Medical Image Classification via Integrated Attention Mechanisms and Dynamic Weighted Loss Function

Jingyi Lyu[1, 2], Benxiang Jiang[1, 2], Songze Zhang[1, 2], Hongjian Shi[1, 2, *]
[1] *Hong Kong Baptist University, Hong Kong, China*
[2] *Beijing Normal-Hong Kong Baptist University, Zhuhai, China*

***Abstract*—A concise and efficient network is the goal pursued in the field of medical image classification. In this study, we developed a lightweight medical image classification module for efficient classification of medical images. This module integrates spatial attention mechanism and channel attention mechanism to effectively capture channel and spatial information. In order to alleviate the problem of data imbalance in medical image classification without increasing the training data, we proposed a novel loss function DWDI. The loss function dynamically adjusts the influence of category weights during the training process, thereby enhancing the model's learning ability for minority categories. We validated our method on seven public datasets, and the results demonstrated that our method significantly improved classification accuracy while maintaining strong competitiveness in model complexity.**

***Index Terms*—Medical image, classification, data imbalance, loss function**

I. Introduction

Accurate medical image classification can greatly improve the accuracy and efficiency of diagnosis, and provide patients with more timely and effective treatment plans [1]. In addition, automated image classification systems can serve as a second opinion for doctors, providing auxiliary decision support [2].

Deep learning models, especially Convolutional Neural Networks (CNNs), can automatically learn complex features in images, thereby improving the accuracy and efficiency of classification [4]. However, medical image classification faces the challenge of data imbalance [5]. The uneven distribution of these data may cause classification models to lean towards the majority class, thereby affecting the accuracy of identifying minority classes [6]. Although some generative models have been designed to alleviate the imbalance problem this approach entails an increase in both training data volume and time consumption [7].

This article focuses on designing a lightweight network for extracting medical image features without the need for additional data generation and training to improve recognition accuracy. In this article, we proposed a hybrid attention block, and then we combined the channel attention module with the proposed hybrid attention block as the basic structure of the classification network. In addition, to address the negative impact of data imbalance on the network, we proposed a dynamic weight allocation loss function, and the proposed dynamic weight loss function is combined with dice loss as the guiding loss for model training. The contribution of this article can be summarized as follows:

(1) We have designed a lightweight CHyA block. This block serves as the foundational structure for image classification network.

(2) To address the issue of data imbalance, we introduced a novel loss function DWDI to guide the network training stage.

(3) We performed experiments on 7 datasets to validate the effectiveness of our proposed module, and numerous experiments have demonstrated the effectiveness of our method.

* Corresponding author: shihj@bnbu.edu.cn

This work was supported in part by the Guangdong Higher Education Key Platform and Research Project under Grant 2020ZDZX3039, in part by the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science under Grant 2022B1212010006.

II. Related work

CNN-based medical image classification. Deep CNN networks have been widely applied in the recognition and classification of various diseases due to their powerful feature extraction capabilities. Many authors have demonstrated the effectiveness of deep convolutional networks in the field of medical image classification, disease diagnosis, cancer detection, etc [3] [4] [5]. Among them, Lei et al. [4] proposed combining depth descriptors with manual features of ultrasound images for the classification of placental images. In [5], a hybrid model based on CNN and Transformer was proposed, which achieved comparable classification results for the recognition of skin cancer.

Technologies for alleviating data imbalance. The popular method is to use GAN networks to generate images or to apply SMOTE (Synthetic Minority Oversampling Technique) [9] to synthesize few class samples to alleviate data imbalance. Golhar et al. [8] proposed using GAN to generate colonoscopy images, and the generated images were used as training data for the polyp classification model to improve their classification performance. Ding et al. [10] combined the iForest algorithm with a GAN model to identify sparse and boundary samples and generate images for training. The core idea of SMOTE is to generate new composite samples by interpolating between minority class samples [9]. Chamseddine et al. [11] applied SMOTE technology to the COVID-19 chest X-ray image classification task, and achieved comparable results.

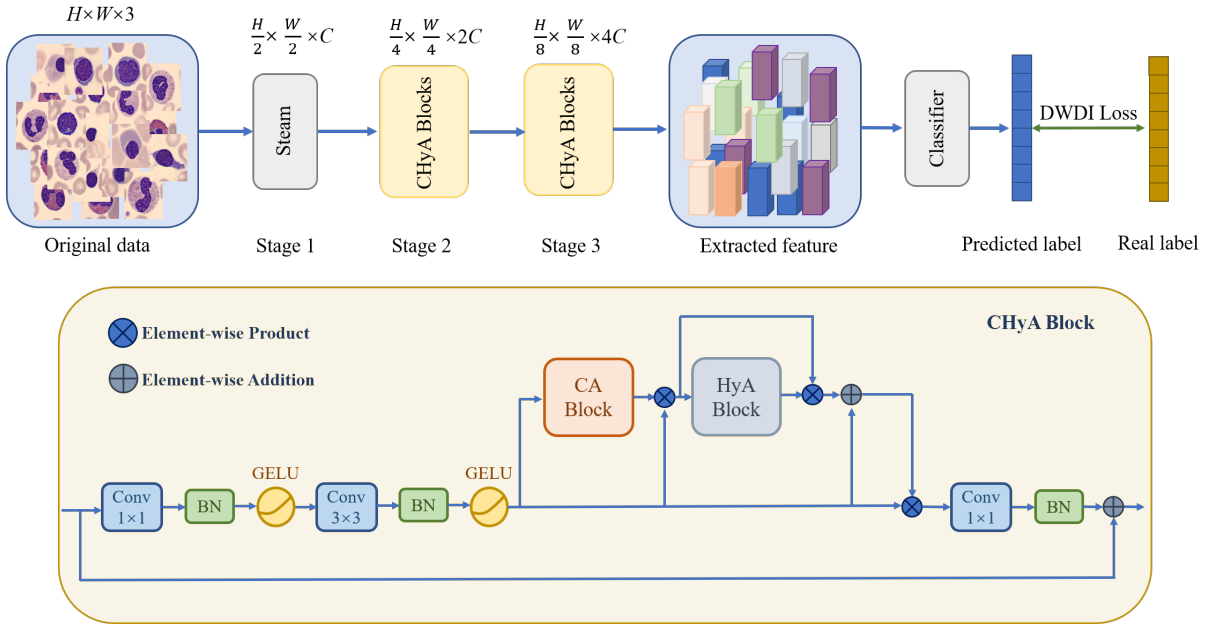

Fig. 1. The proposed CHyA block and feature extraction framework. Where '⊗' represent the element-wise multiplication operation and '⊕' is element-wise add operation.

Motivation. Although the above methods have achieved some classification results, data generation can lead to an increase in training time. Another problem is that the model's performance is affected by data imbalance, and this impact theoretically changes as the training process progresses, which the methods mentioned above have not taken into account. This article aims to design a lightweight neural network architecture that reduces the computational burden of model training while ensuring classification performance. In addition, to further improve the model's performance on imbalanced data, we propose a loss function that dynamically adjusts weights. Our approach does not require additional training data, reducing the training time complexity of the model.

III. Methodology

A. Overview of MBConv block

The MBConv block [12] is an inverse residual module proposed in 2018, which includes a 1x1 convolutional layer to expand the input channel, followed by a 3x3 convolutional layer to encode local spatial interactions, and then a squeeze and exclusion (SE) module [13] to learn the feature dependency relationships between channels, allowing the network to focus on discriminative features. Finally, a 1x1 convolution is used to restore the channels to their initial dimensions and perform residual connections with the input image. The MBConv block can learn channel information and feature dependencies of images, which can be used for different computer vision tasks.

B. Proposed CHyA block and framework

The CHyA block, as depicted in Fig.1. It is designed to enhance the feature extraction capabilities of medical images through a two-pronged approach: a **C**hannel Attention block and a **Hy**brid **A**ttention block. When a medical image is fed into the system, it undergoes a three-step process to transform the raw image data into a rich feature representation.

In the first stage, the goal is to transform the original image (H×W×3) into a more meaningful embedding that can be effectively processed by the network. This is achieved through a sequence of operations: a convolution layer with a 3×3 kernel size, which helps to extract spatial features; a batch normalization (BN) layer, which standardizes the inputs to the next layer; and the application of the Gaussian Error Linear Units (GELU) activation function. The GELU function introduces non-linearity, allowing the network to capture complex patterns. The output of this stage is an image embedding F, which has been down sampled to half the original dimensions (H/2×W/2×C), where C is the number of channels.

In the second and third stages, we make use of several CHyA blocks. The embedding F, which is produced in the first stage, undergoes further processing through two convolutional layers. This is followed by batch normalization and GELU function. We also use F as the processed internal embedding. Next, we feed F into both the CA block and the HyA block to harness the deeper features. We represent the functions of these blocks using P_{CA} for the CA block and P_{HyA} for the HyA block. The procedures of these two blocks can be outlined as follows:

$$P_{CA}(F) = \mathrm{Sig}\left(\mathrm{FC}(\mathrm{MaxPool}(F)) + \mathrm{FC}(\mathrm{AvgPool}(F))\right) \tag{1}$$

$$\mathrm{FC}(x) = \mathrm{Conv}_{1\times1}\left(\mathrm{GELU}\left(\mathrm{Conv}_{1\times1}(x)\right)\right) \tag{2}$$

Where Sig is the sigmoid function used to produce probabilistic values of CA, Conv function is used to create convolutional layers, each with a size of 1×1. AvgPool and MaxPool represent the average pooling max pooling operation, respectively. CA block learning the weights of each channel, in this way, useful features for the current task are highlighted while unimportant features are suppressed.

For the operation of P_{HyA}, it combines the spatial attention mechanism and feature enhanced representation operation (FER), the P_{HyA} operation is given as follows:

$$P_{HyA} = \mathrm{FER}\left(\mathrm{Conv}_{7\times7}\left(\mathrm{Concat}\left(\mathrm{AvgPool}(F), \mathrm{MaxPool}(F)\right)\right)\right) \tag{3}$$

$$\mathrm{FER}(x) = \mathrm{Sig}\left(\mathrm{Linear}_{\mathrm{expand}}\left(\mathrm{GELU}\left(\mathrm{Linear}_{\mathrm{shrink}}\left(\mathrm{AvgPool}(x)\right)\right)\right)\right) \tag{4}$$

In equation (4), $\mathrm{Linear}_{\mathrm{expand}}$ and $\mathrm{Linear}_{\mathrm{shrink}}$ represent the first linear layer shrink the input data while the second linear layer expand the data back. The P_{HyA} operation first performs global average pooling and global maximum pooling on the spatial dimension embedding *F*, and then concatenates the results of these two pooling operations. This operation can capture the global statistical information of the feature map, including the average and maximum values. The concatenated feature map is processed through a 7×7 convolutional layer. After all convolution operations and linear transformations, use the sigmoid function to obtain the probabilistic values of the HyA block. And the output F' of HyA block is:

$$F' = P_{HyA}(P_{CA}(F) \otimes F) \otimes F \tag{5}$$

C. Proposed dynamic weighted loss function DWDI

Weighted cross entropy loss can effectively alleviate the imbalance problem in the classification process of imbalanced datasets. However, the impact of data imbalance on model training gradually decreases. Therefore, we use dynamically adjusted weighted cross entropy loss (WCE) and cross entropy loss (CE) to guide the learning of the network, gradually reducing the weight of WCE loss and increasing the weight of CE loss as training progresses. **D**ynamic **w**eighted **c**ross **e**ntropy loss function is defined as follows:

$$l_{DWCE} = \left(1 - \left(\frac{1}{i}\right)\right) l_{CE}(\textit{Real}, \textit{Pred}) + \left(\frac{1}{i}\right) l_{WCE}(\textit{Real}, \textit{Pred}) \tag{6}$$

In equation (6), *Real* and *Pred* represent the real label and the predicted label of test image. Parameter *i* refers to the model being in the *i*-th training epoch.

Dice Loss can effectively handle class imbalance problems by calculating the ratio of the intersection and union predicted and true labels, as it is very sensitive to samples from a few classes, which can promote the model to achieve better performance on these classes. In this article, we first calculate the dice loss for each category and calculate the average dice loss for all categories:

$$l_{Dice} = \frac{1}{N}\sum_{j=1}^{N}\left(1 - \frac{2|\textit{Real}_j \cap \textit{Pred}_j|}{|\textit{Real}_j \cup \textit{Pred}_j|}\right) \tag{7}$$

Where *N* is the number of categories in the dataset, $Real_j$ and $Pred_j$ represent the real label distribution and the predicted probability distribution of category j. In this study, we use the weighted fusion value of l_{DWCE} and l_{Dice} as the loss function for model training. The weighted **D**ynamic **w**eighted cross entropy loss and **Di**ce loss is given as follows:

$$l_{DWDI} = \lambda l_{DWCE} + (1-\lambda) l_{Dice} \tag{8}$$

In equation (8), the parameter λ is the weight coefficient between the two loss functions.

IV. Experiments

A. Datasets

We use 7 publicly available datasets to validate the effectiveness of the model. Fig. 2. shows some examples of 7 datasets.

MedMNIST [14] is a large-scale medical image classification dataset that includes multiple modalities of images such as ultrasound, X-rays, CT, and so on. The experiment in this article mainly focuses on multi classification tasks, so we selected DermaMNIST, RetinaMNIST, BloodMNIST, and OrganaMNIST datasets to verify the proposed method.

PAD-UFES-20 [15] dataset contains 2298 skin disease images taken by various smartphones from 1373 patients. This dataset involves 6 categories, including 3 types of skin cancer, namely Squamous Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and Melanoma (MEL); The other three are skin diseases, namely actinic keratosis (ACK), seborrheic keratosis (SEK), and nevus (NEV).

Fetal-Planes-DB [16] is a publicly available maternal fetal ultrasound image dataset, which is a large dataset of routinely collected maternal fetal screening ultrasound images collected by several operators and ultrasound machines from two different hospitals. It contains 12400 images from six categories: fetal abdomen, fetal brain, fetal femur, fetal thorax, material cervix, and 'other' category.

ISIC2018 [17] contains a large number of annotated skin lesion images, which are suitable for training various classification models. It contains skin images from 7 categories, with a training dataset of over 10000 images and a validation and testing set of 193 and 1512, respectively.

B. Experimental details

We tested the effectiveness of our method on 7 multi class datasets, including DermaMNIST, RetinaMNIST, BloodMNIST, OrganaMNIST, PAD-UFES-20, ISIC2018, and Fetal-Planes-DB. To ensure fairness in comparison, we set all images to 224×224×3 before inputting them into the network. All experiments are conducted on NVIDIA GeForce RTX 4090 GPU with a batch size of 16. We used the Adam optimizer with a fixed learning rate of 0.001. The parameters λ is set

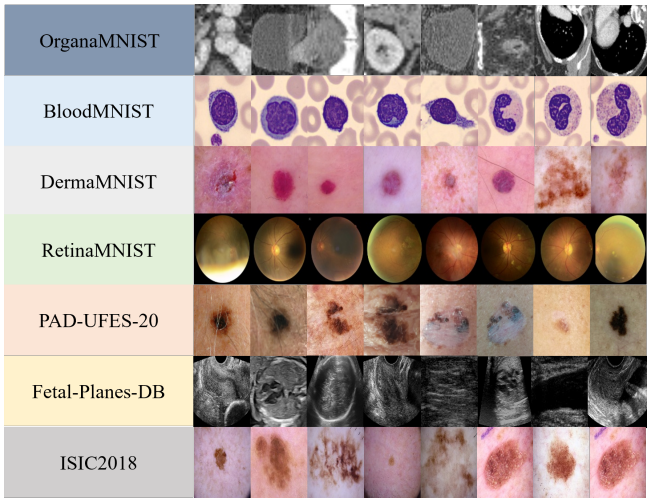

Fig. 2. Some samples of experimental datasets.

to 0.7. We fixed the training epochs of MedMNIST dataset to 200 and other datasets to 300. For the MedMNIST dataset, we recorded the percentages of Accuracy (ACC) metric and Area Under the ROC Curve (AUC) metric of the model. For other datasets, we recorded the percentages of precision (Pre), sensitivity (Sens), specificity (Spe), F1-score (F1) and ACC metrics of the model to validate its effectiveness.

C. Results and analysis

In this section, we present the experimental results of the proposed method on 7 publicly available datasets. In the following text, 'ours (336)' indicates that the first stage of the model includes 3 stem blocks, the second stage includes 3 CHyA blocks, and the third stage includes 6 CHyA blocks. The meanings expressed by (669) is the same as (336).

1) Results of MedMNIST dataset: The experimental results of our proposed method on the MedMNIST dataset are presented in Table I. Compared to the well-known residual network, our approach demonstrates significantly better performance across multiple datasets, including DermaMNIST, RetinaMNIST, BloodMNIST, and OrganaMNIST. Specifically, our minimum model outperforms ResNet-50 (224) on the MedMNIST dataset. Notably, the accuracy (ACC) of DermaMNIST achieved by our method exceeds the ResNet-18 (224) by 10.1%. Similarly, on the RetinaMNIST dataset, both the AUC and ACC metrics of the method ours (336) show significant improvements, increasing by 8.8% and 8%, respectively.

When compare with state-of-the-art (SOTA) methods, our proposed model demonstrates superior performance in both AUC and ACC metrics. For instance, our method ours (669) achieves 96.9% AUC and 85.5% ACC on the DermaMNIST dataset, outperforming MedViTV2-L by 1.9% and 3.8%, respectively. Additionally, our method ours (669) surpasses the latest methods on the RetinaMNIST dataset, with AUC and ACC reaching 84.4% and 64.8%, respectively, which represents improvements of 5.9% and 7% over existing approaches. Similarly, our method achieves state-of-the-art performance on the BloodMNIST and OrganaMNIST datasets.

2) Results of PAD-UFES-20 and ISIC2018 datasets: The results of the PAD-UFES-20 dataset are shown in Table II. Our smallest model ours (336) outperforms the currently best performing Mamba models VMamba-T and MedMamba-T in all five metrics and outperforms VMamba-T by 7% in ACC metrics, while its parameters are also much smaller than VMamba-T. Compared with the best performing MedViTV2-T model, ours (669) model's results exceeded 2.5%, 5.2%, 0.7%, 4.9%, and 4.1% in Precision, Sensitivity, Specificity, F1 score, and ACC, respectively. Overall, our model achieved competitive performance on the PAD-UFES-20 dataset. For dataset ISIC2018, our model achieved the best results in five metrics, surpassing the current best performing model MedViTV2-T. It is worth noting that our smallest model ours (336), with a parameter size of 11M, achieved competitive results with a sensitivity of 78.2%, and ACC of 78.1%. In addition, ours (336) model achieves the best classification results among all compared SOTA models while maintaining the lowest FLOPs.

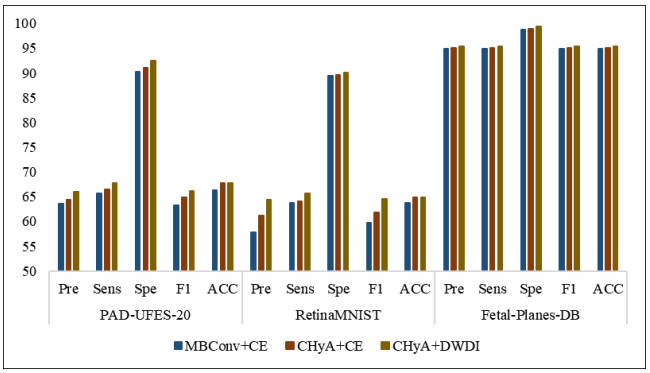

Fig. 3. Comparison results of different modules on three datasets.

V. Ablation Study

A. Verify the effectiveness of CHyA block and DWDI

We recorded 5 metrics for three datasets, PAD-UFES-20, RetinaMNIST, and Fetal-Planes-DB, under different models and loss functions: precision(Pre), sensitivity(Sen), specificity(Spe), F1-score(F1), AUC, and ACC. MBConv+CE represents we use MBConv block as the basic structure and cross entropy loss function to train the model, CHyA+CE means the basic structure of the feature extraction framework and use the cross entropy loss function, and CHyA+DWDI represent we use the proposed CHyA block and the proposed DWDI loss function. From the results shown in Fig. 3, the values of the bar charts for each indicator gradually increase from left to right, especially for the precision and F1 score of the RetinaMNIST dataset, which exhibit a clear upward trend. In addition, the F1 score and AUC of PAD-UFES-20 also showed an increasing trend and reached their maximum values on the CHyA+DWDI model. This indicates that our method has a certain degree of competitiveness compared to MBConv, and

TABLE I
THE RESULTS OF PROPOSED METHOD ON MEDMNIST DATASET.

Methods	DermaMNIST		RetinaMNIST		BloodMNIST		OrganaMNIST	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18(28) [27]	91.7	73.5	71.7	52.4	99.8	95.8	99.7	93.5
ResNet-18(224) [27]	92.0	75.4	71.0	49.3	99.8	96.3	99.8	95.1
ResNet-50(28) [27]	91.3	73.5	72.6	52.8	99.8	96.3	99.7	93.8
ResNet-50(224) [27]	91.2	73.1	71.6	51.1	99.8	96.3	99.7	94.0
auto-sklearn [28]	90.2	71.9	69.0	51.5	97.3	90.7	98.3	89.6
AutoKeras [29]	91.5	74.9	71.9	50.3	99.8	96.2	99.7	93.7
MedViTV1-L [3]	92.0	77.3	75.4	55.2	99.8	96.5	99.8	95.1
MedMamba-B [2]	92.5	75.7	71.5	55.3	99.9	98.3	99.9	96.4
MedViTV2-L [26]	95.0	81.7	78.5	57.8	99.9	98.7	99.9	97.3
Ours (336)	96.7	84.6	81.4	60.8	99.9	99.3	99.9	96.7
Ours (669)	**96.9**	**85.5**	**84.4**	**64.8**	**99.9**	**99.4**	**99.9**	**97.4**

TABLE II
THE RESULTS OF PROPOSED METHOD ON PAD-UFES-20 AND ISIC2018 DATASET.

Methods	PAD-UFES-20					ISIC2018					Param	FLOPs
	Pre	Sens	Spe	F1	ACC	Pre	Sens	Spe	F1	ACC		
Swin-T [18]	38.2	41.1	90.6	39.5	60.5	60.7	66.1	91.5	61.9	66.1	27.5M	4.5G
ConvNext-T [19]	37.2	33.6	88.9	33.7	54.3	65.3	67.1	91.6	63.2	67.1	27.8M	4.5G
Mobilevitv2-200 [20]	33.9	32.9	88.0	32.2	49.9	66.4	68.1	92.0	65.2	68.1	17.4M	5.6G
EdgeNext-base [21]	35.0	36.4	89.9	34.6	57.6	64.3	67.7	91.7	64.5	67.7	17.9M	2.9G
Nest-tiny [22]	49.9	45.5	91.3	42.3	63.5	67.6	69.1	91.3	64.2	69.1	16.7M	5.8G
Mobileone-s4 [23]	35.9	32.2	87.9	32.3	49.3	70.0	72.2	93.0	70.0	72.2	12.9M	3.0G
Cait-xxs36 [24]	37.1	37.8	90.0	37.0	58.6	56.6	63.9	90.1	68.4	63.9	17.1M	3.8G
VMamba-T [25]	53.2	40.6	90.0	41.6	59.3	70.5	72.5	92.8	70.3	72.5	22.1M	4.4G
MedMamba-T [2]	38.4	36.9	89.9	35.8	58.8	72.2	74.1	93.4	72.3	74.0	14.5M	4.5G
MedViTV1-T [3]	53.3	59.8	90.4	56.2	59.8	71.5	72.4	92.4	69.4	72.4	31.1M	11.7G
MedViTV2-T [26]	63.6	62.5	91.7	61.2	63.6	76.1	77.1	94.4	76.2	77.1	12.3M	5.1G
Ours (336)	64.5	66.6	91.3	64.2	66.3	77.5	78.2	**94.6**	77.3	78.1	11.0M	2.9G
Ours (669)	**66.1**	**67.7**	**92.4**	**66.1**	**67.7**	**77.7**	**78.6**	94.5	**77.5**	**78.6**	18.1M	5.8G

also verifies that our proposed DWDI loss function has a good effect on alleviating data imbalance problems.

B. Results of different λ

We determined the optimal λ values on four datasets: BloodMNIST, DermaMNIST, RetinaMNIST, and OrganaMNIST, and applied the results to other datasets. We record the indicator changes of λ from 0 to 1 at intervals of 0.1. Fig. 4 shows the results of AUC and ACC for four datasets with different λ values. From the line graph, we can observe that the performance of the two indicators is the worst when λ value is 0. When λ value is 0.1, AUC and ACC show significant improvement on the three datasets of DermaMNIST, RetinaMNIST, and OrganaMNIST, but there is no significant change in the AUC indicator of BloodMNIST. As λ gradually increases, the ACC index of the four datasets shows a certain fluctuation trend, while the AUC index is relatively stable. It should be noted that when the value of λ is 0.7, all indicators except for the AUC of DermaMNIST and the ACC of RetinaMNIST reach their maximum values. Therefore, this article decides to use 0.7 as the value of λ and apply it to all datasets.

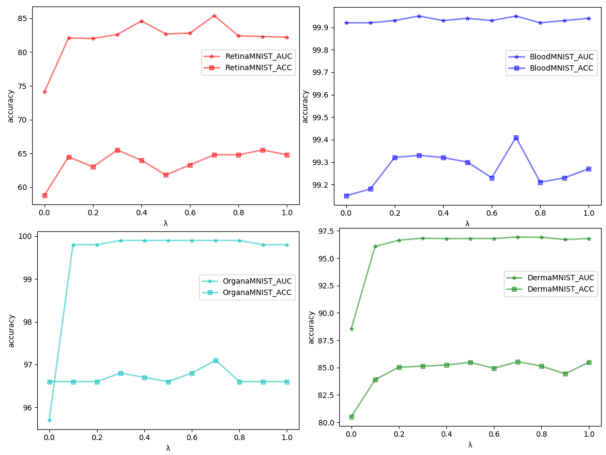

Fig. 4. Comparison results of different λ values on four datasets.

C. Verify the effectiveness of DWDI on minority category

We calculated the classification results of the category with the least number of images in datasets Fetal-Planes-DB (Fetal), RetinaMNIST (Retina), and PAD-UFES-20 (PAD). We

TABLE III
THE RESULTS OF MINORITY CATEGORY.

Method	Fetal		Retina		PAD	
	ACC	F1	ACC	F1	ACC	F1
CHyA+CE	90.8	91.6	62.5	30.7	54.5	46.1
CHyA+DWDI	92.4	92.0	63.3	33.4	56.2	47.3

calculated the ACC and F1-score (F1) for the Fetal abdomen category in dataset Fetal-Planes-DB, the number 4 category in dataset RetinaMNIST, and the MEL category in dataset PAD-UFES-20, As shown in Table III, our method has improved the classification performance of minority classes to a certain extent compared to CE loss.

VI. CONCLUSION

In this paper, we proposed a lightweight block and uses it as the infrastructure to design a feature extraction network for the classification task of medical images. At the same time, a novel loss function is proposed to alleviate the problem of data imbalance. We conducted experiments on 7 publicly available datasets and compared our proposed method with the SOTA method. Our method outperforms existing methods in terms of metrics, and the parameters of the model maintain a certain level of competitiveness.

REFERENCES

[1] Xuan Chen, Zijian Wang, Qianzi Shen, Yanting Zhang, and Cairong Yan, "Mccvm: Multi-scale cross-axes conv-vmamba for medical image classification," in International Joint Conference on Neural Networks, 2025, pp. 1-8.

[2] Yubiao Yue and Zhenzhang Li, "Medmamba: Vision mamba for medical image classification," arXiv preprint arXiv:2403.03849v5, 2024.

[3] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B. Shokouhi, and Ahmad Ayatollahi, "Medvit: A robust vision transformer for generalized medical image classification," Computers in Biology and Medicine, vol. 157, pp. 106791–106801, 2023.

[4] Baiying Lei, Feng Jiang, Feng Zhou, Dong Ni, Yuan Yao, Siping Chen, and Tianfu Wang, "Hybrid descriptor for placental maturity grading," Multimedia Tools and Applications, vol. 79, no. 1, pp. 21223–21239, 2020.

[5] Omar S. EL-Assiouti, Ghada Hamed, Dina Khattab, and Hala M. Ebied, "Hdkd: Hybrid data-efficient knowledge distillation network for medical image classification," Engineering Applications of Artificial Intelligence, vol. 138, pp. 109430–109442, 2024.

[6] Candelaria Mosquera, Luciana Ferrer, Diego H. Milone, Daniel Luna, and Enzo Ferrante, "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance," European Radiology, vol. 34, no. 12, pp. 7895-7903, 2024.

[7] Stenzel Cackowski, Emmanuel L. Barbier, Michel Dojat, and Thomas Christen, "Imunity: A generalizable vae-gan solution for multicenter mr image harmonization," Medical Image Analysis, vol. 88, pp. 102799–102807, 2023.

[8] Mayank V. Golhar, Taylor L. Bobrow, Saowanee Ngamruengphong, and Nicholas J. Durr, "Gan inversion for data augmentation to improve colonoscopy lesion classification," IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 6, pp. 3864-3873, 2025.

[9] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "Smote: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 324–357, 2002.

[10] Hongwei Ding, Nana Huang, and Xiaohui Cui, "Leveraging gans data augmentation for imbalanced medical image classification," Applied Soft Computing, vol. 165, pp. 112050–112062, 2024.

[11] Ekram Chamseddine, Nesrine Mansouri, Makram Soui, and Mourad Abed, "Handling class imbalance in covid-19 chest x-ray images classification: Using smote and weighted loss," Applied Soft Computing, vol. 129, pp. 109588–109601, 2022.

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of Conference on Computer Vision and Pattern Recognition. IEEE, 2018, pp. 4510–4520.

[13] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011–2023, 2020.

[14] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni, "MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification". Scientific data, vol. 10, pp. 1-10, 2023.

[15] Andre G.C. Pacheco, Gustavo R. Lima, Amanda S. Salom˜ao, et al., "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," Data in Brief, vol. 32, pp. 106221–106230, 2020.

[16] Xavier P. Burgos-Artizzu, David Coronado-Guti´errez, Brenda Valenzuela-Alcaraz, et al., "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," scientific reports, vol. 10, no. 1, pp. 10200–10211, 2020.

[17] The International Skin Imaging Collaboration (ISIC) Website, "Isic 2018: Skin lesion analysis towards melanoma detection," 2018, [Online]. Available:https://www.kaggle.com/datasets/huyf4007/isic2018.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 9992-10002.

[19] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A ConvNet for the 2020s." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022): 11966-11976.

[20] Sachin Mehta and Mohammad Rastegari, "Separable self-attention for mobile vision transformers," Trans. Mach. Learn. Res., vol. 2023, 2023.

[21] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. "EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications." ECCV Workshops (2022).

[22] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö. Arik, and Tomas Pfister. "Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding." AAAI Conference on Artificial Intelligence (2021).

[23] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan, "Mobileone: An improved one millisecond mobile backbone," in Proceedings of CVPR. IEEE, 2023, pp. 7907–7917.

[24] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou, "Going deeper with image transformers," in Proceedings of ICCV, 2021, pp. 32–42.

[25] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu, "Vmamba: Visual state space model," in Proceedings of the 38th International Conference on Neural Information Processing System, 2025, vol. 3273, pp. 103031-103063.

[26] Omid Nejati Manzari, Hojat Asgariandehkordi, Taha Koleilat, Yiming Xiao, and Hassan Rivaz, "Medical image classification with kan-integrated transformers and dilated neighborhood attention," Applied Soft Computing, vol. 186, pp. 114045-114055, 2026.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in Proceedings of CVPR, 2016, pp. 770-778.

[28] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter, "Auto-sklearn: Efficient and Robust Automated Machine Learning," in Proceedings of the 29th International Conference on Neural Information Processing Systems, 2015, vol. 2, pp. 2755–2763.

[29] Haifeng Jin, Qingquan Song, and Xia Hu, "Auto-keras: An efficient neural architecture search system," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 1946–1956.