mammogram classification, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, (2017), 169–177. https://doi.org/10.1007/978-3-319-67558-9_20

29. A. A. Nahid, M. A. Mehrabi, Y. Kong, Histopathological breast cancer image classification by deep neural network techniques guided by local clustering, *Biomed Res. Int.*, **2018** (2018). https://doi.org/10.1155/2018/2362108

30. H. K. Mewada, A. V. Patel, M. Hassaballah, M. H. Alkinani, K. Mahant, Spectral–spatial features integrated convolution neural network for breast cancer classification, *Sensors*, **20** (2020), 4747. https://doi.org/10.3390/s20174747

31. W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, *Data Brief*, **28** (2020), 104863. https://doi.org/10.1016/j.dib.2019.104863

32. P. S. Rodrigues, Breast ultrasound image, *Mendeley Data*, **1** (2017). https://doi.org/10.17632/wmy84gzngw.1

33. J. Virmani, R. Agarwal, Deep feature extraction and classification of breast ultrasound images, *Multimedia Tools Appl.*, **79** (2020), 27257–27292. https://doi.org/10.1007/s11042-020-09337-z

34. W. Al-Dhabyani, M. Gomaa, H. Khaled, F. Aly, Deep learning approaches for data augmentation and classification of breast masses using ultrasound images, *Int. J. Adv. Comput. Sci. Appl.*, **10** (2019), 1–11. https://doi.org/10.14569/IJACSA.2019.0100579

35. N. Vigil, M. Barry, A. Amini, M. Akhloufi, X. P. Maldague, L. Ma, et al., Dual-intended deep learning model for breast cancer diagnosis in ultrasound imaging, *Cancers*, **14** (2022), 2663. https://doi.org/10.3390/cancers14112663

36. T. Xiao, L. Liu, K. Li, W. Qin, S. Yu, Z. Li, Comparison of transferred deep neural networks in ultrasonic breast masses discrimination, *Biomed Res. Int.*, **2018** (2018). https://doi.org/10.1155/2018/4605191

37. W. X. Liao, P. He, J. Hao, X. Y. Wang, R. L. Yang, D. An, et al., Automatic identification of breast ultrasound image based on supervised block-based region segmentation algorithm and features combination migration deep learning model, *IEEE J. Biomed. Health. Inf.*, **24** (2019), 984–993. https://doi.org/10.1109/JBHI.2019.2960821

38. W. K. Moon, Y. W. Lee, H. H. Ke, S. H. Lee, C. S. Huang, R. F. Chang, Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks, *Comput. Methods Programs Biomed.*, **190** (2020), 105361. https://doi.org/10.1016/j.cmpb.2020.105361

39. S. Acharya, A. Alsadoon, P. Prasad, S. Abdullah, A. Deva, Deep convolutional network for breast cancer classification: enhanced loss function (elf), *J. Supercomput.*, **76** (2020), 8548–8565. https://doi.org/10.1007/s11227-020-03157-6

40. E. Y. Kalafi, A. Jodeiri, S. K. Setarehdan, N. W. Lin, K. Rahmat, N. A. Taib, et al., Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks, *Diagnostics*, **11** (2021), 1859. https://doi.org/10.3390/diagnostics11101859

41. G. S. Tran, T. P. Nghiem, V. T. Nguyen, C. M. Luong, J. C. Burie, Improving accuracy of lung nodule classification using deep learning with focal loss, *J. Healthcare Eng.*, **2019** (2019). https://doi.org/10.1155/2019/5156416

42. L. Ma, R. Shuai, X. Ran, W. Liu, C. Ye, Combining dc-gan with resnet for blood cell image classification, *Med. Biol. Eng. Comput.*, **58** (2020), 1251–1264. https://doi.org/10.1007/s11517-020-02163-3

43. C. Zhao, R. Shuai, L. Ma, W. Liu, D. Hu, M. Wu, Dermoscopy image classification based on stylegan and densenet201, *IEEE Access*, **9** (2021), 8659–8679. https://doi.org/10.1109/ACCESS.2021.3049600

44. D. Sarwinda, R. H. Paradisa, A. Bustamam, P. Anggia, Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer, *Procedia Comput. Sci.*, **179** (2021), 423–431. https://doi.org/10.1016/j.procs.2021.01.025

45. Y. Chen, Q. Zhang, Y. Wu, B. Liu, M. Wang, Y. Lin, Fine-tuning resnet for breast cancer classification from mammography, in *Proceedings of the 2nd International Conference on Healthcare Science and Engineering 2nd*, Springer, (2019), 83–96. https://doi.org/10.1007/978-981-13-6837-0_7

46. F. Almalik, M. Yaqub, K. Nandakumar, Self-ensembling vision transformer (sevit) for robust medical image classification, in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2022*, Springer, (2022), 376–386. https://doi.org/10.1007/978-3-031-16437-8_36

47. B. Gheflati, H. Rivaz, Vision transformers for classification of breast ultrasound images, in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, (2022), 480–483. https://doi.org/10.1109/EMBC48229.2022.9871809

48. L. Yuan, X. Wei, H. Shen, L. L. Zeng, D. Hu, Multi-center brain imaging classification using a novel 3d cnn approach, *IEEE Access*, **6** (2018), 49925–49934. https://doi.org/10.1109/ACCESS.2018.2868813

49. J. Zhang, Y. Xie, Y. Xia, C. Shen, Attention residual learning for skin lesion classification, *IEEE Trans. Med. Imaging*, **38** (2019), 2092–2103. https://doi.org/10.1109/TMI.2019.2893944

50. B. Xu, J. Liu, X. Hou, B. Liu, J. Garibaldi, I. O. Ellis, et al., Attention by selection: A deep selective attention approach to breast cancer classification, *IEEE Trans. Med. Imaging*, **39** (2019), 1930–1941. https://doi.org/10.1109/TMI.2019.2962013

51. Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, *Adv. Neural Inf. Process. Syst.*, **31** (2018).

52. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 618–626. https://doi.org/10.1109/ICCV.2017.74

53. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

54. A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, et al., Searching for mobilenetv3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

55. X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 6848–6856. https://doi.org/10.1109/CVPR.2018.00716

56. S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2019), 652–662. https://doi.org/10.1109/TPAMI.2019.2938758

57. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 10012–10022.

58. A. Trockman, J. Z. Kolter, Patches are all you need?, preprint, arXiv:2201.09792. https://doi.org/10.48550/arXiv.2201.09792

59. Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, et al., Conformer: Local features coupling global representations for visual recognition, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 367–376. https://doi.org/10.1109/ICCV48922.2021.00042

AIMS Press