# Research on Medical Image Classification Based on Triple Fusion Attention

Y. G. Wang, L. Wang and Y. X. Geng

*Abstract*—**Attention mechanism is very important in the task of medical image classification. In medical image classification, different types of images may have different lesion morphology and size, and lesion characteristics are not obvious; however, the existing attention mechanism has the problem of insufficient feature diversity and ignoring small lesions, which seriously affects the classification performance. In order to solve these problems, Triple Fusion Attention (TFA) was proposed. Through convolutional fusion, attention fusion and adaptive fusion, TFA improves the model's perception ability of subtle structures and features in medical images, suppresses the noise in the image, and enhances the representation of key features, which can effectively solve the problem of insufficient sensitivity of important features in medical images. Experiments have shown that TFA enables the model to focus on the lesion area more accurately, so that multi-scale features can be fused and classified, which significantly improves the overall performance and outperforms other attention mechanisms. In addition, TFA is able to improve training efficiency and ease of deployment while maintaining good performance, while improving the accuracy and effectiveness of computer-aided diagnosis.**

*Index Terms*—**Medical Image Classification, Attention Mechanism, Feature Fusion, Triple Fusion Attention.**

## I. Introduction

SIGNIFICANT advancements have been made in the discipline in recent years due to the ongoing development of deep learning technology and the expansion of computing capacity for classifying medical images [1]. Convolutional neural networks (CNN) are one of the most popular model architectures, and deep neural networks (DNN) are particularly good at classifying medical images [2]. CNN can effectively extract characteristics from medical images and categorize them using fully linked layers by utilizing multi-layer convolution and pooling procedures [3]. When this strategy is successfully implemented, it offers strong instruments and assistance for medical research and diagnostics.

The attention mechanism is crucial in deep learning because it directs convolutional neural networks to extract pertinent elements while suppressing uncorrelated ones, growing in significance as a plug-and-play element of convolutional network models [4]. Currently, the widely utilized SE (Squeeze-and-Excitation) attention technique [5] mostly uses global average pooling [6] to compress the channel characteristics via the fully linked layer. The global average pooling process, however, is unable to completely account for the variations in different positions within the feature map and loses positional information. In addition to the SE attention method, the CBAM (Convolutional Block Attention Module) [7] introduces the maximum pooling operation [8] to overcome this problem. This improves attention performance by synthesizing the maximum feature information. For small-sized lesions in medical image classification tasks, the CBAM attention mechanism may perform poorly in terms of detection and classification because the maximum pooling operation concentrates more on the global maximum eigenvalue [9]. Furthermore, CBAM limits attention performance by primarily focusing on channel feature adjustment and failing to completely account for feature changes at other locations in the image. By incorporating location information into features and activating them through completely connected operations, CA (Coordinated Attention) [10] gets around these restrictions and improves the attention mechanism's efficacy by making greater use of location information. The SK (Selective Kernel) [11] attention mechanism processes features of different scales using a dynamic selection mechanism and parallel branching, enabling the network to adaptively select receptive fields of different sizes to model features. This further enhances the SE attention mechanism's capacity to model features at different scales. Convolutional kernels of various sizes are used to extract features, and the dynamic selection module automatically chooses the optimum feature combination. By decreasing the number of channels in the fully linked layer and introducing spatial attention based on the SE mechanism, the Bottleneck Attention Module (BAM) [12] addresses the issue of an excessive number of parameters in the SE attention module.

The attention mechanism in medical image classification can lessen the influence of redundant information and interfering elements, providing interpretable classification results [13]. By applying the attention mechanism, the model can better understand and utilize the image information for medical image classification [14]. However, existing attention mechanisms suffer from insufficient feature diversity and neglect small lesions, which seriously affects classification performance [15].

Unlike natural colour images, most medical images are greyscale images, and the lesion area is usually less contrasted with adjacent normal tissue. As shown in Figure 1, chest X-rays with lung disease are not significantly different from healthy chest X-rays, and areas with lung disease show very low contrast to surrounding normal tissue. As a result, it is difficult to extract features as diverse as those extracted from natural colour images. Detail in medical images is also important when critical information in medical images, such as lesion areas, often occupy far fewer pixels than normal tissue, and lesion areas differ in detail from normal

Manuscript received Aug 4, 2024; revised Nov 28, 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 71472081.

Y. G. Wang is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: wyigeng@163.com).

L. Wang is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: wangli9966@ustl.edu.cn).

Y. X. Geng is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: gen9yanx1n@ustl.edu.cn).

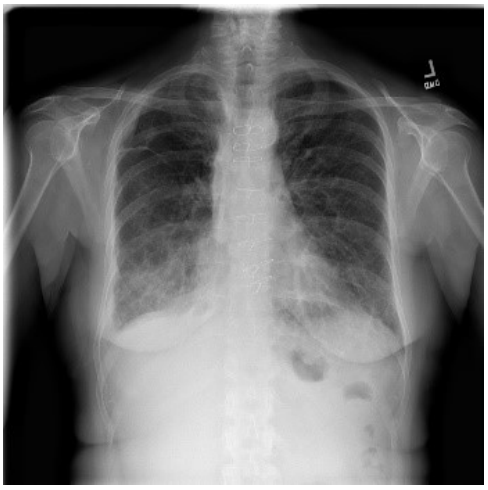

Fig. 1. Chest X-ray image of lung disease

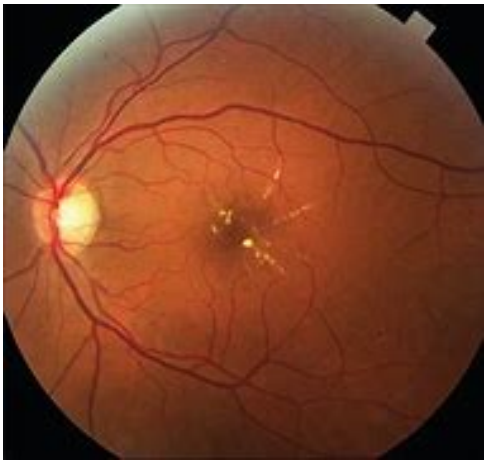

Fig. 2. Picture of DME eye disease

tissues. An image of a DME (diabetic macular oedema) eye disease is shown in Figure 2. The lesion area is too small and inconspicuous compared to normal tissue. Because small lesions are omitted in the high-level feature map, it is difficult to learn discriminant features from small, non-obvious lesions.

## II. Related Work

### A. Spatial attention

Spatial attention [16] is a technique used in the field of computer vision and deep learning to process image data by simulating the attention mechanism of the human visual system. Traditional neural networks usually process images through fully connected layers, which cannot effectively utilize the spatial information between pixels in the image. Therefore, the spatial attention mechanism is introduced to enable the model to better focus on important regions and structures in the image, improving its ability to understand and classify image information. Through spatial attention, the model can assign different attention weights based on the importance of various positions in the image, more effectively capturing and classifying image features, thereby enhancing the accuracy and efficiency of visual tasks [17].

The significance of spatial attention is to improve the comprehension and generalization ability of deep learning models. By fully utilizing the spatial relationships between pixels in an image, it helps the model better understand the structure and content of the image, thereby enhancing the model's generalization ability across different scenes and datasets. The expression formula for the spatial attention mechanism is shown in Equation 1, where X is the input image data, $W_{spatial}$ is the spatial attention weight, and $F_{spatial}$ is the feature representation obtained after processing by the spatial attention mechanism. Sigma is an activation function, usually ReLU or Sigmod.

$$F_{spatial}=\sigma(W_{spatial}\cdot X) \tag{1}$$

In medical image classification, spatial attention can enhance the model's ability to recognize detailed anatomical structures, such as distinguishing different parts of the heart or identifying the direction of blood vessels. Additionally, spatial attention improves the model's ability to process noise, increasing the accuracy and robustness of classification. It enhances the model's detection of edge and contour information in medical images, helping to more accurately locate and classify structures and abnormalities, such as tumors or abnormal tissues. This provides doctors with more precise diagnostic and treatment options.

At present, spatial attention has made remarkable research progress in many fields. Various forms of spatial attention mechanisms have been proposed and applied to image classification, object detection, semantic segmentation, and other domains, yielding impressive results. With the ongoing development of deep learning technology, spatial attention remains an active research area, and its potential to enhance model performance and practical applications continues to be explored. The spatial attention structure is illustrated in Figure 3.

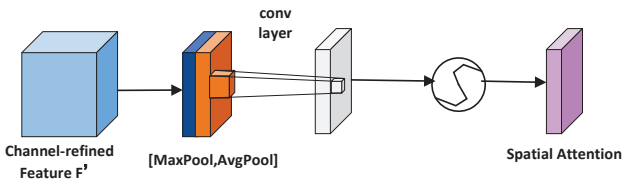


Fig. 3. Schematic diagram of spatial attention structure

### B. Channel Attention

Channel attention[18] is a key technology in deep learning that enhances a model's ability to understand and classify input data by automatically learning and assigning channel weights. Traditional neural networks process multi-channel data relatively simply and fail to fully utilize the information differences between different channels. With the advancement of deep learning technology, channel attention has become an important approach to address this issue. Its aim is to enable the model to automatically focus on the channel information most useful for a particular task, thereby improving the model's representation and adaptability. The introduction of channel attention significantly enhances the model's ability to understand and generalize input data, improving its robustness and resistance to interference. This results in better performance in tasks such as image classification, object detection, and semantic segmentation[19].

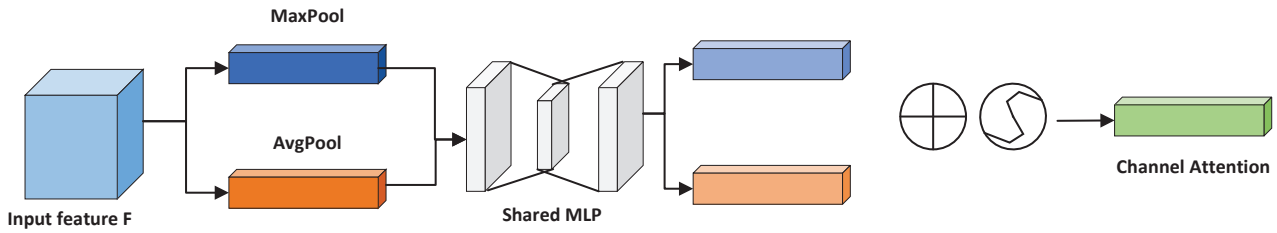


Fig. 4. Schematic diagram of channel attention structure

The channel attention mechanism can be represented by the following equation 2, where $X_i$ is the i-th channel in the input feature map, $W_i$ is the attention weight of the i-th channel, $C$ is the number of channels, $F_{channel}$ is the feature representation obtained after processing by the channel attention mechanism, and sigma is the activation function, usually ReLU or Sigmod.

$$F_{channel}=\sigma(\sum_i c = 1^{1/c} \cdot W_i \cdot X_i) \tag{2}$$

In medical image classification, the application of the channel attention mechanism can help the model better understand the features of medical images and improve classification accuracy. By reducing redundant information and highlighting key features, channel attention can optimize the performance of models in medical image classification tasks, thereby assisting doctors in diagnosing diseases more quickly and accurately.

In current research, the channel attention mechanism has been widely studied and applied. Various forms of channel attention modules, such as the Squeeze-and-Excitation (SE) module and the Channel Attention (CA) module, have been proposed and applied to tasks such as image classification, object detection, and semantic segmentation. These modules have yielded remarkable results and progress. The channel attention structure diagram is shown in Figure 4.

## III. Triple Fusion Attention

To more accurately focus on the lesion area for the fusion and classification of multi-scale features, this paper proposes a new attention method called TFA (Triple Fusion Attention). TFA is designed for diversity and discriminative feature learning in medical image classification, aiming to solve two key problems: insufficient feature diversity and the neglect of small lesions. This method achieves significant performance improvements without adding additional parameters to the model.

To efficiently collect multi-scale features in medical pictures, TFA uses numerous convolutional layers with varying kernel sizes and group counts. This feature extraction method enables models to comprehend and process both minor alterations (small lesions) and larger structures (tissue architecture) in medical pictures. First, the module uses multi-scale convolutional layers to record visual features, with changing kernel sizes and group numbers, in order to extract rich feature information at various scales. Following the capture of multi-scale data, TFA uses the Convolutional Block Attention Module (CBAM) to focus on critical image areas and features. The CBAM module is applied independently to the feature group derived from each convolutional path, finding and emphasizing information-rich channels through the channel attention mechanism and focusing on critical spatial locations in the image via the spatial attention mechanism. This improves feature representation and efficiently combines multi-scale data.

Furthermore, TFA includes adaptive convolutional layers that can dynamically alter the weights of the convolutional kernels based on the content of the feature map. This enables the model to more easily adapt to medical images with changing resolutions and content changes, increasing the model's robustness and generalization ability while also improving its performance in processing complex medical images. Before the output layer, TFA includes a residual connection. This approach preserves significant information from the original features while also assisting in the effective propagation of the gradient during training, hence improving the model's training efficiency and performance. Residual connections facilitate a smooth passage of information by connecting inputs to outputs, preventing vanishing gradient difficulties, and allowing for more efficient feature learning. Figure 5 shows the general structure of TFA.

The input feature map has dimensions (B, C, H, W), where B is the batch size, C is the number of channels, and H and W are the height and breadth. TFA uses many convolutional paths to capture features at varying sizes. To obtain the feature map $F_i$, the kernel size $k_i$ and the number of groups $g_i$ are used for each convolutional path $i$, respectively. $F_i$ denotes the feature map derived from the i-th convolution layer. Equation 3 illustrates how feature extraction is achieved.

$$F_i = Conv(x, k_i, g_i) \tag{3}$$

The CBAM output is pooled using global average pooling and normalized with Softmax to achieve a dynamic weighted fusion of multiscale feature maps ($A_i$). The Softmax function normalizes the values of each channel into an attention vector $A_i$, which is then used to weight the feature map $F_i$ from the convolutional path. Each convolutional path's features are weighted and summed based on the attention vector $A_i$, resulting in weighted features that synthesize information from many convolution paths. Equation 4 shows the formula for the weighted features.

$$F_{weighted} = \sum_i^N = l(F_i \odot A_i) \tag{4}$$

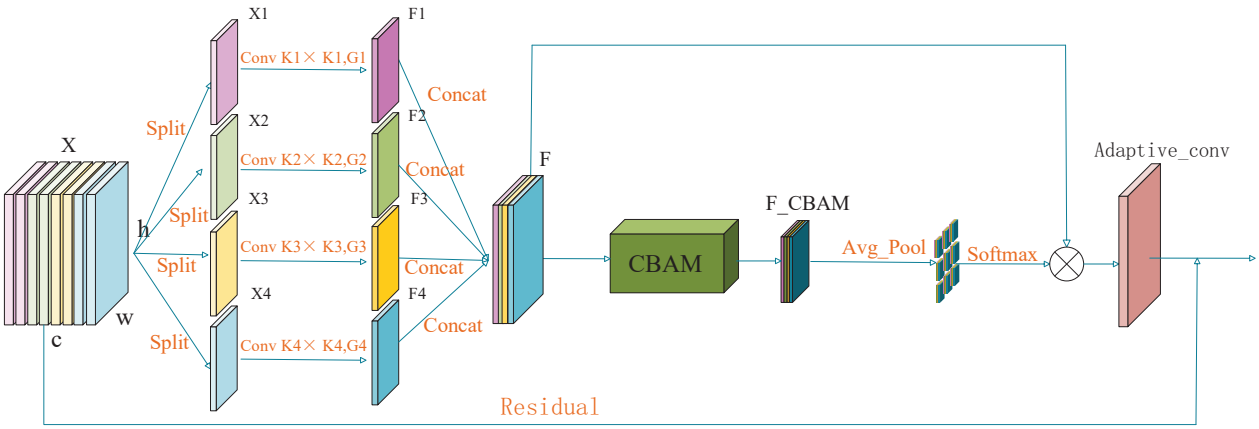


Fig. 5. Schematic diagram of TFA structure

The adaptive convolutional layer optimizes the weighted features by constantly altering the weights of the convolution kernels to improve the feature extraction process. The weighted fused feature is transformed into the final output feature using an adaptive convolutional layer. Equation 5 shows the formula for the output feature.

$$Output = AdaptiveConv(Fweighted) \tag{5}$$

In TFA, $1 \times 1$ convolution is frequently used to modify the number of channels or dimensions of the feature map to suit the model structure or the needs of a specific task. This module adjusts the shape of the input feature x to match the shape of the output feature created by the weighted feature. The residuals from the $1 \times 1$ convolution change the size and number of channels of the input features. After $1 \times 1$ convolution, the residuals and weighted features are added element by element to get the final output. This residual connection aids in the maintenance and propagation of key information in the input features, as well as the effective promotion of gradient propagation, which is beneficial to the model's optimization and convergence during training. Equations 6 and 7 represent this process.

$$Residual = Conv1 \times 1(x) \tag{6}$$

$$Output = Output + Residual \tag{7}$$

TFA is important in medical image classification since it uses multipath convolution and the CBAM technique to capture pictures' minutiae as well as broad trends. This strategy improves the model's capacity to find lesion sites, increasing classification accuracy and reliability. This is critical for jobs that demand attention to both subtle changes in small lesions and overall structure. The addition of CBAM enables the module to focus on critical image regions and features while applying channel and spatial attention to multi-scale features simultaneously. This capacity aids in gathering lesion information at diverse scales, increasing the model's ability to recognise various lesion forms and sizes and, eventually, improving the model's accuracy in medical image classification.

Medical images frequently have varying resolutions and content fluctuations, and TFA's adaptive convolution technology enables it to adapt to these changes, improving the model's stability and generalisation ability. The residual connection allows the gradient to be backpropagated directly through the path across layers, alleviating the gradient vanishing problem. It also effectively combines the original key features and the enhanced features, allowing the deeper network to learn features more efficiently and optimising the model's training effect and performance. TFA operates well in diverse types and qualities of medical images, adapts to medical images under different imaging equipment and situations, and considerably improves its application value and dependability in clinical practice.

## IV. Experimental Results

### A. Experimental datasets

The Kvasir v1, Kvasir v2, HAM10000 and Brain Tumor MRI datasets were used in this experiment. The Kvasir datasets are primarily utilized for digestive tract image analysis and are divided into two versions, v1 and v2, each containing 8 categories. Both the v1 and v2 datasets were randomly split into training and validation sets with a ratio of 8:2. The Kvasir v1 dataset, developed by researchers at University Hospitals in Norway, includes more than 4,000 gastroscopic images taken in clinical settings. These images cover a variety of pathological conditions, lesion types, and lesion severity. The Kvasir v2 dataset, created by the University of Bergen in collaboration with major Norwegian hospitals, is an upgraded version designed specifically for image classification tasks related to digestive tract diseases. It contains approximately 8,000 high-definition gastrointestinal endoscopic images, derived from procedures such as gastroscopy and colonoscopy, and annotated and classified by professional doctors. To promote research in medical image analysis and computer-aided diagnosis, the Kvasir datasets, v1 and v2, are employed to automatically identify and locate digestive tract lesions, thereby enhancing the early detection and diagnosis of diseases.

The HAM10000 ("Human Against Machine with 10000 training images") dataset was created in collaboration with dermatologists and computer scientists at the University of

Graz in Austria. It contains 10,015 high-quality dermoscopic images of skin lesions collected from diverse populations, each marked by a professional dermatologist. The dataset covers 7 common skin lesion types: actinic keratoses and intraepithelial carcinoma (AKIEC), melanoma (MEL), benign keratosis (BKL), basal cell carcinoma (BCC), dermatofibroma (DF), vascular lesions (VASC), and melanocytic nevi (NV). The dataset is randomly divided into a training set and a validation set at a ratio of 8:2. Accurate classification of skin lesions is crucial for the early detection of serious diseases such as skin cancer. By providing public datasets like HAM10000, research into the automatic detection and diagnosis of skin lesions is promoted, facilitating the automatic identification and localization of skin cancer lesions.

The Brain Tumor MRI dataset was jointly created by multiple medical research institutions and data challenges, and it includes multimodal MRI images from different patients, totaling 7023 high-quality brain tumor images. Each image is annotated with tumor regions by a professional radiologist. The dataset was randomly divided into training and validation sets in an 8:2 ratio, covering various types of brain tumors, including gliomas (both low-grade and high-grade), meningiomas, neuroglial tumors, and other types (such as metastatic tumors). The precise classification of brain tumors is of great significance for the early detection and diagnosis of serious diseases such as brain tumors. By making this dataset public, we aim to advance research on the automatic detection and diagnosis of brain tumors, thereby promoting progress in the field of medical image analysis.

### *B. Experimental test of the fusion of Confnikster and Tefa*

Five types of attention modules, namely channel attention (abbreviation ConvNeXT + CA), spatial attention(abbreviation ConvNeXT+SA), spatial channel combined attention(abbreviation Convnext+SA+CA and ConvNeXT +SA+CA), CBAM attention(abbreviation ConvNeXT+CBAM) and TFA(abbreviation ConvNeXT+TFA), were integrated into the Conxnext model for comparative experiments. The output of these modules is used as the input to the final inspection head. The test results of Conxnext and the five attention mechanism models on the Kvasir dataset are shown in Table 1, and the results show that the addition of TFA significantly improves the Top1 accuracy by 3.81%.

TABLE I
COMPARISON OF EXPERIMENTAL RESULTS OF CONVNEXT COMBINED WITH ATTENTION MODEL.

| Settings | Top-1(%) | Top-5(%) |
|---|---|---|
| ConvNeXT | 84.83 | 100 |
| ConvNeXT + CA | 86.88 | 100 |
| ConvNeXT + SA | 85.34 | 100 |
| ConvNeXT + CA + SA | 86.12 | 100 |
| ConvNeXT + SA + CA | 85.76 | 100 |
| ConvNeXT + CBAM | 86.25 | 100 |
| ConvNeXT + TFA | 89.03 | 100 |

From the experimental results, it can be observed that the classification accuracy of ConvNext with different attention modules is 86.88%, 85.34%, 86.12%, 85.76%, 86.25%, and 89.03%, respectively. These results show that ConvNext+TFA has the highest Top-1 accuracy, indicating that TFA performs best among these types of attention and achieves a higher accuracy rate when applied to medical image classification tasks.

### *C. Comparative Tests*

To confirm and further strengthen the evidence of the proposed optimization scheme's effectiveness, rigorous tests were carefully performed using four datasets and three models within the context of this study.

*1) Comparative experiments with Confnicst:* ConvNeXT is a new convolutional neural network architecture proposed by Stanford University and the Google AI research team in 2022. It uses a block-like design similar to the Vision Transformer but replaces the Self-Attention mechanism with a more efficient standard convolutional layer. This innovative network structure not only maintains strong model performance but also greatly reduces complexity and computational overhead. Since medical images usually contain rich detailed information, the hierarchical downsampling mechanism of ConvNeXT can effectively capture multi-scale features, resulting in a significant improvement in classification accuracy.

The results of training and validation using the Kvasir v1 dataset are shown in Table 2. The batch size is set to 8, the number of epochs is set to 100, the learning rate is set to 5e-4, and the weight decay (regularization) factor is set to 5e-2. After the experiments, the model achieved an accuracy of 84.83%, indicating its ability to correctly identify and classify gastrointestinal abnormalities.

TABLE II
EXPERIMENTAL RESULTS OF CONVNEXT IN THE KVASIR V1 DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| ConvNeXT | 10.98 | 9.40 | 84.83 | 100 |
| ConvNeXT + SE | 11.14 | 9.40 | 85.63 | 100 |
| ConvNeXT + EMA | 11.18 | 10.64 | 86.13 | 100 |
| ConvNeXT + CBAM | 11.14 | 9.41 | 86.25 | 100 |
| ConvNeXT + GAM | 13.44 | 12.28 | 85.50 | 100 |
| ConvNeXT + TRI | 16.02 | 13.72 | 86.63 | 100 |
| ConvNeXT + PSA | 12.42 | 11.11 | 87.88 | 100 |
| ConvNeXT + TFA | 13.61 | 12.72 | 89.03 | 100 |

The results of training and validation using the Kvasir v2 dataset are shown in Table 3. The parameters used by the model are the same as those set when training and validation were performed using the Kvasir v1 dataset. After experiments, the accuracy of the model is 81.93%.

TABLE III
EXPERIMENTAL RESULTS OF CONVNEXT IN THE KVASIR V2 DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| ConvNeXT | 10.98 | 9.40 | 81.93 | 99.69 |
| ConvNeXT + SE | 11.14 | 9.40 | 82.49 | 100 |
| ConvNeXT + EMA | 11.18 | 10.64 | 89.81 | 100 |
| ConvNeXT + CBAM | 11.14 | 9.41 | 83.93 | 100 |
| ConvNeXT + GAM | 13.44 | 12.28 | 88.69 | 100 |
| ConvNeXT + TRI | 16.02 | 13.72 | 82.24 | 99.69 |
| ConvNeXT + PSA | 12.42 | 11.11 | 90.75 | 100 |
| ConvNeXT + TFA | 13.61 | 12.72 | 91.25 | 100 |

The outcomes of training and validation using HAM10000 datasets are displayed in Table 4. Following trials, the model's accuracy was 81.68%, meaning it could accurately detect and categorize atypical skin lesions.

TABLE IV
EXPERIMENTAL RESULTS OF CONVNEXT IN THE HAM10000 DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| ConvNeXT | 10.98 | 9.40 | 81.68 | 99.62 |
| ConvNeXT + SE | 11.14 | 9.40 | 81.89 | 99.50 |
| ConvNeXT + EMA | 11.18 | 10.64 | 81.80 | 99.56 |
| ConvNeXT + CBAM | 11.14 | 9.41 | 81.93 | 99.69 |
| ConvNeXT + GAM | 13.44 | 12.28 | 82.58 | 99.81 |
| ConvNeXT + TRI | 16.02 | 13.72 | 82.24 | 99.69 |
| ConvNeXT + PSA | 12.42 | 11.11 | 82.46 | 99.79 |
| ConvNeXT + TFA | 13.61 | 12.72 | 84.88 | 100 |

The results of training and validating using the Brain Tumor MRI dataset are shown in Table 5.The parameters used in the model are consistent with those set during training and validation using the Kvasirv1 dataset.After the experiment, the model achieved an accuracy rate of 98.29%, indicating that it can correctly identify and classify brain tumor lesions.

TABLE V
EXPERIMENTAL RESULTS OF CONVNEXT IN THE BRAIN TUMOR MRI DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| ConvNeXT | 10.98 | 9.40 | 98.29 | 100 |
| ConvNeXT + SE | 11.14 | 9.40 | 98.58 | 100 |
| ConvNeXT + EMA | 11.18 | 10.64 | 99 | 100 |
| ConvNeXT + CBAM | 11.14 | 9.41 | 98.43 | 100 |
| ConvNeXT + GAM | 13.44 | 12.28 | 98.86 | 100 |
| ConvNeXT + TRI | 16.02 | 13.72 | 99 | 100 |
| ConvNeXT + PSA | 12.42 | 11.11 | 99.29 | 100 |
| ConvNeXT + TFA | 13.61 | 12.72 | 99.37 | 100 |

Through the experimental results of these four datasets, it was found that TFA achieves the best performance in medical image classification tasks within the ConvNeXt model. In the Kvasir v1 dataset, TFA's Top-1 accuracy is 4.2% higher than the original model, in the Kvasir v2 dataset, TFA's Top-1 accuracy is 9.82% higher than the original model, in the HAM10000 dataset, TFA's Top-1 accuracy is 5.59% higher than the original model, and in the Brain Tumor MRI dataset, TFA's Top-1 accuracy is 1.08% higher than the original model. This shows that TFA is more effective than other attention mechanisms in the application of medical image classification tasks, performing the best.

*2) Comparative experiment with Mobilenetv3:* MobileNetV3 is a lightweight convolutional neural network model designed for efficient image recognition and classification tasks on computationally resource-constrained mobile devices. MobileNetV3 features a network design based on the inverted residual structure and incorporates an efficient convolutional block with an h-swish activation function, significantly reducing computational complexity. Additionally, the model includes a mechanism to adaptively adjust the depth and width of the network, allowing the network structure to be dynamically tailored to different tasks and hardware conditions. MobileNetV3 also utilizes neural architecture search technology to automatically optimize the network design, further enhancing model performance.

MobileNetV3 offers two variants: MobileNetV3-Large for more powerful devices and MobileNetV3-Small for resource-constrained devices. After the final convolutional layer, the model employs global average pooling to convert the feature map into a vector representation and uses the Softmax function for classification prediction.

In this experiment, MobileNetV3 is trained and validated using the same parameter settings as the Convnext model.The experiments were conducted on the MobileNetV3-Large model and tested on the Kvasir v1, Kvasir v2, HAM10000, and Brain Tumor MRI datasets, with the results shown in Tables 6, 7, 8, and 9, respectively.

TABLE VI
EXPERIMENTAL RESULTS OF MOBILENETV3 IN THE KVASIR V1 DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| MobileNetV3 | 4.21 | 1.21 | 84.27 | 100 |
| MobileNetV3 + EMA | 3.17 | 2.84 | 89.13 | 100 |
| MobileNetV3 + CBAM | 3.08 | 1.21 | 88.25 | 100 |
| MobileNetV3 + GAM | 7.80 | 3.68 | 86.50 | 100 |
| MobileNetV3 + TRI | 2.70 | 1.22 | 89.88 | 100 |
| MobileNetV3 + PSA | 2.70 | 1.20 | 86.88 | 100 |
| MobileNetV3 + TFA | 2.70 | 1.20 | 90.25 | 100 |

TABLE VII
EXPERIMENTAL RESULTS OF MOBILENETV3 IN THE KVASIR V2 DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| MobileNetV3 | 4.21 | 1.21 | 87.63 | 100 |
| MobileNetV3 + EMA | 3.17 | 2.84 | 91.19 | 100 |
| MobileNetV3 + CBAM | 3.08 | 1.21 | 90.75 | 100 |
| MobileNetV3 + GAM | 7.80 | 3.68 | 91.25 | 100 |
| MobileNetV3 + TRI | 2.70 | 1.22 | 92.69 | 100 |
| MobileNetV3 + PSA | 2.70 | 1.20 | 90.56 | 100 |
| MobileNetV3 + TFA | 2.70 | 1.20 | 93.13 | 100 |

TABLE VIII
EXPERIMENTAL RESULTS OF MOBILENETV3 IN HAM10000 DATASETS.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| MobileNetV3 | 4.21 | 1.21 | 80.49 | 99.62 |
| MobileNetV3 + EMA | 3.17 | 2.84 | 81.99 | 99.56 |
| MobileNetV3 + CBAM | 3.08 | 1.21 | 81.11 | 99.25 |
| MobileNetV3 + GAM | 7.80 | 3.68 | 83.36 | 99.62 |
| MobileNetV3 + TRI | 2.70 | 1.22 | 84.55 | 99.62 |
| MobileNetV3 + PSA | 2.70 | 1.20 | 80.55 | 99.12 |
| MobileNetV3 + TFA | 2.70 | 1.20 | 85.55 | 99.91 |

Through the experimental results of these four datasets, it was found that the baseline accuracies of the MobileNetV3 model on the Kvasir v1, Kvasir v2, HAM10000, and Brain Tumor MRI datasets were 84.27%, 87.63%, 80.49%, and 98.58%, respectively. This indicates that the MobileNetV3 model can be applied to medical image classification tasks. After experimental comparisons, it was found that the highest Top-1 accuracy with TFA added on the Kvasir v1 dataset

TABLE IX
EXPERIMENTAL RESULTS OF MOBILENETV3 IN BRAIN TUMOR MRI DATASETS.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| MobileNetV3 | 4.21 | 1.21 | 98.58 | 100 |
| MobileNetV3 + EMA | 3.17 | 2.84 | 99.22 | 100 |
| MobileNetV3 + CBAM | 3.08 | 1.21 | 98,79 | 100 |
| MobileNetV3 + GAM | 7.80 | 3.68 | 98.86 | 100 |
| MobileNetV3 + TRI | 2.70 | 1.22 | 99.07 | 100 |
| MobileNetV3 + PSA | 2.70 | 1.20 | 98.72 | 100 |
| MobileNetV3 + TFA | 2.70 | 1.20 | 99.39 | 100 |

was 90.25%, which is 5.98% higher than the baseline model accuracy. On the Kvasir v2 dataset, the Top-1 accuracy with TFA added was 93.13%, which is 5.5% higher than the baseline model accuracy. On the HAM10000 dataset, the Top-1 accuracy with TFA added was 85.55%, exceeding the baseline model accuracy by 5.06%. On the Brain Tumor MRI dataset, the Top-1 accuracy with TFA added was 99.85%, exceeding the baseline model accuracy by 1.27%. The accuracy with TFA added on the Kvasir v1, Kvasir v2, HAM10000, and Brain Tumor MRI datasets was the highest.It indicates that the TFA classification model has the best medical image classification capability.

*3) Regent's comparative experiments:* RegNet (Regularized Network) is a series of convolutional neural network (CNN) models used for image classification tasks.The model structure of RegNet is a stacked architecture based on bottleneck blocks and depthwise separable convolutions. By controlling the number and width of the basic blocks, the model's complexity can be adjusted. This structure maintains efficiency while providing good image classification performance.In RegNet, the activation function is usually ReLU (Rectified Linear Unit) as the default choice.The ReLU function maintains linear growth in the positive range and outputs zero for negative values.This function is simple, computationally efficient, and performs well in practice.The parameters used in this model are consistent with those set during the training and validation of the Convnext model.The experimental results of RegNet on the Kvasir v1, Kvasir v2, HAM10000, and Brain Tumor MRI datasets are shown in Tables 10, 11, 12, and 13.

TABLE X
RESULTS OF REGNET EXPERIMENTS IN THE KVASIR V1 DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| RegNet | 4.78 | 2.14 | 83.88 | 100 |
| RegNet + EMA | 5.11 | 3.17 | 88.50 | 100 |
| RegNet + CBAM | 5.05 | 2.16 | 84.50 | 100 |
| RegNet + GAM | 4.78 | 2.14 | 84.62 | 100 |
| RegNet + TRI | 4.78 | 2.17 | 87.25 | 100 |
| RegNet + PSA | 4.78 | 2.14 | 84.25 | 100 |
| RegNet + TFA | 4.78 | 2.14 | 88.75 | 100 |

Through experiments, the baseline accuracy of the model is 83.88%, 88.38%, 77.99%, and 98.64%, respectively, indicating that it can correctly classify lesions in medical images and can be easily used in low-resource environments. Experiments on the Kvasir v1, Kvasir v2, HAM10000, and Brain Tumor MRI datasets show that the accuracy with TFA added reaches the highest, achieving 88.75%, 91.81%, 85.57%, and 99.41%, respectively.The comparison reveals that TFA still performs excellently across different network models.

TABLE XI
EXPERIMENTAL RESULTS OF REGNET IN THE KVASIR V2 DATASET.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| RegNet | 4.78 | 2.14 | 88.38 | 100 |
| RegNet + EMA | 5.11 | 3.17 | 90.75 | 100 |
| RegNet + CBAM | 5.05 | 2.16 | 90.25 | 100 |
| RegNet + GAM | 4.78 | 2.14 | 88.94 | 100 |
| RegNet + TRI | 4.78 | 2.17 | 90.13 | 100 |
| RegNet + PSA | 4.78 | 2.14 | 88.88 | 100 |
| RegNet + TFA | 4.78 | 2.14 | 91.81 | 100 |

TABLE XII
RESULTS OF REGNET EXPERIMENTS IN HAM10000 DATASETS.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| RegNet | 4.78 | 2.14 | 77.99 | 99.12 |
| RegNet + EMA | 5.11 | 3.17 | 83.68 | 99.62 |
| RegNet + CBAM | 5.05 | 2.16 | 83.61 | 99.81 |
| RegNet + GAM | 4.78 | 2.14 | 78.92 | 99.50 |
| RegNet + TRI | 4.78 | 2.17 | 84.68 | 99.87 |
| RegNet + PSA | 4.78 | 2.14 | 78.49 | 99.25 |
| RegNet + TFA | 4.78 | 2.14 | 85.57 | 99.72 |

TABLE XIII
RESULTS OF REGNET EXPERIMENTS IN BRAIN TUMOR MRI DATASETS.

| Settings | Param.(M) | FLOPs(M) | Top-1(%) | Top-5(%) |
|---|---|---|---|---|
| RegNet | 4.78 | 2.14 | 98.64 | 100 |
| RegNet + EMA | 5.11 | 3.17 | 99.29 | 100 |
| RegNet + CBAM | 5.05 | 2.16 | 98.86 | 100 |
| RegNet + GAM | 4.78 | 2.14 | 98.79 | 100 |
| RegNet + TRI | 4.78 | 2.17 | 99.22 | 100 |
| RegNet + PSA | 4.78 | 2.14 | 98.86 | 100 |
| RegNet + TFA | 4.78 | 2.14 | 99.41 | 100 |

## V. CONCLUSION

The fundamental concept of this research is to extract features at various levels using convolutional layers of multiple scales, then adaptively weight these features using the CBAM mechanism, fuse the weighted features into an adaptive convolutional layer, and add residual connections to improve the model's learning capacity. The experimental findings show that by successfully combining features from various levels, TFA greatly increases the classification accuracy of medical images. TFA solves the problems of inadequate feature diversity and the failure of current attention methods to detect minor lesions, achieving the greatest accuracy across several datasets. This enhancement makes it easier for the model to concentrate on key elements, improves performance and generalization, and opens the door to investigating the TFA module's potential applications in other fields.

## REFERENCES

[1] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future

promises," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, 2021.

[2] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of Medical Systems*, vol. 42, pp. 1–13, 2018.

[3] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–18, 2019.

[4] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[6] A. Al-Sabaawi, H. M. Ibrahim, Z. M. Arkah, M. Al-Amidie, and L. Alzubaidi, "Amended convolutional neural network with global average pooling for image classification," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2020, pp. 171–180.

[7] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[8] N. Murray and F. Perronnin, "Generalized max pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2473–2480.

[9] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*. IEEE, 2014, pp. 844–848.

[10] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 713–13 722.

[11] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[12] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[13] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, and S. Yin, "Deep learning attention mechanism in medical image analysis: Basics and beyonds," *International Journal of Network Dynamics and Intelligence*, pp. 93–116, 2023.

[14] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine*, vol. 8, no. 11, 2020.

[15] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[16] Z. Pylyshyn, "Some primitive mechanisms of spatial attention," *Cognition*, vol. 50, no. 1-3, pp. 363–384, 1994.

[17] L. Xu, J. Huang, A. Nitanda, R. Asaoka, and K. Yamanishi, "A novel global spatial attention mechanism in convolutional neural network for medical image classification," *arXiv preprint arXiv:2007.15897*, 2020.

[18] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 783–792.

[19] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.