Received 3 June 2024, accepted 11 June 2024, date of publication 14 June 2024, date of current version 21 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3414675

RESEARCH ARTICLE

Tnc-Net: Automatic Classification for Thyroid Nodules Lesions Using Convolutional Neural Network

JUNYANG CAO[1], YANGPING ZHU[2], XING TIAN[1], AND JUAN WANG[1]

[1]School of Computer Science, China West Normal University, Nanchong, Sichuan 637009, China
[2]Department of Radiology, Nanjiang TCM Hospital, Bazhong, Sichuan 636600, China

Corresponding author: Juan Wang (wj20221213@126.com)

This work was supported by the Innovation Team Funds of China West Normal University under Grant KCXTD2022-3. This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Management Committee of Nanjiang County Traditional Chinese Medicine Hospital.

ABSTRACT Automatic and accurate classification of thyroid nodules is of great significance to doctors for clinical diagnosis and subsequent treatment recommendations. Since there are no obvious features between benign and malignant nodules, enlarging or reducing the image will result in blurred edges and image distortion, thus limiting the accuracy of clinical diagnosis. Furthermore, the prevalence of sample class imbalance in medical images poses significant challenges to applying convolutional neural networks in thyroid nodule classification methods. This paper proposes a network Tnc-Net for thyroid nodule classification. The network backbone can adapt to the problem of small data volume, capture global features with the help of simple channel attention, and effectively extract image information. The branch network supplements the feature extraction from the backbone network, and the information extracted from the backbone and branch networks is effectively utilized through the fusion module. In addition, this article designs training strategies suitable for this network to deal with category imbalance, improve model classification performance, and make classification results more clinically referenceable. The method test accuracy is 0.902, which exceeds other classic deep learning models in classification. This result demonstrates the effectiveness of our method in achieving automatic classification of thyroid nodules.

INDEX TERMS Thyroid nodule, data imbalance, convolutional neural network, classification, attention mechanism.

I. INTRODUCTION

Thyroid nodules are a manifestation of thyroid lesions, most of which are benign, but 7% are malignant [1]. Timely and accurate detection of thyroid nodules can greatly reduce patient risk and medical costs [2]. As a simple and inexpensive noninvasive examination, Doppler ultrasonography has been used as the main method of thyroid nodule examination. Clinically, doctors diagnose and classify thyroid nodules according to the size, edge, high-low echo, adjacency, blood flow, and location of thyroid nodules in imaging, to assess the risk level of thyroid lesions [3].

Figure 1 shows three different types of thyroid lesions in a thyroid Doppler ultrasound image. In Figure 1(a), the benign nodules show a regular nodule shape. In contrast, Figure 1(c) shows the characteristics of malignant nodules. Figure 1(b) shows inflammation of the thyroid, with clear boundaries between normal and diseased areas within the epithelium. However, due to the lack of fully representative characteristics of benign and malignant thyroid nodules [4], and problems such as low resolution and noise interference in ultrasound imaging [5], this is more hardship to accurately diagnose and classify thyroid nodules by Doppler ultrasound imaging alone in the clinic. This can lead to overdiagnosis and even unnecessary needle biopsies [6]. Therefore, the deep learning method can effectively extract

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

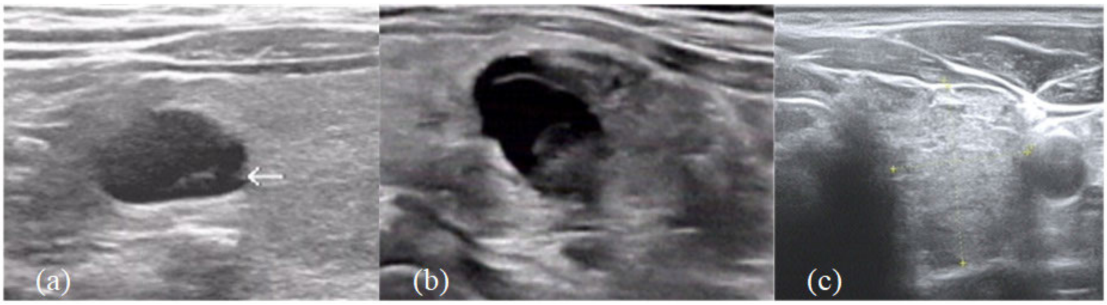


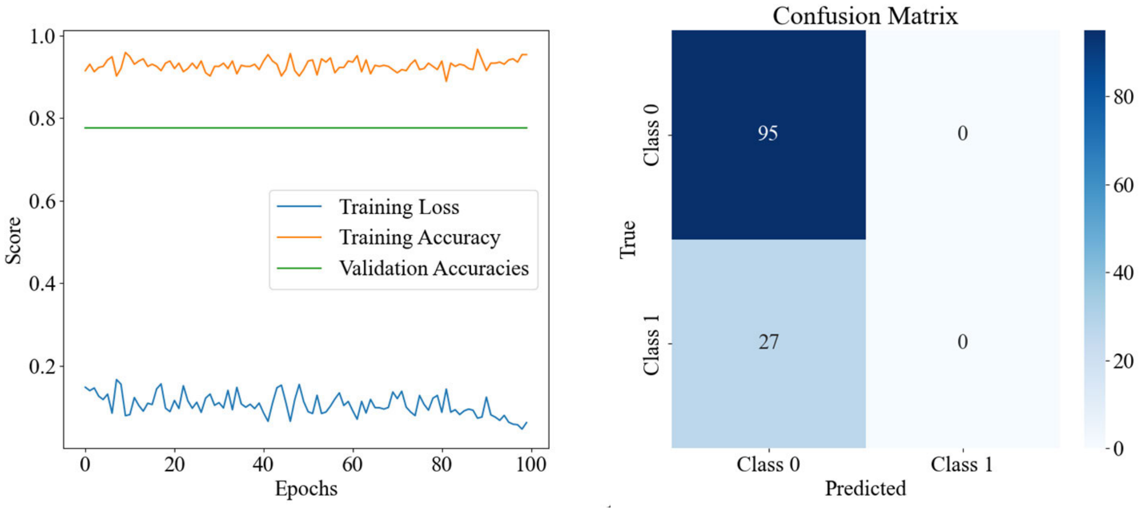
FIGURE 1. Three types of thyroid lesions: (a) benign nodules. (b) Thyroid inflammation. (c) Malignant nodules.

the features of Doppler ultrasound images, to assist doctors in diagnosing thyroid nodule types and improve diagnostic accuracy.

Thyroid nodules were clinically classified according to TI-RADS levels. When TI-RADS levels are below 2 (including 2), the thyroid is considered to have no significant lesions or benign nodules. When TI-RADS is above 4 (including 4), thyroid nodules are considered to have a higher risk of malignancy. When classified as TI-RADS 3, it indicates that the current nodules are atypical benign nodules with a risk of malignancy of less than 5%. As clinical judgment is so dependent on the doctor's experience, the judgment has a high degree of uncertainty, patients usually need to undergo multiple examinations, follow-up follow-ups, and even needle biopsy for diagnosis [7]. These tests are expensive and time-consuming; At the same time, puncture biopsy is invasive and carries greater risk than imaging diagnosis [8]. The purpose of this paper is to classify thyroid nodules to assist physicians in diagnosing thyroid nodule types more effectively and accurately, thereby reducing unnecessary follow-up and needle biopsies.

With the excellent performance of deep learning in the field of image classification, many researchers have begun to apply deep learning classification methods to the classification of thyroid nodules. Zhang et al. used CNN-based multiple cross-validation to select ultrasound images and specifically trained the binary classification module to distinguish between obscure images by adding new labels to the images and putting them back [9]. Baima et al. proposed a dense lymph node Swin-Transformer (DST) system to improve the diagnostic performance of thyroid nodules through a dense connection mechanism and full use of multilayer features [10]. Ali et al. developed a highly differentiated predictive model, called AIPs-SnTCN, to accurately predict anti-inflammatory peptides [11]. Zhao et al. proposed a local and global feature Unwrapping network (LOCH-NET) to classify benign and malignant thyroid nodules by mimicking the dual pathway structure of human vision [12]. Lim et al. evaluated the performance of artificial neural network (ANN) and binary logistic regression (BLR) in the diagnosis of benign and malignant thyroid nodules in ultrasound examination and found that it was superior to the diagnostic performance of radiologists [13]. Ma et al. suggested that a pre-trained convolutional neural network (CNN) with two different convolutional layers and a fully connected layer be fused for thyroid nodule diagnosis [14]. Chi et al. used pre-trained GoogLeNet to extract image features and classify thyroid nodules by random forest [15]. Song et al. developed a multi-task cascades Convolutional neural network (MC-CNN) framework designed to classify thyroid nodules using contextual information [16]. Shi et al. applied the knowledge-guided adversarial enhanced synthetic medical image method to thyroid nodule classification [17]. Ullah et al. developed a highly efficient antiviral peptide recognition method named DeepAVP-TPPred, which achieved accurate prediction of antiviral peptides through image feature extraction, evolutionary information integration, and deep neural network construction [18]. Peng et al. trained three branch networks on the same training set and adopted a voting mechanism to improve the model's performance in classifying thyroid nodules [19]. Akbar et al. proposed an improved deep learning model, iAFPs-MvBiTCN, which successfully predicted AFPs, providing a new reliable tool for solving the problem of fungal infections [20]. Sun et al. combined Vision-Transformer with comparative learning to propose a thyroid nodule classification model [21].

The above convolutional neural network-based methods all perform well on classification tasks. However, the inherent locality of convolution operators keeps CNNS lack focused on the fusion of local features [22]. Swin Transformer proposed by Liu et al. uses Patch Merging to replace pooling for subsampling, thereby reducing the loss of context and spatial information caused by pooling and enabling better extraction of local features [23]. Liu et al. proposed the ConvNeXt model [24], which expands the model receptive field through large nuclear convolution and removes pooling in the Block to reduce the loss of context and spatial information, thus making the model perform better in feature extraction. However, both of these methods have some limitations in global feature extraction. In the Xception network proposed by Chollet [25], separable convolution is used to reduce the complexity of the model, and a pooling layer is added to each Block to improve the performance of the model when extracting global features. However, pooling often leads to the loss of context and spatial information [26], [27], reducing the accuracy of model classification.



(a)

(b)

FIGURE 2. ConvNeXt training images. (a) ConvNeXt images for loss and accuracy during training and verification. (b) Confusion matrix for ConvNeXt validation set.

In contrast to natural images, medical images typically exhibit a constrained data volume, complicating the construction of expansive, well-annotated datasets and often encountering issues of class imbalance. Deep learning methodologies necessitate substantial data volumes for effective training, posing a significant hurdle when applied to medical image analysis with limited datasets. Moreover, the inherent data distribution in medical images is often skewed, with certain categories containing significantly more samples than others. Consequently, an imbalanced distribution of categories in the dataset input to the model predisposes the model to a phenomenon known as class imbalance, wherein minority categories are dominated by majority categories during training [28], leading the model to erroneously classify minority categories into the dominant ones to minimize training loss.

Although models may exhibit low training loss under such conditions, their performance on the validation set may suffer, indicating poor generalization, particularly for underrepresented categories. For instance, as illustrated in Figure 2(a), ConvNeXt demonstrates marginal improvements in accuracy during training while validation accuracy remains stagnant. An examination of the confusion matrix of the validation set (Figure 2(b)) reveals the model's tendency to misclassify certain classes as dominant, potentially reducing the clinical reference value of its predicted results. Such occurrences are commonplace in deep learning models, particularly in domains like medical imaging characterized by data imbalance.

Consequently, addressing class imbalance becomes paramount when dealing with medical image data sets. Mitigating strategies may involve leveraging ensemble learning techniques like the random forest method [29]. However, in scenarios with scant data and imbalanced categories, a singular approach often proves inadequate. Thus, practical applications necessitate tailored preprocessing techniques or parameter adjustments to optimize model performance following specific dataset characteristics. Therefore, we propose a novel training strategy and a new backbone branch network to mitigate the effects of class imbalance [30] while fully extracting global and local information from images. Specifically, we first divide the image into a training set, a validation set, and a test set, and resampling the training set to ensure that the number of benign and malignant nodule images is roughly equal. However, excessive use of raw images for random data augmentation can lead to the generation of inappropriate pairs, making the training set too homogeneous and reducing the model's generalization ability. To further balance the disparity between the two classes, we employ a weighted loss function to make the model more attentive to the less frequent classes.

To fully extract global and local features from images, we design a backbone branch network. We introduce a simplified channel attention mechanism into the Xception network to facilitate information exchange between channels, enhance feature representation, and improve the model's robustness. To prevent overfitting during training, we reduce the complexity of the model by decreasing the depth and width of the intermediate flow in Xception. The branch network employs the ConvNeXt model to augment the features extracted from the backbone network. In this model, a large 7 × 7 convolutional kernel is introduced to achieve a larger receptive field, and the pooling layer is replaced to avoid feature loss associated with pooling. Additionally, we introduce a feature aggregation module to fully utilize the features extracted by the backbone network and branch network, ultimately achieving accurate diagnosis of benign and malignant thyroid nodules. Our contributions are summarized below:

1. We proposed a thyroid nodule classification model based on the complementary feature extraction of the trunk branch network and used the feature fusion module to effectively integrate these features, to achieve the accurate diagnosis of benign and malignant thyroid nodules.
2. A new training strategy is proposed to reduce the effects of class imbalance by resampling images and using a weighted loss function.
3. Compared to other advanced deep models, our Tnc-Net performs better on small, unevenly categorized ultrasonic image datasets.

The rest of this article is organized as follows. Section II reviews the work. Section III introduces data sets, preprocessing, model structure, and training strategies. The experimental setup and results are given in Section IV. In Section V, the advantages and limitations of this model are discussed. Section VI is the conclusion.

II. RELATED WORK

Medical image classification has always been an important research topic to assist clinical diagnosis. However, limited data set size and unbalanced categories have always been problems in medical image classification. This paper reviews the latest research work related to medical image classification and the treatment of data imbalance.

A. MEDICAL IMAGE CLASSIFICATION BASED ON CONVOLUTIONAL NEURAL NETWORK

With the excellent performance of convolutional neural networks in the field of natural images, many medical image researchers have begun to pay attention to the use of deep learning methods to process medical images. Zhu et al. used the pre-trained residual network for transfer learning to classify thyroid after data set enhancement, proving the effectiveness of convolutional enhancement networks and the application of transfer learning in medical images [31]. Jiang et al. combined the residual-full convolution transformer (Res-FCT) model and the Residual-channel attention module (Res-CAM) to propose an HT-RCM ultrasonic image classification model for Hashimoto thyroiditis [32]. Akbar et al. proposed a new computational model, deepstacking - avp, to accurately distinguish AVPs and successfully identify antiviral peptides, providing an efficient tool for drug design and research [33]. Shakeel et al. used regularization to enhance the performance of the basic model, evaluated the model with isolated data, and achieved an accurate classification of thyroid nodules [34]. Huang et al. realized the classification of thyroid nodules by sampling video frames at the same time interval, capturing different semantic information, and using two different feature extraction branches to extract local and global features [35]. Akbar et al. used three different classification models to evaluate prediction rates, using individual and heterogeneous vectors respectively. Then the prediction label of a single classifier is used by genetic algorithm to integrate the depth model, which enhances the prediction and training ability of the proposed model. Good results have been obtained in the identification of anti-tuberculosis peptides [36].To enable the model to capture both local details and global dependencies, Zhao et al. designed a new FCST model that extracts local information by residual convolution and compensates for missing information of local details by Swin Transformer, which is used to extract deep features of ultrasonic nodule images [37]. Anissa et al. developed a method to help experts define component characteristics [38]. The modified VGG16 classifies the pre-processed data into four categories: cystic, solid, complex, and spongy to help the doctor or specialist identify the characteristics of the nodules.

B. METHODS TO DEAL WITH THE IMBALANCE OF MEDICAL IMAGE CATEGORIES

Due to the uneven distribution of medical image data, a few categories are often dominated by the majority. Therefore, to make the model cope with this problem, many researchers focus on transfer learning, ensemble learning, and so on. Park and Ha used the Hadoop framework to efficiently process and analyze large traffic data and adopted a sampling method to solve the problem of data imbalance [39]. Chen et al. proposed a joint prototype alignment (CPA) loss to address the class imbalance problem to promote balanced optimization of the FL framework to eliminate unbalanced gaps with such balance objectives [40]. Guo et al. developed a globally optimal label fusion algorithm to deal with the problem that MRI segmentation performance usually deteriorates when training with small data sets and sparse annotations [41]. Zhao et al. proposed self-integrated dual-course learning in feature space to enhance feature distribution through feature distillation and feature reweighting, thereby improving the problem caused by data imbalance in glaucoma diagnosis [42]. Suk et al. combined sparse regression with deep learning and proposed a deep integrated sparse regression network to solve the problem of scarcity of training samples [43]. Pizarro et al. addressed the problem of MRI class imbalance by expanding the dataset size and the proportion of non-artificial data instances [44]. Therefore, we first expand the number of images of a few classes in the training set by off-line data enhancement and make the model pay more attention to the fewer classes by weighted loss, to alleviate the influence of class imbalance. We use the two-stream network to extract the global and local features of the image and then use the feature fusion module to make full use of the global and local features extracted by the model, reduce the training cost of the model, and improve its robustness and generalization.

III. METHOD

To fully extract image features and alleviate the impact of category imbalance, we proposed the Tnc-Net model. The model consists of three parts, including the backbone network, branch network and feature aggregation module. The backbone network SP-LeNet of Tnc-Net. The modified network uses separable convolution to reduce model

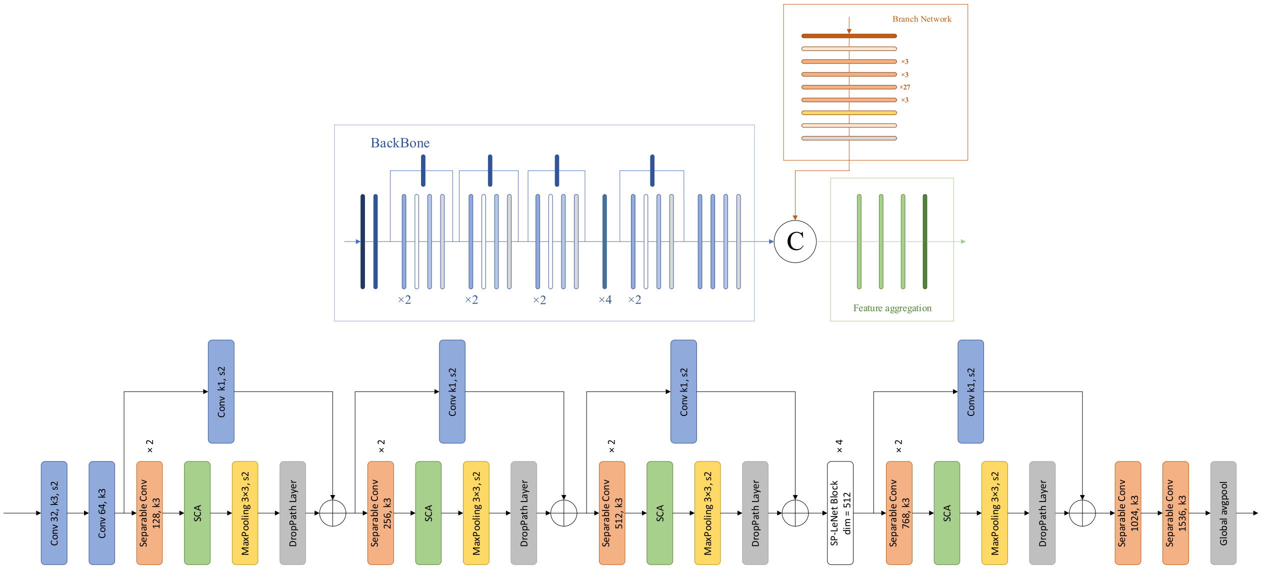


FIGURE 3. Architecture of Tnc-Net model, '⊕', '©' denote addition, concatenation.

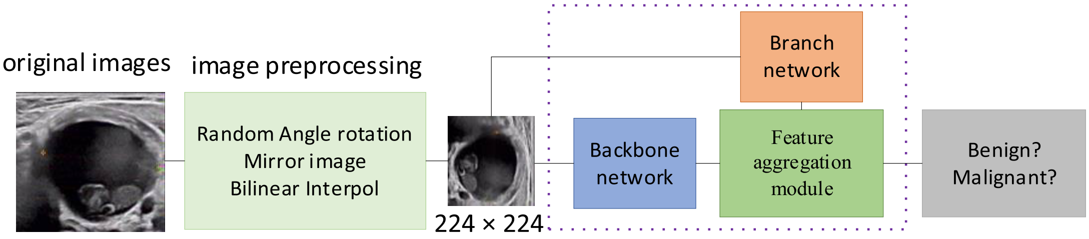


FIGURE 4. Main working flow chart.

complexity. Its pooling layer can better extract global features in images and introduces simple channel attention [45] to capture global information. As a result, the model was overfitted. We reduced the model complexity and introduced the drop_path layer to prevent the model from overfitting and improve the model's generalization ability. The branch network is ConvNeXt, which can effectively extract local features in the image and supplement the features extracted by the backbone network Sp-LeNet. To make full use of the features extracted from the backbone network and branch networks, we designed a feature aggregation module. Compared with direct splicing, the training cost is lower, and it has better robustness and generalization. The flowchart and model structure diagram are shown in Figure 3.

In the training process, we adopted the strategy of weighting the categories to further alleviate the problem of category imbalance, to achieve the best performance of the model. The workflow of the method is shown in Figure 4.

A. BACKBONE NETWORKS

To allow the neural network to better extract image features, we have made changes to Xception. Specifically, we introduce Simplified Channel Attention to realize information exchange between channels. Channel attention controls the correlation between different channels by weighting the channel dimensions of the feature map. It can be expressed as:

$$CA(X) = X \times \sigma(W_2 \max(0, W_1 pool(X))) \tag{1}$$

The X in the expression represents the feature map, and the pool represents the global average pooling operation. σ is the nonlinear activation function Sigmoid, and W1 and W2 are fully connected layers. The ReLU activation function is used between fully connected layers. * Is a channel-wise operation. By retaining the two most important roles of channel attention, namely aggregating global information and channel information interaction [45], a simplified channel attention mechanism (SCA) is obtained to realize information exchange between channels. Simple channel attention (SCA) is defined by (2):

$$SCA(X) = X \times Wpool(X) \tag{2}$$

To prevent model overfitting due to small data sets, we reduced the model complexity, reduced the number of middle flows in Xception, and changed the number of channels to 512.

Due to the modification of the number of channels of the middle flow. Therefore, the number of channels of the four separable convolutions of the exit flow is also changed to 512, 768, 1024, and 1536 respectively. In addition, the activation function is changed to GeLU [46], and a drop_path layer [47] is introduced in each Block. We named the improved network Sp-LeNet. The Block of Sp-LeNet is shown in Figure 5(b). The branch network uses the ConvNeXt model to supplement the features extracted from the backbone network. This model introduces a 7 × 7 large convolution kernel to obtain a larger receptive field and replaces the pooling layer to avoid feature loss caused by this layer. The Block of ConvNeXt is shown in Figure 5(a). In the training phase, we removed the last fully connected layer of ConvNeXt and Sp-LeNet and input the thyroid nodule data set into the backbone network and branch network respectively to extract features.

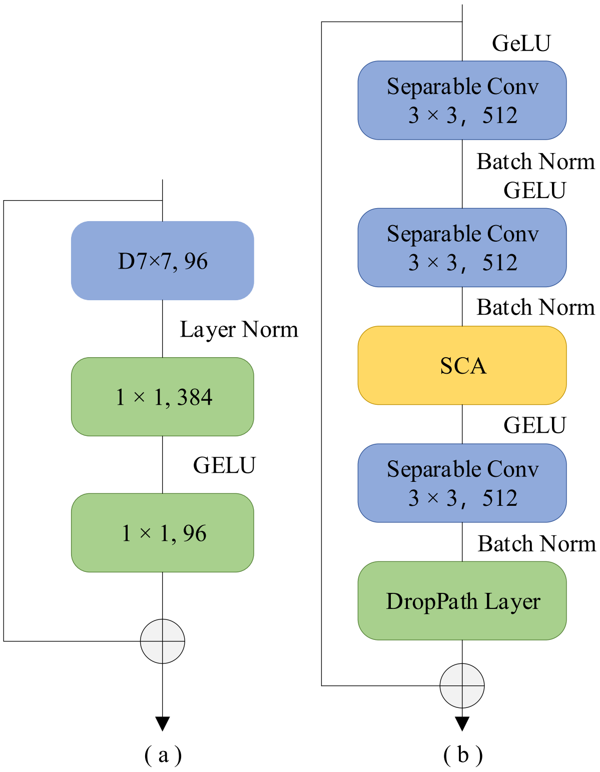


FIGURE 5. blocks of the network. (a)ConvNeXt block. (b) block block of SP-LeNet.

B. FEATURE FUSION MODULE

To fully extract and utilize the information in the image, Wang et al. [48] used the powerful representation ability of multi-layer neural networks to perform feature fusion and achieved good performance in the overall classification of pests. Wu et al. [49] used a network composed of two complementary streams to automatically integrate dual-view contextual lesion information to extract global features and local features for esophageal lesion classification. However, this method directly performs feature splicing, which leads to a significant increase in the number of channels and thus higher training costs. Peng et al. [50] used a parallel structure to fuse global and local images at different resolutions in an interactive manner, but this structure also means high training costs. Feedforward neural networks can enhance the expression ability of features through activation functions [51] to extract richer semantic information. A feature aggregation module is designed through a feedforward neural network to complement the features extracted by the branch network with the features extracted by the backbone network. Compared with existing combination models, the feature aggregation module can make full use of features extracted from the backbone network and branch networks without significantly increasing the number of channels. The feature aggregation module consists of four parts, including a feature merging module, a layer of downsampling, a feature feedforward network, and a classification layer. The feature aggregation module is shown in Figure 6.

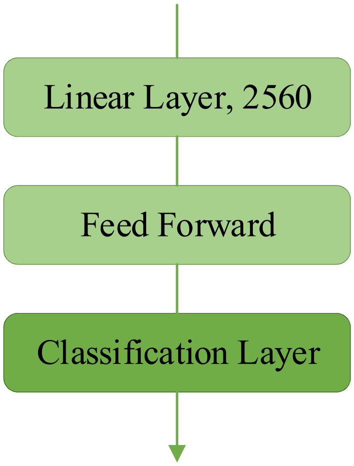


FIGURE 6. Feature aggregation module. The spliced features are downsampled, input into the feedforward neural network to further integrate the features, and finally connected to the classification layer to output the results.

Among them, the first linear is used to downsampling, filter out the repeated features extracted from the two backbone networks, and reduce the model training cost. The feedforward neural network improves the expression ability of features through nonlinear transformation, and then extracts richer semantic information and learns more complex deep features.

At the same time, the drop_path layer is introduced to help prevent overfitting of the model. The subsequent ablation experiments also verify that the design of the module is effective and reasonable. The nonlinear layer of the feedforward neural network uses the Gaussian error linear unit (GELU) [46] as the activation function, and the activation function is formulated as follows:

$$GeLU(X) = X \times \frac{1}{2}[1 + erf(\frac{x}{\sqrt{2}})] \tag{3}$$

where erf is the error function defined by (4):

$$\tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)) \tag{4}$$

C. MODEL TRAINING STRATEGY

Due to the scarcity and imbalance of medical images, overfitting, poor generalization, and the majority of classes dominated by a few classes often occur in the training of deep models. Thus, we designed a training strategy for Tnc-Net to mitigate the effects of data imbalance. In the pre-processing stage, the number of malignant nodules in the training set was expanded by off-line data enhancement. However, overuse of raw images for random enhancement can result in the generation of inappropriate pairs, making the training set too similar. To further balance the ratio of benign and malignant nodules during training, we used the weighted loss to make the model pay more attention to the fewer classes. When capturing images, first calculate the proportion of the number of classes in each round of training, so that the weights of the classes are inversely proportional to their numbers and apply these weights to the loss function. The pseudo-code is shown below:

Algorithm 1 Framework of ensemble learning for our system.

```
Input: Training set S = s1, s2 ..., sN, C represents the class
Output: Weighted loss function F;
1: Calculate the number of samples S for each class
2: The number of samples S is inversely proportional as the sample weight W for each class;
3: WeightedRandomSampler assigns weights Wn to each sample according to W;
4: The sample weights Wn are passed in when the DataLoader is created;
5: Create a loss function F and enter the weights Wn into the function
6: return F;
```

During the training, set the learning rate to 1e-4, batch_size to 16, and epoch to 100. The model is evaluated by a 50 percent verification method, and the weight with the highest accuracy of model verification is saved for prediction.

IV. EXPERIMENTS AND RESULTS

A. DATASET AND PRE-PROCESSING

In this study, 606 thyroid Doppler ultrasound images from 537 patients were retrospectively studied. All images are from Nanjiang TCM Hospital, Bazhong City, Sichuan Province. Most of the ultrasonic probes used are high-frequency linear array probes with a frequency of 7-12MHZ. For images containing abnormally enlarged goiter, a convex array probe was used. The data set contained 473 benign nodules and 133 malignant nodules. Among them, the level of TI-RADS below 2 and the level of partial TI-RADS is 3 are classified as benign nodules, while the level of partial TI-RADS is 3 and the level of TI-RADS is 4 and above are classified as malignant nodules. The diagnosis of TI-RADS level 3 and above is based on pathological biopsy. The dataset is split into three parts: 386 training sets (including 302 benign nodules and 84 malignant nodules), 98 validation sets, and 122 test sets. To reduce the influence of class imbalance, offline data enhancement is used to expand the training set. Detailed statistics of thyroid images collected are shown in Table 1.

TABLE 1. Thyroid image data distribution.

	Train	Validation	Test	Total
Benign	302	76	95	473
Malignant	84	22	27	133

TABLE 2. The hyperparameters of the Tnc-net.

Parameter	Value
epochs	100
optimizer	AdamW
learning rate	1e-5
drop path rate	0.4
batch size	8
loss function	cross-entropy loss

In clinical practice, doctors judge thyroid nodules based on factors such as nodule size, edge condition, high and low echo, and location. Therefore, in the experiment, the number of malignant nodules in the training set was expanded to 252 through random Angle rotation and mirror operation. To maintain the consistency of the input model, bilinear interpolation is used to adjust the image size to 224 × 224, and the image is normalized. The data preprocessing process is shown in Figure 7.

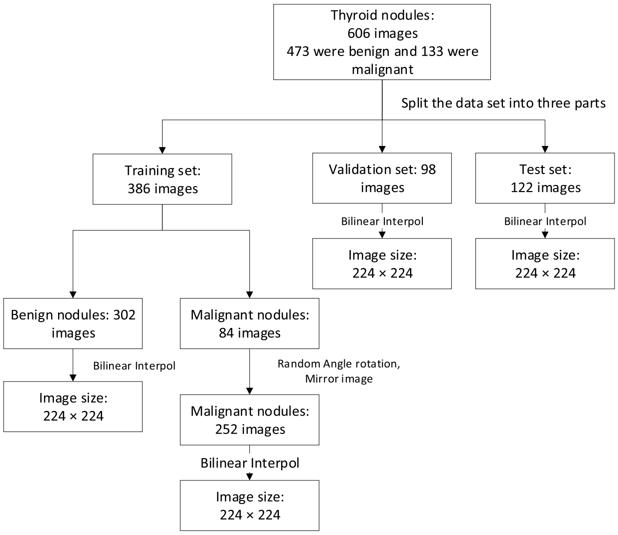


FIGURE 7. Divides the image into a training set, a verification set, and a test set. Off-line data enhancement was performed for malignant nodules in the training set, including random Angle rotation and mirroring, to reduce the influence of class imbalance. The size of all images is adjusted to 224 × 224 using bilinear interpolation to ensure consistency of the input model.

B. TRAINING SETTINGS

The computer platform configuration is as follows: The CPU is InterI CoreI i3-12100f 3.3GHz four core eight threads; The

GPU is NVIDIA GeForce RTX 3060 12GB, and the memory size is 8G x 2. All code is written under Python 3.9, using PyTorch [52] as a deep learning library, and CUDA version 12.0. The dataset was divided into three parts: 554 images (offline data enhancement) for training sets, 98 images for validation sets, and 122 images for test sets. The input size is 224 × 224. The hyperparameter values employed in this investigation are presented in Table 2. The basic learning rate is 1e-5, using the AdamW optimizer, and weighted cross-entropy loss function, and the batch size is 8. All models are trained through 100 iterations. In training, the weight parameters with the highest verification accuracy are saved for testing. The accuracy and loss of training are shown in Figure 8.

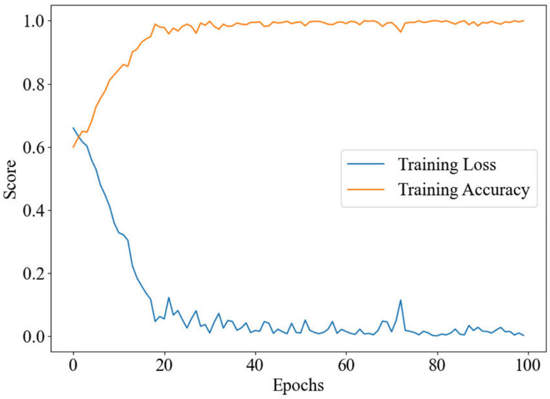


FIGURE 8. Accuracy and loss graphs for model training.

C. EVALUATION METRICS

To objectively evaluate our proposed models, we calculate the following indicators for analysis and comparison. Accuracy: The proportion of the total number of correctly classified samples to the total number of samples is the most used index to evaluate the classification performance. However, in the presence of class imbalance, accuracy is easily affected, so other indicators are needed for collaborative analysis. The precision is also called the precision ratio, and the study area is the sample predicted to be positive, indicating the proportion of samples predicted correctly in the study area. Sensitivity, also known as recall rate, refers to the proportion of true positive samples that are correctly predicted as positive samples to the total true positive samples. Specificity is the proportion of the number of negative examples correctly classified to the total number of negative examples. The F-score is a harmonic average of precision and recall rates, considering the precision and sensitivity of the model. When F1 is higher, the test method is more effective. Accuracy, precision, sensitivity, specificity, and F1-score are defined by (5) - (8):

$$accuracy = \frac{TP + FN}{TP + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$specificity = \frac{TN}{FP + TN} \tag{8}$$

- TP, true positive: The real class is positive, and the predicted result is also a positive class.
- FP, false positive: The true class is negative, and the predicted result is positive.
- TN, The true class is negative, and the predicted result is also a negative class.
- FN, false negative: The true class is positive, and the predicted result is negative.

D. RESULTS

We evaluated our proposed Tnc-Net and other classical classification methods as well as advanced deep learning models on the same dataset, These include AlexNet [53], VGG-19 [54], ResNet34 [55], Xception [25], EfficientNet-B3 [56], Vision-Transformer [57](ViT), Swin-Transformer [23](ST), ConvNeXt [24], FocalNet [58], UniFormer [59], SGFormer [60], BiFormer [61]. The training strategy proposed by us is used to save the model with the highest verification accuracy for testing. Table 2 lists the evaluation indicator values for each model. Tnc-Net had the highest accuracy, precision, sensitivity, and F1 scores, but lacked specificity compared with EfficientNet-B3. It can be seen from the data in the table that when other classical networks make predictions, the results are biased toward the majority class and the minority class are dominated by the majority class. For example, EfficientNet_b3 has a sensitivity of only 0.222. Except for AlexNet and BiFormer, other classification methods are less than half accurate in predicting a small number of classes. Even BiFormer, which performs well on natural images, is inferior in every respect to our proposed model. In contrast, our proposed Tnc-Net can classify a few classes more accurately, which proves that our proposed method can effectively mitigate the impact of class imbalance.

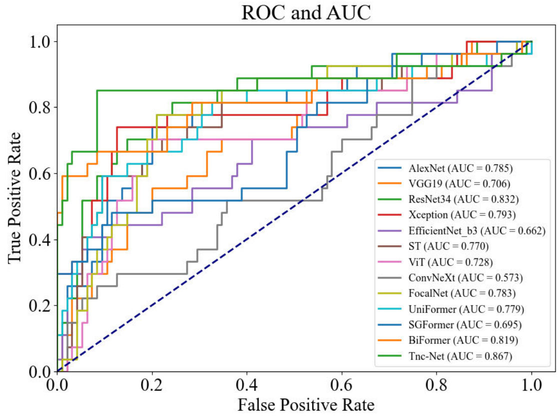


FIGURE 9. AUC diagram of Tnc-Net and other classical classifi-cation methods proposed in this paper.

TABLE 3. Comparison results between Tnc-Net proposed in this paper and other classical classification methods.

Model	Accuracy	Precision	Sensitivity	Specificity	F1_score	AUC
AlexNet	0.795	0.536	0.556	0.863	0.545	0.785
VGG19	0.762	0.455	0.370	0.874	0.408	0.706
ResNet34	0.803	0.667	0.222	0.968	0.333	0.832
Xception	0.820	0.632	0.444	0.926	0.522	0.793
EfficientNet_b3	0.811	**0.750**	0.222	**0.979**	0.343	0.662
ST	0.811	0.591	0.481	0.905	0.531	0.770
ViT	0.779	0.500	0.444	0.874	0.471	0.728
ConvNeXt	0.680	0.286	0.296	0.789	0.291	0.573
FocalNet	0.770	0.481	0.481	0.853	0.481	0.783
UniFormer	0.811	0.591	0.481	0.905	0.531	0.779
SGFormer	0.787	0.522	0.444	0.884	0.480	0.695
BiFormer	0.836	0.621	0.667	0.884	0.643	0.819
Tnc-Net	**0.902**	**0.742**	**0.852**	0.916	**0.793**	**0.867**

TABLE 4. Ablation experiments of the feature extraction module and feature aggregation module.

Model	Accuracy	Precision	Sensitivity	Specificity	F1_score	AUC
SP-LeNet	0.844	0.700	0.519	**0.937**	0.596	0.812
Xception	0.820	0.632	0.444	0.926	0.522	0.793
ConvNeXt	0.680	0.286	0.296	0.789	0.291	0.573
No-fusion_block	0.869	0.704	0.704	0.916	0.704	0.839
Tnc-Net	**0.902**	**0.742**	**0.852**	0.916	**0.793**	**0.867**

The AUC-ROC curve of Tnc-Net and other classical classification models is shown in Figure 9. The Area under the curve (AUC), that is, the area under the ROC curve, is represented by FPR (False Positive Rate) as the horizontal axis and TPR (True Positive Rate) as the vertical axis [62]. The closer the ROC curve is to the point (0,1), the better the discriminant ability of the model. Therefore, the AUC is also an important indicator to evaluate the model. Compared with other classical networks, Tnc-Net still performs best on the AUC, indicating that our method can efficiently and accurately classify thyroid nodules.

To evaluate the performance of each module in Tnc-Net, we conducted ablation studies on the feature extraction module and feature aggregation module respectively. ConvNeXt, Xception, and Sp-LeNet are all single-path architectures. As can be seen from Table 3, the single-path network performs extremely poorly when faced with minority class classification. Except for SP-LeNet, the accuracy of the other networks for minority class classification is less than half. The dual-path network also performs better when faced with a small number of category scores, which shows that global and local features can complement each other, alleviate the impact of category imbalance, and achieve better classification performance.

Compared with Xception, Sp-LeNet has a smaller network structure, can realize information exchange between channels through SCA, and has a better extraction effect of global features. No-fusion_block removes the feature fusion module in Tnc-Net and uses a direct splicing method for fusion. As can be seen from the experimental data in Table 3, Tnc-Net is better than No-fusion_block in all indicators, which shows that our proposed model has better robustness and generalization and can achieve thyroid nodule automatic classification.

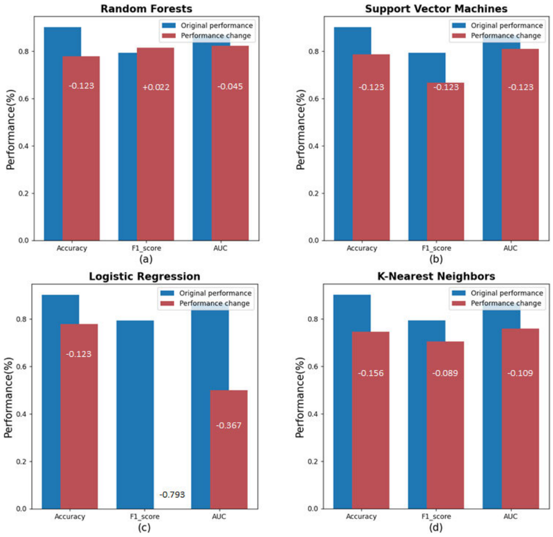


FIGURE 10. Performance differences under different classifiers. (a) Random Forests. (b) Support Vector Machines. (c) Logistic Regression. (d) K-Nearest Neighbors.

We also explore the performance difference when replacing SoftMax with a traditional classifier, as shown in Figure 10. The logistic regression classifier has a situation where the minority class is completely dominated by the majority class, and the f1-score is 0. When predicting the test set, all images are classified as the majority class. The F1-score of the random forest classifier is slightly higher than SoftMax. In future experiments, optimization of the random forest classifier can be considered. Other classifiers are completely inferior to SoftMax.

V. DISCUSSION

In this study, we aimed to make an efficient and accurate determination of the type of thyroid nodules. Due to the low definition of Doppler ultrasound images, image enlargement or reduction will lead to blurred edges and image distortion, and there is a lack of representative features between benign and malignant nodules. The classification of benign and malignant nodules remains a challenging issue, especially for TI-RADS type 3. Early studies were conducted on manual features for classification, but their performance was limited by the performance capabilities of manual features. With the rise of deep learning, data-driven methods have attracted the attention of medical imaging researchers. Compared with traditional methods, convolutional neural networks can extract deeper and richer semantic information from images during the training process, which makes it more competitive than traditional manual feature-based recognition schemes.

Fully extracting and utilizing the global and local features extracted by convolutional neural networks will help improve the model's classification of benign and malignant thyroid nodules. Most techniques are usually single-path architectures that cannot fully extract image features. Training strategies suitable for natural images are also difficult to apply directly to medical images. The model we propose can fully mine image information and use backbone networks and branch networks to extract features to complement each other. The feature aggregation module can enhance the expression ability of features, extract richer semantic information, and reduce model training costs. Faced with the common problems of small data volume and imbalanced categories in medical images, we use offline data enhancement and weighted loss to balance the gap between data volumes, reduce the impact caused by category imbalance, and improve the performance of model classification. Experiments show that the method we proposed can assist doctors in the clinical diagnosis and classification of thyroid nodules accurately and efficiently.

However, this study has certain limitations. Firstly, collaboration was confined to a single hospital, resulting in our ability to collect only 606 images for model training, validation, and testing. This also prevents us from doing multi-center experiments. Secondly, thyroid nodules can be further classified based on TI-RADS, while thyroid disorders also include conditions such as hyperthyroidism and Hashimoto's thyroiditis. A more precise classification of thyroid nodules would benefit physicians in devising subsequent treatment plans. In the future, we intend to collaborate with other hospitals to expand the dataset size and refine dataset categories to achieve more accurate classification of thyroid disorders, thereby aiding physicians in more precisely diagnosing thyroid nodules in clinical practice.

VI. CONCLUSION

The classification of benign and malignant thyroid nodules is still a challenging task due to the problems of low resolution, noise interference, lack of representative features, and unbalanced categories. We propose a new framework for trunk branch network modeling and design training strategies suitable for this framework to accurately identify thyroid nodules. Tnc-Net can not only fully extract and utilize global and local features, but also alleviate the problem of class imbalance. The designed training strategy can better cope with the imbalance of medical image categories and enhance the performance of the model in thyroid nodule classification. Extensive experimental analysis has demonstrated the superior performance of the algorithm, with the Tnc-Net model achieving an accuracy of 0.902 on the test set. This shows that our model can perform well in processing small and unevenly classified ultrasonic image datasets. However, this study has certain limitations. Currently, our collaboration is restricted to a single hospital, which precludes multi-center comparisons. Moreover, the TI-RADS criteria could be further refined. Future research will involve collaborating with additional hospitals to expand the dataset and refine classification standards, thereby enabling more precise classification of thyroid nodules. This effort is anticipated to provide valuable insights for subsequent treatment strategies. We foresee that the Tnc-Net model will play a significant role in aiding physicians in the diagnosis of benign and malignant thyroid nodules.

DATA AVAILABILITY

The raw data analyzed in the study are available from the corresponding author on reasonable request.

REFERENCES

[1] A. E. Ebeed, M. A. E.-H. Romeih, M. M. Refat, and N. M. Salah, "Role of ultrasound, color Doppler, elastography and micropure imaging in differentiation between benign and malignant thyroid nodules," *Egyptian J. Radiol. Nucl. Med.*, vol. 48, no. 3, pp. 603–610, Sep. 2017.

[2] G. Naga Sujini and S. Balakrishna, "Machine learning based computer aided diagnosis models for thyroid nodule detection and classification: A comprehensive survey," in *Proc. 2nd Int. Conf. Autom., Comput. Renew. Syst. (ICACRS)*, Dec. 2023, pp. 1283–1287.

[3] J. Sun, C. Li, Z. Lu, M. He, T. Zhao, X. Li, L. Gao, K. Xie, T. Lin, J. Sui, Q. Xi, F. Zhang, and X. Ni, "TNSNet: Thyroid nodule segmentation in ultrasound imaging using soft shape supervision," *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106600.

[4] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 919–923.

[5] B. Wildman-Tobriner, M. Buda, J. K. Hoang, W. D. Middleton, D. Thayer, R. G. Short, F. N. Tessler, and M. A. Mazurowski, "Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: Diagnostic accuracy and utility," *Radiology*, vol. 292, no. 1, pp. 112–119, Jul. 2019.

[6] M. Buda, B. Wildman-Tobriner, J. K. Hoang, D. Thayer, F. N. Tessler, W. D. Middleton, and M. A. Mazurowski, "Management of thyroid nodules seen on US images: Deep learning may match performance of radiologists," *Radiology*, vol. 292, no. 3, pp. 695–701, Sep. 2019.

[7] A. Persichetti, E. Di Stasio, R. Guglielmi, G. Bizzarri, S. Taccogna, I. Misischi, F. Graziano, L. Petrucci, A. Bianchini, and E. Papini, "Predictive value of malignancy of thyroid nodule ultrasound classification systems: A prospective study," *J. Clin. Endocrinol. Metabolism*, vol. 103, no. 4, pp. 1359–1368, Apr. 2018.

[8] D. T. Nguyen, T. D. Pham, G. Batchuluun, H. S. Yoon, and K. R. Park, "Artificial intelligence-based thyroid nodule classification using information from spatial and frequency domains," *J. Clin. Med.*, vol. 8, no. 11, p. 1976, Nov. 2019.

[9] H. Zhang, C. Zhao, L. Guo, X. Li, Y. Luo, J. Lu, and H. Xu, "Diagnosis of thyroid nodules in ultrasound images using two combined classification modules," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Suzhou, China, Oct. 2019, pp. 1–5.

[10] N. Baima, T. Wang, C.-K. Zhao, S. Chen, C. Zhao, and B. Lei, "Dense Swin transformer for classification of thyroid nodules," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–4.

[11] A. Raza, J. Uddin, A. Almuhaimeed, S. Akbar, Q. Zou, and A. Ahmad, "AIPs-SnTCN: Predicting anti-inflammatory peptides using fasttext and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks," *J. Chem. Inf. Model.*, vol. 63, no. 21, pp. 6537–6554, Nov. 2023.

[12] S.-X. Zhao, Y. Chen, K.-F. Yang, Y. Luo, B.-Y. Ma, and Y.-J. Li, "A local and global feature disentangled network: Toward classification of benign-malignant thyroid nodules from ultrasound image," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1497–1509, Jun. 2022.

[13] K. J. Lim, C. S. Choi, D. Y. Yoon, S. K. Chang, K. K. Kim, H. Han, S. S. Kim, J. Lee, and Y. H. Jeon, "Computer-aided diagnosis for the differentiation of malignant from benign thyroid nodules on ultrasonography," *Academic Radiol.*, vol. 15, no. 7, pp. 853–858, Jul. 2008.

[14] J. Ma, F. Wu, J. Zhu, D. Xu, and D. Kong, "A pre-trained convolutional neural network based method for thyroid nodule diagnosis," *Ultrasonics*, vol. 73, pp. 221–230, Jan. 2017.

[15] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *J. Digit. Imag.*, vol. 30, no. 4, pp. 477–486, Aug. 2017.

[16] W. Song, S. Li, J. Liu, H. Qin, B. Zhang, S. Zhang, and A. Hao, "Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1215–1224, May 2019.

[17] G. Shi, J. Wang, Y. Qiang, X. Yang, J. Zhao, R. Hao, W. Yang, Q. Du, and N. G.-F. Kazihise, "Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105611.

[18] M. Ullah, S. Akbar, A. Raza, and Q. Zou, "DeepAVP-TPPred: Identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm," *Bioinformatics*, vol. 40, no. 5, May 2024.

[19] S. Peng, Y. Liu, W. Lv, L. Liu, Q. Zhou, H. Yang, J. Ren, G. Liu, X. Wang, X. Zhang, and Q. Du, "Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: A multicentre diagnostic study," *Lancet Digit. Health*, vol. 3, no. 4, pp. 250–259, Apr. 2021.

[20] S. Akbar, Q. Zou, A. Raza, and F. K. Alarfaj, "IAFPs-Mv-BiTCN: Predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks," *Artif. Intell. Med.*, vol. 151, May 2024, Art. no. 102860.

[21] J. Sun, B. Wu, T. Zhao, L. Gao, K. Xie, T. Lin, J. Sui, X. Li, X. Wu, and X. Ni, "Classification for thyroid nodule using ViT with contrastive learning in ultrasound images," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106444.

[22] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102357.

[23] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002.

[24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11966–11976.

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.

[26] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," 2021, *arXiv:2106.16031*.

[27] P. Tang, Q. Liang, X. Yan, D. Zhang, G. Coppola, and W. Sun, "Multi-proportion channel ensemble model for retinal vessel segmentation," *Comput. Biol. Med.*, vol. 111, Aug. 2019, Art. no. 103352.

[28] X. Pan, J. Cheng, F. Hou, R. Lan, C. Lu, L. Li, Z. Feng, H. Wang, C. Liang, Z. Liu, X. Chen, C. Han, and Z. Liu, "SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102867.

[29] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Montreal, QC, Canada, 1995, pp. 278–282.

[30] C. Gan, Q. Feng, and Z. Zhang, "Scalable multi-channel dilated CNN–BiLSTM model with attention mechanism for Chinese textual sentiment analysis," *Future Gener. Comput. Syst.*, vol. 118, pp. 297–309, May 2021.

[31] Y. Zhu, Z. Fu, and J. Fei, "An image augmentation method using convolutional network for thyroid nodule classification by transfer learning," in *Proc. 3rd IEEE Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2017, pp. 1819–1823.

[32] W. Jiang, K. Chen, Z. Liang, T. Luo, G. Yue, Z. Zhao, W. Song, L. Zhao, and J. Wen, "HT-RCM: Hashimoto's thyroiditis ultrasound image classification model based on res-FCT and res-CAM," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 2, pp. 941–951, Feb. 2024.

[33] S. Akbar, A. Raza, and Q. Zou, "Deepstacked-AVPs: Predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model," *BMC Bioinf.*, vol. 25, no. 1, p. 102, Mar. 2024.

[34] M. Faizan Shakeel, M. Hasan Khan, and Y. Uzzaman Khan, "Deep learning empowering diagnosis of thyroid nodule malignancy through ultrasound imaging," in *Proc. Int. Conf. Recent Adv. Sci. Eng. Technol. (ICRASET)*, Nov. 2023, pp. 1–6.

[35] J. Huang, T. Chen, W. Jiang, H. Zhang, and R. Wang, "Thyroid nodule classification in ultrasound videos by combining 3D CNN and video transformer," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2023, pp. 5273–5278.

[36] S. Akbar, A. Raza, T. Al Shloul, A. Ahmad, A. Saeed, Y. Y. Ghadi, O. Mamyrbayev, and E. Tag-Eldin, "PAtbP-EnC: Identifying anti-tubercular peptides using multi-feature representation and genetic algorithm-based deep ensemble model," *IEEE Access*, vol. 11, pp. 137099–137114, 2023.

[37] X. Zhao, H. Li, J. Xu, and J. Wu, "Ultrasonic thyroid nodule benign-malignant classification with multi-level features fusions," in *Proc. 8th Int. Conf. Image, Vis. Comput. (ICIVC)*, Dalian, China, Jul. 2023, pp. 907–912.

[38] T. Anissa, H. A. Nugroho, and I. Soesanti, "Improved VGG-16 for classifying thyroid nodule on thyroid ultrasound images," in *Proc. Int. Conf. Comput. Sci., Inf. Technol. Eng. (ICCoSITE)*, Jakarta, Indonesia, Feb. 2023, pp. 95–99.

[39] S. H. Park and Y. G. Ha, "Large imbalance data classification based on MapReduce for traffic accident prediction," in *Proc. 8th Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput.*, Birmingham, U.K., Jul. 2014, pp. 45–49.

[40] Z. Chen, C. Yang, M. Zhu, Z. Peng, and Y. Yuan, "Personalized retrogress-resilient federated learning toward imbalanced medical data," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3663–3674, Dec. 2022.

[41] F. Guo, M. Ng, G. Kuling, and G. Wright, "Cardiac MRI segmentation with sparse annotations: Ensembling deep learning uncertainty and shape priors," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102532.

[42] R. Zhao, X. Chen, Z. Chen, and S. Li, "Diagnosing glaucoma on imbalanced data with self-ensemble dual-curriculum learning," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102295.

[43] H.-I. Suk, S.-W. Lee, and D. Shen, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Med. Image Anal.*, vol. 37, pp. 101–113, Apr. 2017.

[44] R. Pizarro, H.-E. Assemlal, S. K. B. Jegathambal, T. Jubault, S. Antel, D. Arnold, and A. Shmuel, "Deep learning, data ramping, and uncertainty estimation for detecting artifacts in large, imbalanced databases of MRI images," *Med. Image Anal.*, vol. 90, Dec. 2023, Art. no. 102942.

[45] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. Eur. Conf. Comput. Vis.*, vol. 13667. Tel Aviv, Israel, 2022.

[46] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[47] G. Larsson, M. Maire, and G. Shakhnarovich, "FractalNet: Ultra-deep neural networks without residuals," 2016, *arXiv:1605.07648*.

[48] C. Wang, J. Zhang, J. He, W. Luo, X. Yuan, and L. Gu, "A two-stream network with complementary feature fusion for pest image classification," *Eng. Appl. Artif. Intell.*, vol. 124, Sep. 2023, Art. no. 106563.

[49] Z. Wu, R. Ge, M. Wen, G. Liu, Y. Chen, P. Zhang, X. He, J. Hua, L. Luo, and S. Li, "ELNet: Automatic classification and segmentation for esophageal lesions using convolutional neural network," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101838.

[50] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, and Q. Ye, "Conformer: Local features coupling global representations for recognition and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9454–9468, Aug. 2023.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[52] P. Adam et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Buc, E. Fox, and R. Garnett, Eds. New York, NY, USA: Curran Associates, 2019, pp. 8024–8035.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[56] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[58] J. Yang, C. Li, X. Dai, L. Yuan, and J. Gao, "Focal modulation networks," 2022, *arXiv:2203.11926*.

[59] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023.

[60] L. Weng, K. Pang, M. Xia, H. Lin, M. Qian, and C. Zhu, "Sgformer: A local and global features coupling network for semantic segmentation of land cover," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6812–6824, 2023.

[61] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 10323–10333.

[62] K. Liu, W. Wang, R. Wang, X. Cui, L. Zhang, X. Yuan, and X. Li, "CDF-LS: Contrastive network for emphasizing feature differences with fusing long- and short-term interest features," *Appl. Sci.*, vol. 13, no. 13, p. 7627, Jun. 2023.

JUNYANG CAO received the B.S. degree in software engineering from the College of Engineering and Technical, Chengdu University of Technology, in 2022. He is currently pursuing the master's degree in software engineering with the School of Computer Science, China West Normal University. His research interest includes classification and segmentation of medical images.

YANGPING ZHU received the degree from Chengdu Medical College, in 2023. He currently works with the Department of Radiology, Nanjiang TCM Hospital. He studied in West China Hospital, Sichuan University, in 2019, and Affiliated Hospital of North Sichuan Medical College, in 2022. He is mainly engaged in clinical research and practice of medical imaging technology.

XING TIAN received the bachelor's degree in software engineering from the School of Computer Science, China West Normal University, in 2022, where he is currently pursuing the master's degree in software engineering. His research interest includes classification and segmentation of medical images.

JUAN WANG received the Ph.D. degree in earth exploration and information technology from Chengdu University of Technology, in 2015. She is currently a Professor with the School of Computer Science, China West Normal University. Her current research interests include image processing.