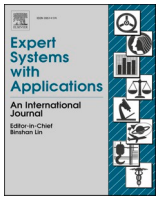
Expert Systems With Applications 275 (2025) 127145

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Multiple teachers are beneficial: A lightweight and noise-resistant student model for point-of-care imaging classification

Yucheng Song [a,1], Anqi Song [a,2], Jincan Wang [a,3], Zhifang Liao [a,*,4]

[a] School of Computer Science and Engineering, Central South University, Changsha, China

ARTICLE INFO

Keywords:
Medical image classification
Noisy label
Multi-teacher knowledge distillation
Deep learning
Point-of-care

ABSTRACT

In recent years, the development of medical imaging technology has shifted imaging solutions from laboratories to point-of-care imaging using instant imaging. However, these point-of-care devices are often limited by environmental factors such as ambient light and noise, resulting in poor imaging quality, which in turn affects the diagnostic accuracy of point-of-care devices. Additionally, since nursing devices require more lightweight models, traditional models cannot meet the requirements in terms of computing resources, model parameters, and inference time. Therefore, in response to these issues, this paper proposes a lightweight student model that optimizes residual information. A lightweight structure based on Shift MLP is designed on the residual branch of the model to enhance the model's ability to capture multi-scale spatial feature information. At the same time, we design a multi-teacher distillation strategy to improve the accuracy and noise resistance of the student model. We first propose an adaptive learning method based on auxiliary teachers, which utilizes unlabeled and noisy data for adaptive learning to enhance the model's robustness. Then, we design a global teacher model to improve the accuracy of the student model and indirectly enhance the auxiliary teaching ability of the auxiliary teacher model, thereby achieving global knowledge transfer. We evaluated our approach on three public medical image classification datasets. The results show that compared with existing state-of-the-art methods, we reduced the number of parameters by 38 times and the computation by 11 times, with an inference time of only 18.94 ms on a CPU. Our lightweight student model not only significantly reduces the number of parameters and computational complexity but also maintains competitive classification accuracy. These results demonstrate that our proposed model can achieve high-precision classification while meeting the requirements of lightweight and efficient deployment in point-of-care imaging devices, providing a more practical solution for medical image classification tasks in clinical settings. Our Github code link is: https://github.com/wangprocess/Mul-Teachers-KD.

1. Introduction

Medical images play a key role in the diagnosis and treatment of the healthcare sector (Li et al., 2023). A major task in medical imaging is diagnostic classification, as it is crucial for assisting doctors in diagnosis, disease prediction, and treatment planning (Liao et al., 2023). Currently, a large number of researchers have developed powerful deep learning models to cope with the complexity and diversity of medical imaging data, and have shown strong capabilities in disease diagnosis and classification tasks (Yuan et al., 2023; Zeng et al., 2024). For example, Sivapriya et al. (Sivapriya et al., 2024) proposed a new ResEAD2Net with two encoding and decoding stages to preserve spatial and spectral domain features for better classification of diabetic retinopathy. Zeng et al. (Zeng, 2018) developed a deep discriminative autoencoder network that learns site-shared brain connectivity features through deep learning methods for cross-site classification of schizophrenia. Veeramani et al. (Veeramani et al., 2024) proposed a new convolutional neural network model DM-CNN, which mainly consists of four sub-modules: Dynamic Multi-scale Feature Fusion (DMFF), Hierarchical Dynamic Uncertainty Quantization Attention (HDUQ-Attention),

* Corresponding author.
E-mail addresses: 234703024@csu.edu.cn (Y. Song), 8202220712@csu.edu.cn (A. Song), 8209210530@csu.edu.cn (J. Wang), zfliao@csu.edu.cn (Z. Liao).
[1] 0009-0009-4987-1783.
[2] 0009-0008-6181-5795.
[3] 0009-0007-9928-980X.
[4] 0000-0002-5525-904X.

https://doi.org/10.1016/j.eswa.2025.127145
Received 16 December 2024; Received in revised form 19 February 2025; Accepted 2 March 2025
Available online 3 March 2025

Multi-scale Fusion Pooling (MF Pooling) and Multi-Objective Loss (MO Loss). They use the information of multiple modules to improve the accuracy of medical image classification. Tan et al. (Tan et al., 2024) proposed a Self-Supervised Learning and Self-Distillation Method for COVID-19 (SSSD-COVID) medical image classification. In addition to calculating the reconstruction loss of mask image patches, SSSD-COVID also calculates the self-distillation loss of the latent representations of the encoder and decoder outputs to improve the overall classification effect of the model. Note that almost all of the above works focus on improving the performance of the model, but pay less attention to model complexity, inference time, or the number of parameters. In addition, most models are used for analysis in laboratory settings, so they are tested using machines with high computing power (such as GPUs). The above laboratory environment helps to speed up inference and accommodate a large number of parameters, but it also means that the requirements of lightweight and timeliness cannot be met.

In recent years, medical imaging solutions have shifted from laboratories to point-of-care imaging in bed-side settings. Point-of-care imaging refers to the process of performing instant imaging using point-of-care devices at the site of healthcare or at the patient's bedside. This imaging method moves medical imaging technology into clinical practice, enabling doctors to obtain images immediately at the patient's bedside or clinical site without the need to transfer the patient to a location where specialized imaging equipment is located (Vashist, 2017; Rasheed et al., 2024). Currently, a large number of point-of-care devices are being used in clinical settings. For example, Point-Of-Care Ultrasound (POCUS) has been used for imaging in emergency departments, wards, preoperative and post-anesthesia care units, and ICU environments, providing basic information for rapid, real-time diagnosis and evaluation of treatment responses in patients with life-threatening diseases (Grotberg et al., 2024; Huang & Palmeri, 2024). With the development of magnetic resonance imaging technology, point-of-care magnetic resonance imaging has also been used for bedside operations and rapid analysis (Samardzija, 2024). At the same time, images based on smartphone cameras can also be used for instant diagnosis and analysis of skin diseases (Khan et al., 2024). The above-mentioned point-of-care imaging helps improve the quality and efficiency of medical services and enhance patients' treatment experience and satisfaction by integrating functions such as instant diagnosis. However, point-of-care imaging devices are usually used at the clinical site or at the patient's bedside and may be affected by ambient light and environmental noise. Specifically, ambient light can cause uneven illumination in medical images. For example, in bedside skin lesion imaging using smartphone cameras, varying light intensities can lead to overexposed or underexposed areas, obscuring the true characteristics of the lesions. Environmental noise, including electrical interference and background clutter, can introduce random pixel − level fluctuations. In ultrasound images, this noise can make it difficult to distinguish between normal and abnormal tissue boundaries, reducing the image's contrast and clarity. These interferences may affect the contrast and clarity of the image, resulting in reduced image quality (Dutta, 2019). At the same time, point-of-care imaging devices are usually required to be lightweight and portable, so device performance is sacrificed during device design to meet the requirements of portability and immediacy, which results in the device being unable to run complex models with large numbers of parameters. These limitations in existing work pose significant challenges for the practical application of medical image classification in point − of − care scenarios. The high − complexity models cannot be directly deployed on resource − constrained devices, and their lack of robustness to environmental noise and variable lighting conditions further restricts their effectiveness. As a result, there is an urgent need for a new model that can not only achieve high − accuracy classification but also adapt to the harsh environmental and resource − limited conditions of point − of − care devices. This new model should be lightweight enough to run efficiently on these devices without sacrificing too much accuracy, and it should be able to handle noisy and low − quality images to ensure reliable diagnosis in real − world clinical settings.

To address the above issues, we developed a comprehensive solution. First, we designed a lightweight student model with a novel architecture. The key innovation lies in the use of a Shift MLP − based structure on the model's residual branch. This design serves a dual purpose: it reduces the model's complexity by minimizing the number of convolutional layers, while simultaneously enhancing its ability to capture multi − scale spatial feature information. By doing so, the model can efficiently extract relevant features from medical images, even with a reduced number of parameters. Second, we introduced a multi − teacher distillation strategy. Given the challenges of noisy data and the need for high − accuracy classification, this strategy plays a crucial role. We first proposed an adaptive learning method based on auxiliary teachers. This method enables the model to leverage unlabeled and noisy data for training, improving its robustness. Then, we designed a global teacher model. The global teacher transfers deep − feature knowledge to the student model, enhancing its classification accuracy. Additionally, the knowledge transfer from the global teacher to the auxiliary teacher indirectly improves the auxiliary teacher's teaching ability, forming a comprehensive knowledge − transfer framework. This integrated approach effectively combines lightweight design, noise resistance, and high − accuracy classification, providing a practical solution for point − of − care medical image classification.

The lightweight student model we proposed, which optimizes residual information, demonstrates significant performance enhancement. Compared with traditional models, the number of parameters is reduced by 38 times, and the computational complexity is decreased by 11 times. The inference time on a CPU is only 18.94ms, which greatly improves the efficiency of the model and enables it to be easily deployed on resource-constrained point-of-care imaging devices. Through the multi-teacher distillation strategy, the model achieves an average accuracy of 56.85% and an average AUC of 83.00% in complex noisy environments, significantly enhancing the model's accuracy and noise resistance. This provides more reliable support for clinical diagnosis. These achievements effectively address the key challenges faced by point-of-care imaging devices and have important application value and innovation in the field of medical image classification. The main contributions of this work are as follows:

1. We proposed a lightweight student model that optimizes residual information. On the residual branch of the model, a lightweight structure based on Shift MLP was designed to obtain multi-scale spatial feature information to ensure the lightweight model while maintaining good performance of the model.
2. We proposed an adaptive learning method based on auxiliary teachers, using unlabeled and noisy data for adaptive learning to improve the robustness of the model.
3. We designed a multi-teacher distillation strategy to achieve global knowledge transfer by using the global teacher model to improve the accuracy of the student model and indirectly improve the auxiliary teaching ability of the auxiliary teacher model.
4. We conducted experimental verification on multiple datasets. The experimental results show that the proposed lightweight student model has better robustness and faster inference time with almost no loss of accuracy.

The remainder of the paper is organized as follows: In Section II, we introduce the proposed teacher model, student model, and the distillation learning process. In Section III, we analyze the experiments and their results. In Section IV, we discuss the model and the process. Finally, in Section V, we provide a conclusion.

2. Relate work

Since the discovery of neural networks, the exploration of network

architectures has been an essential part of their research (Mazumdar et al., 2023). Among these, knowledge distillation, multi-teacher distillation, and lightweight network structures have played a crucial role in medical image classification and clinical applications.

2.1. Knowledge distillation

Knowledge distillation was introduced by Hinton et al. (Hinton et al., 2015) to train a smaller student model to achieve performance comparable to that of a larger teacher model by using soft targets and feature alignment. In recent years, knowledge distillation has been widely applied in the field of medical imaging. For instance, Gou et al. (Gou et al., 2021) reviewed the application of knowledge distillation in medical image classification and pointed out that feature-matching-based distillation can effectively enhance the generalization ability of the student model. Additionally, Tian et al. (Tian et al., 2022) proposed a distillation method based on contrastive learning for lung CT image classification, which significantly improved the robustness of the model. Liu et al. (Liu et al., 2025) introduced multi-layer feature distillation, which aligns features at different network levels to improve the performance of medical image segmentation tasks.

2.2. Multi-teacher learning strategy

Multi-teacher distillation enhances the learning ability and generalization of the student model by integrating knowledge from multiple teacher models. Sarıyıldız et al. (Sariyildiz et al., 2024) combined the complementary strengths of multiple teacher models to train a universal encoder that performs well across multiple tasks. In the context of medical image segmentation, Gu et al. (Gu, 2025) proposed Dual Structure-Aware Image Filtering (DSAIF) for semi-supervised medical image segmentation. By utilizing the dual contrast-invariant representations of Max-tree and Min-tree, DSAIF generates two images with different appearances but preserving the original image's topological structure. This approach reduces overfitting to erroneous pseudo-labels and improves segmentation performance. Mekonnen et al. (Mekonnen, 2024) combined diffusion models with Generative Adversarial Networks (GANs) and used knowledge distillation to achieve more efficient training and evaluation, reducing the number of parameters and denoising steps and speeding up the sampling process.

2.3. Lightweight model design

Driven by the real-time requirements of medical imaging processing, researchers have been dedicated to designing lightweight neural networks to reduce computational overhead. Sandler et al. (Sandler et al., 2018) proposed MobileNetV2, which employs depthwise separable convolutions and inverted residual structures to reduce computational costs while maintaining performance. Howard et al. (Howard, 2019) further advanced this with MobileNetV3, which integrates neural architecture search to optimize the network structure and enhance adaptability. Meanwhile, Transformer architectures have been extensively studied for lightweight optimization. For example, Mehta et al. (Mehta & Rastegari, 2022) introduced MobileViT, which combines CNN and Transformer structures to reduce computational complexity while retaining the global modeling capabilities of Transformers. Additionally, Qiu and Shi (Qiu & Shi, 2023) proposed L-MLP, which uses a lightweight hybrid MLP structure and has achieved good performance in medical image segmentation tasks.

Knowledge distillation is a widely used model compression method in deep learning, which optimizes the performance of a smaller student model by having it learn from the knowledge of a larger teacher model. However, the knowledge from a single teacher may be limited, affecting the generalization ability of the student model. To address this, researchers have proposed multi-teacher distillation strategies to enhance the learning capabilities of the student model. Even with multi-teacher distillation, highly lightweight models are still required in scenarios such as Point-of-Care (Dutta, 2019) imaging to meet real-time requirements. Therefore, further exploration of lightweight model design is necessary to reduce computational complexity while maintaining high task performance.

3. Method

This section provides a detailed introduction to the models and strategies designed to address the challenges associated with point-of-care imaging devices. By elaborating on the lightweight student model, the global teacher model, the multi-teacher knowledge distillation strategy, and the loss function, this section lays the theoretical foundation for the subsequent experimental validation and result analysis.

We designed a lightweight student model and a global teacher model respectively, and proposed a multi-teacher distillation strategy to train the student model. Specifically, we first designed a lightweight student model with fewer convolutional layers, and optimized the residual information of the student model through a lightweight MLP structure. Then we designed a powerful global teacher model that achieves powerful feature information extraction capabilities through dual attention. The multi-teacher distillation strategy we proposed transfers knowledge to the student model, so that the model can not only acquire knowledge from the global teacher, but also robustly use noisy unlabeled data for adaptive learning. The overall process of our distillation strategy is shown in Fig. 1.

3.1. Lightweight student model for optimizing residual information

This section focuses on the design of the lightweight student model, which aims to reduce the number of parameters and computational complexity while maintaining performance by optimizing residual information extraction and designing a specific MLP structure to meet the deployment requirements of point-of-care devices.

This paper proposes a lightweight student model with only 0.18M parameters to ensure that it can be easily deployed on point-of-care imaging devices. To ensure the lightweight of the model, we have made the following designs: 1) Use fewer convolutional layers to avoid too many parameters in the model. 2) Optimize the feature extraction process of the model's residual information. By designing a lightweight MLP structure, multi-scale spatial information extraction is performed on the residual information of each layer, avoiding the performance degradation caused by the low depth of the model. The overall structure of the model is shown in Fig. 2.

3.1.1. Optimizing residual feature information extraction

Since the residual information still contains feature information of different scales and complexities (Ye et al., 2020), we optimized the feature extraction process of the residual. Specifically, we used three convolution blocks in the feature extraction stage, each of which contains a convolution layer, a batch normalization layer, and a maximum pooling layer with a window pool size of 2×2. Reducing the number of convolutional layers is a key step in optimizing the residual information while maintaining the model's lightweight nature. Convolutional layers are computationally expensive and contribute significantly to the model's parameter count. By using fewer convolutional layers, we reduce the overall complexity of the model. However, this could potentially lead to a loss of feature − extraction ability. To address this, we focus on optimizing the residual information. Residual information often contains valuable features that can be exploited to enhance the model's performance. When we reduce the number of convolutional layers, the residual connections become even more important. They allow the model to preserve the information that might be lost due to the reduction in convolutional layers. We use the designed lightweight MLP structure to extract features from the residual information of each

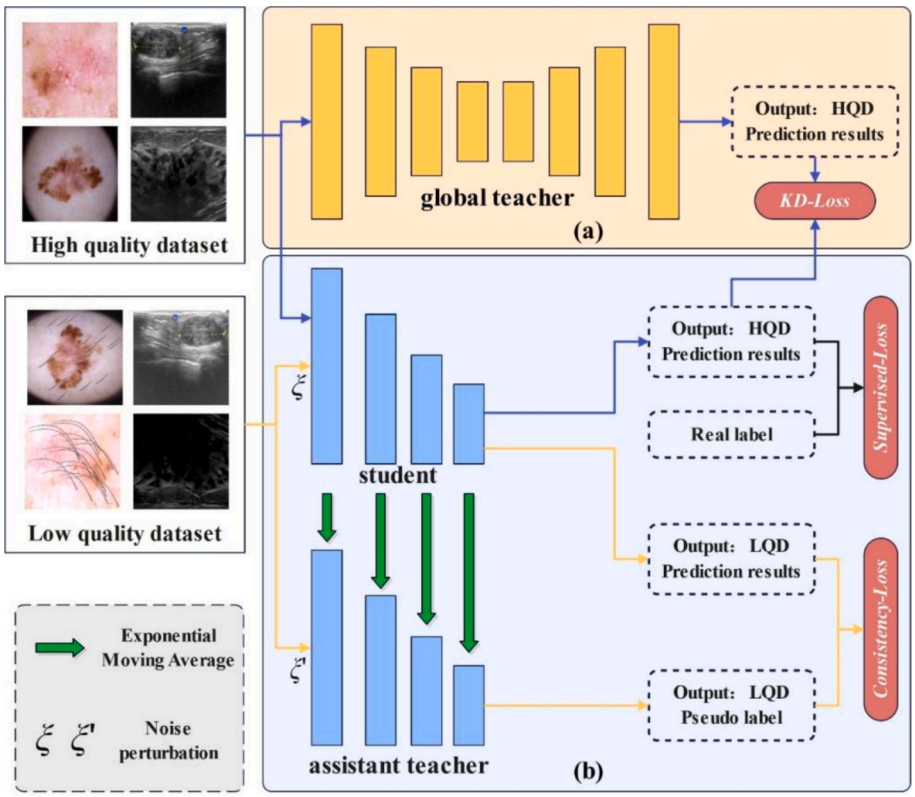


Fig. 1. Overall flow chart of the multi-teacher knowledge distillation strategy. (a)The knowledge transfer process of the global teacher model. (b)The adaptive knowledge transfer process of the auxiliary teacher.

convolution block, and use linear interpolation to unify the scale of the output information of each part, and finally merge the input into the classification layer for the result output. The optimized residual feature extraction process is described as follows.

$$D_i = MP^i\left(Relu\left(BN\left(conv^i(D_{i-1})\right)\right)\right) \tag{1}$$

$$E_i = interp^i(M_{MLP}(D_i)) \tag{2}$$

where D_i represents the output of the i-th convolutional block, E_i represents the output of the i-th convolutional block, MP^i represents the maximum pooling operation of the i-th convolutional block, $conv^i$ represents the convolution operation of the i-th layer, $interp^i$ represents the linear interpolation operation of the i-th layer, BN represents batch normalization, Relu represents the rectified linear unit, and $M_{MLP}(\cdot)$ represents the lightweight MLP stage. The specific operations will be described in detail in the next section.

3.1.2. Lightweight MLP structure

In order to enhance the nonlinear expression ability and representation ability of the network while keeping the model parameter count low, we designed a lightweight structure based on MLP for the residual of the model (as shown in the right half of Fig. 2). For each Block block, specifically, in order to make the MLP focus on the position information under the feature channel, we first perform patch embedding on the feature data input to the MLP stage to extract local features and increase the receptive field, and then move the axis of the feature channel to make the network focus on the local information in the position block. Inspired by the Swin Transformer (Liu, 2021), we perform feature shift in the direction of the width and height axes and input them into the respective MLP layers in turn. To better encode positional information, we inserted a convolution layer between the MLPs across the width and height. Additionally, to address the gradient vanishing problem caused by increasing model depth, we also added residual connections between each MLP Block. The calculation process can be summarized as follows:

$$X_s = Shift_W(X) \tag{3}$$

$$h = conv(MLP(X_s)) \tag{4}$$

$$Y_s = Shift_H(h) \tag{5}$$

$$Y = LN(X + MLP(Y_s)) \tag{6}$$

where X represents the input feature vector, $Shift_W(\cdot)$ and $Shift_H(\cdot)$ represent the feature shift of width and height, and LN represents layer normalization. The lightweight MLP structure is designed for efficient extraction of multi − scale spatial features with a low parameter count. This can reduce dimensionality and focuses on local features. Feature shift operations in the width ($Shift_W$) and height ($Shift_H$) directions, implemented via circular padding, capture different spatial relationships. For example, in width − shift operation on an input tensor $X \in R^{H\times W\times C}$, elements in the width dimension are shifted, akin to sliding a window of size W with circular edge shifting. The same applies to the height − shift operation. A 3×3 convolutional layer with stride 1 is inserted between MLPs across width and height. It captures local spatial correlations after shift operations and extracts fine-grained spatial information from width − shifted features. This information is then processed by the MLP for height − shifted features, enhancing multi − scale spatial feature extraction.

Compared with other structures, our Shift MLP structure demonstrates superior performance in lightweight models. Traditional convolutional layers, widely used in many models, typically require a large number of parameters to capture multi-scale features. For example, standard convolutional layers with large kernel sizes can capture global features but at the cost of high computational burden and a large number of parameters. In contrast, our Shift MLP structure achieves

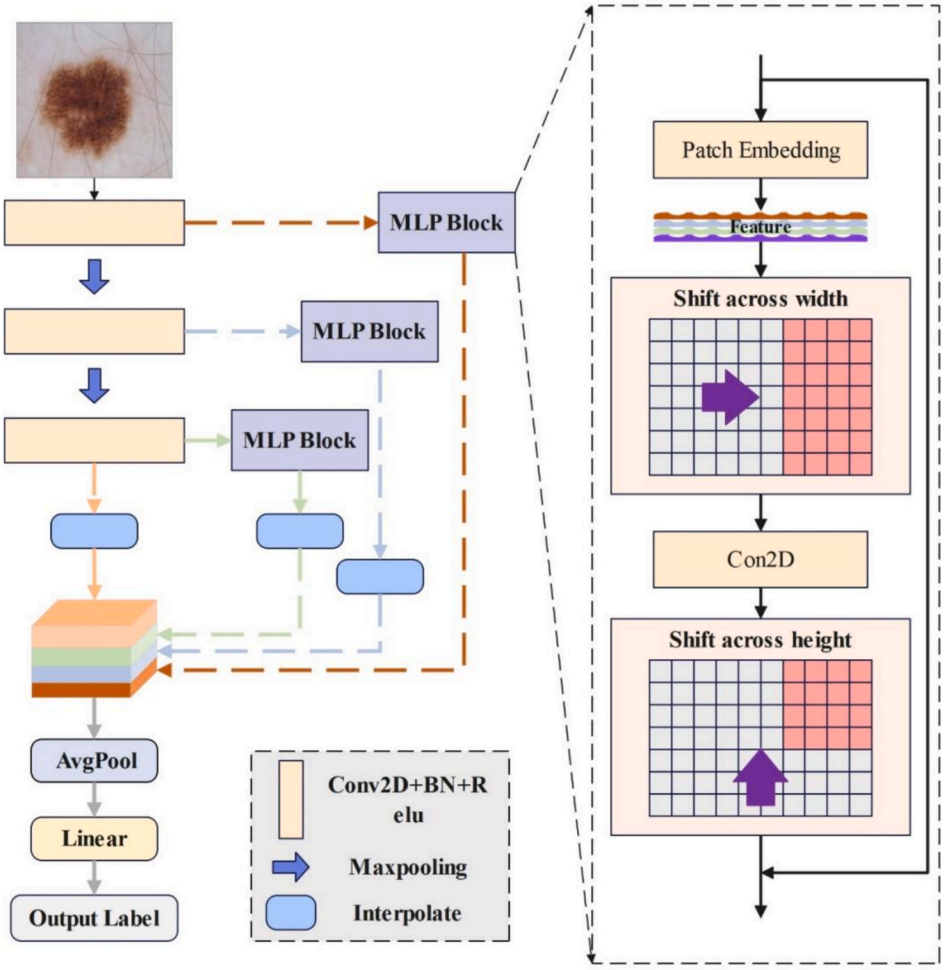


Fig. 2. Overall structure of the lightweight student model that optimizes residual information.

multi-scale feature capture through operations such as patch embedding and feature shifting. It can significantly reduce the number of parameters while maintaining a high level of performance.

3.2. Transformer global teacher model based on dual attention mechanism

This section introduces the global teacher model based on a dual-attention mechanism. The goal is to utilize a complex network

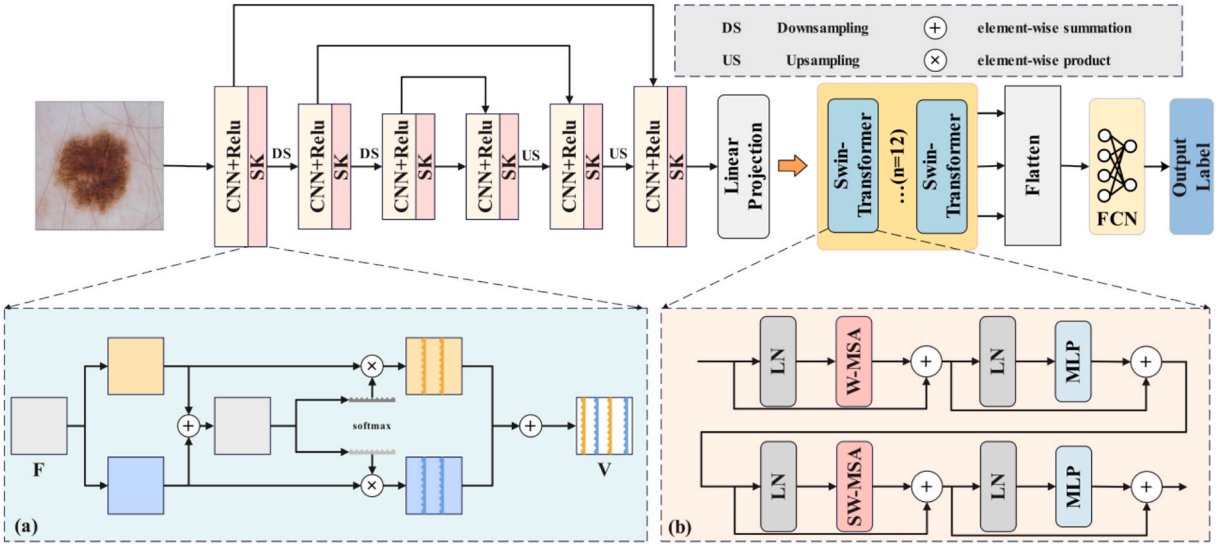


Fig. 3. Transformer teacher model based on dual attention mechanism. (a) Selecting the core attention mechanism. (b) Swin-Transformer structure details.

structure to achieve powerful feature extraction capabilities, thereby providing high-quality knowledge to the student model and enhancing its accuracy.

This paper proposes a global teacher model based on a dual attention mechanism, which achieves powerful feature information extraction capabilities through a complex network structure. The model is divided into two parts: 1) The shallow network uses a selective kernel attention mechanism to better capture the multi-scale features of complex image spaces. 2) In the deep network, we use a more powerful Transformer structure to more comprehensively capture the global information in the feature map. The overall structure of the model is shown in Fig. 3.

3.2.1. Adaptive extraction of multi-scale spatial information

In the shallow network part of the teacher model, due to the up-sampling and down-sampling process, different network layers have feature maps of different scales. Since receptive fields of different sizes have different effects on objects of different scales, in order to better capture the multi-scale features of complex image spaces, we use the Selective Kernel Attention (SKAttention) mechanism (Li et al., 2019) in each sampling stage. SKAttention has two main steps: feature fusion and attention weight generation. First, in the feature fusion stage, for the input feature map X, we use n convolution kernels of different sizes to obtain different feature maps $F_1, F_2, F_3, \cdots F_n$, and then pass through a 1×1 convolution fusion layer to obtain the fused feature map $F = [f_1, f_2, f_3, \ldots, f_n]$, where $f_1, f_2, f_3, \ldots, f_n$ are the fusion results of $F_1, F_2, F_3, \cdots F_n$ respectively. Then, in the attention weight generation stage, the feature dimension is reduced to $\frac{c}{r}$ through the fully connected layer, and then the dimension is restored to the original number of channels C through the full connection, and the weight $\alpha = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n]$ corresponding to the feature map is generated through the *Softmax* function, and finally the output feature map o is obtained by weighted fusion with the original features $F_1, F_2, F_3, \cdots F_n$. The specific process is described as follows.

$$F = Conv_{1*1}(F_1, F_2, F_3, \cdots F_n) \tag{7}$$

$$\alpha = softmax\left(FC_{\frac{C}{r}}(FC_C(F))\right) \tag{8}$$

$$o = \sum_{i=1}^{n} \alpha_i \cdot F_i \tag{9}$$

Among them, *Conv* represents 1×1, *FC* represents the fully connected layer, C represents the number of channels, r represents the reduction ratio, and *softmax* represents the *softmax* activation function.

The selective kernel attention mechanism (SKAttention) is designed to adaptively select the most relevant receptive fields for different scale objects in the image. This is crucial in medical image analysis, where lesions and anatomical structures can vary greatly in size. For example, small micro − lesions in skin or internal organs may require a small receptive field to capture their details, while larger tissue-level changes need a larger receptive field. SKAttention's feature fusion step, as described in formula (7), combines feature maps from different-sized convolution kernels. By doing so, it aggregates information at various scales. The 1×1 convolution fusion layer not only reduces the dimensionality but also combines the features in a way that preserves their scale − related information. In the attention weight generation step (formula (8)), reducing the feature dimension and then restoring it through fully-connected layers helps in capturing the global relationship between different − scale features. The Softmax function then generates weights that represent the importance of each scale's feature map. This ensures that the model can focus on the most relevant scale for each part of the image.

When integrating SKAttention into the Transformer-based global teacher architecture, it enhances the model's ability to capture multi-scale features at an early stage. In the shallow network, SKAttention processes the input image before the data enters the deeper Transformer layers. This pre-processed data, with multi-scale features already emphasized, allows the Transformer structure to better capture global information. The combination of SKAttention and the Transformer architecture enables the global teacher model to have a more comprehensive understanding of the medical image, from fine-grained local features to global context, which is essential for accurate knowledge transfer to the student model.

3.2.2. Deep information extraction based on self-attention

In a deep network, the feature map may lose some important contextual information after being transformed and combined by multiple layers of neural networks. Therefore, we added a Transformer structure to the deep network part. Through the self-attention mechanism, the network can dynamically adjust the correlation between features at each position, thereby enhancing the representation ability of the features and enabling the network to more accurately capture the semantic information of the input data. We divide the feature map $F \in R^{H\times W\times C}$ into a series of flattened 2D patches according to the size of P, and obtain $z_i \in R^{\frac{H}{P}\times\frac{W}{P}\times C} | i = 1, \cdots, N\}$, and then input these vectors into the Swin-Transformer block of the encoder, which consists of L layers of Windows Multi-head Self-Attention and Shifted Windows Multi-Head Self-Attention. The output of each block is as follows

$$\widehat{z}^l = W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \tag{10}$$

$$z^l = MLP\left(LN\left(\widehat{z}^l\right)\right) + \widehat{z}^l \tag{11}$$

$$\widehat{z}^{l+1} = \mathrm{SW} - \mathrm{MSA}\left(LN\left(z^l\right)\right) + z^l \tag{12}$$

$$z^{l+1} = MLP\left(LN\left(\widehat{z}^{l+1}\right)\right) + \widehat{z}^{l+1} \tag{13}$$

Where LN is Layer normalization, $\widehat{z}^l$ and z^l represent the output features of the SW-MSA module and the MLP module of block l respectively.

Our dual-attention mechanism combines selective kernel attention in the shallow network to capture multi-scale spatial features and Transformer-based self-attention in the deep network to capture global information. This combination enables the model to better adapt to the complexity of medical images, as it can consider both local and global features simultaneously.

3.3. Multi-teacher knowledge distillation strategy

This section provides a detailed description of the multi-teacher knowledge distillation strategy. By working together with the auxiliary teacher and the global teacher model, the robustness and accuracy of the student model are enhanced to address the challenges of image quality and model performance faced by point-of-care imaging devices.

As mentioned above, there are two main problems with current point-of-care imaging devices: (1) The image acquisition process is affected by ambient light and environmental noise, resulting in low model accuracy. (2) The limitations of device performance hinder the use of high-accuracy complex models. To address these problems, we propose a new multi-teacher knowledge distillation strategy that uses auxiliary teachers to improve the robustness of the student model, while using a global teacher model to improve the accuracy of the student model and indirectly improve the teaching level of the auxiliary teachers.

3.3.1. Adaptive knowledge transfer for supporting teachers

There are a large number of natural medical image resources on the Internet, especially for certain specific diseases. The quality of these image resources is uneven and lacks diagnostic labels (Abbasi, 2021). Compared with High-Quality Datasets (HQD) dedicated to machine learning with real labels, these data are undoubtedly lower in quality and considered a "burden." However, if we can utilize these Low-Quality

Data (LQD) to assist in model training, it will effectively help the model adapt to image data from various imaging environments in advance, thereby improving the model's robustness. To achieve this, we propose a semi-supervised auxiliary teacher learning method to extract adaptive knowledge from low-quality data and pass it to the student model, transforming it from a "burden" into a "treasure."

The main idea of the semi-supervised auxiliary teacher model is that after adding slight perturbations to the input data, if the current weights of the model capture the distribution characteristics of the dataset, then consistent prediction results will be obtained on the original input and perturbed data. And because the auxiliary teacher model takes into account the information of the entire training history, it can better capture the distribution characteristics of the data and has better generalization ability. Therefore, we believe that its prediction results can be used as pseudo labels. The adaptive knowledge of LQD is passed to the student model, encouraging the student model to obtain prediction results consistent with the auxiliary teacher model. Specifically, the student model and the auxiliary teacher model have the same network structure. We record the weight of the student model at training step t as θ_t, and the weight of the auxiliary teacher model as θ'_t. The weight of the auxiliary teacher model is frozen and updated using the Exponential Moving Average (EMA), where the exponential decay rate is α. In this way, part of the weight of the teacher model comes from its own historical weight, and the other part comes from the current weight of the current student model. The update process is shown in formula (14).

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \tag{14}$$

By adding perturbations ξ and ξ' to two identical input data x, and inputting them into the student model and the auxiliary teacher model respectively, a consistency loss L_{CO} is constructed for the student model prediction and the pseudo label.

$$L_{CO} = E_{x,\xi,\xi'}\left[\left|\left|f\left(x,\theta'_t,\xi'\right) - f(x,\theta_t,\xi)\right|\right|^2\right] \tag{15}$$

where $f(\cdot)$ represents the lightweight model. $E_{x,\xi,\xi'}$ represents the expected distance of the prediction results under different perturbations of the input data.

3.3.2. Knowledge transfer from the global teacher model

In order to make the accuracy of the lightweight student model on the point-of-care imaging device close to that of the complex model, we proposed a Transformer global teacher model based on the dual attention mechanism. The globality of this model is mainly reflected in two aspects: (1) The global model first transfers the deep feature knowledge it extracts to the student model to achieve direct knowledge transfer. (2) The student transfers the knowledge of the global model to the auxiliary teacher model through EMA to improve its auxiliary teaching ability, so that the auxiliary teacher generates pseudo labels containing more adaptive knowledge and provides more accurate guidance information.

The teaching method of the global teacher model is soft label distillation, which allows the student model to imitate the probability distribution of the prediction results of the global teacher model as much as possible, so that the student model can learn the decision boundary of the teacher model and capture the category distribution characteristics of the data from it, thereby improving the accuracy of the student model. We first soften the output $z = [z_1, z_2, z_3, \ldots, z_n]$ of the teacher model by the distillation temperature T, and adjust the hyperparameter T to 2, converting the output into a smoother probability distribution $q = [q_1, q_2, q_3, \ldots, q_n]$, so that students can learn category knowledge in addition to the true value. The softening calculation process is as follows:

$$q_i = \frac{e^{\frac{z_i}{T}}}{\sum_j e^{\frac{z_j}{T}}} \tag{16}$$

We transfer knowledge to the student model by constructing a distillation loss between the global teacher model and the student model. Specifically, we use the consistency of the student model's prediction results for the training data and the soft labels of the global teacher model to construct the loss so that their probability distributions are similar. The calculation process is as follows:

$$L_{KD} = -\sum_j p_j \log\left(q_j\right) \tag{17}$$

Regarding the interaction between the global teacher, student, and auxiliary teacher models. The global teacher model extracts high-level features from the high-quality dataset using its dual-attention mechanism. The soft-label distillation process transfers this knowledge to the student model. As the student model updates its weights based on the distillation loss L_{KD}, it affects the auxiliary teacher model. The student model's updated weights are used to update the auxiliary teacher model's weights via Exponential Moving Average (EMA). This enables the auxiliary teacher model to incorporate knowledge from the global teacher, generating more accurate pseudo-labels for low-quality data. The updated pseudo-labels help the student model improve its robustness to noisy and low-quality images. This cyclic interaction among the three models continues throughout training, enhancing the student model's accuracy and noise resistance.

3.4. Loss function

This subsection primarily focuses on constructing the loss functions required for model training. By integrating the consistency loss from the auxiliary teacher, the distillation loss from the global teacher model, and the supervised loss, the learning process of the model is effectively guided to ensure the optimization of model performance.

In addition to the auxiliary teacher consistency loss L_{CO} and the global teacher model distillation loss L_{KD} mentioned above, we also construct a supervision loss LS between the student model prediction result and the true label to represent the self-learning process of the student model. The supervision loss LS is a weighted integration of the cross entropy loss LCE and the focus loss LFL. β is an adjustable hyperparameter. The calculation process is as follows:

$$LS = \beta LCE + (1-\beta)LFL \tag{18}$$

Since our auxiliary teacher model is derived from the student model EMA, it needs to be "warmed up" to ensure the reliability of adaptive knowledge. Therefore, our total loss function is divided into two sections according to the training step of the model:

$$L = \begin{cases} (1-\lambda)*L_S + \lambda*L_{KD}, step < N \\ (1-\lambda)*L_S + \mu*L_{CO} + (\lambda-\mu)*L_{KD}, step \geq N \end{cases} \tag{19}$$

Where λ and μ are adjustable loss weights, and N is the number of training iterations at which the auxiliary teacher starts to participate in teaching. The pseudocode description of the training process of the multi-teacher distillation strategy is shown in Algorithm 1.

Algorithm 1: Multi-teacher distillation strategy training process

Input: training dataset $\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^{N}$, Where $\boldsymbol{x_i}$ is the image,$\boldsymbol{y_i}$ is the label, the learning rate ***base_lr***, the distillation temperature $\boldsymbol{T}$, the consistency weight $\boldsymbol{\alpha}$, the EMA decay rate ***ema_decay***, the maximum number of iterations ***max_iterations***, the auxiliary teacher starts training round $\boldsymbol{N}$, the number of training rounds $\boldsymbol{E}$, and the global teacher model $\boldsymbol{M_t}$.

Output: Student Model $\boldsymbol{M_s}$ and MEA Model $\boldsymbol{M_e}$.

```
Loading the global teacher model M_t.
Initialize the weight of M_s, initialize M_s → M_e.
For train_epoch in E:
  For batch(X, Y) in D:
    X→ unlabeled X_u
    Adding noise ∈ : X'_u = X_u + ∈
    O_s = M_s(X)
    O_t = M_t(X)
    O_e = M_e(X'_u)
```

(continued on next page)

(*continued*)

Algorithm 1: Multi-teacher distillation strategy training process

Input: training dataset $\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^{N}$, Where $\boldsymbol{x_i}$ is the image,$\boldsymbol{y_i}$ is the label, the learning rate ***base_lr***, the distillation temperature $\boldsymbol{T}$, the consistency weight $\boldsymbol{\alpha}$, the EMA decay rate ***ema_decay***, the maximum number of iterations ***max_iterations***, the auxiliary teacher starts training round $\boldsymbol{N}$, the number of training rounds $\boldsymbol{E}$, and the global teacher model $\boldsymbol{M_t}$.

Output: Student Model $\boldsymbol{M_s}$ and MEA Model $\boldsymbol{M_e}$.

1: Loading the global teacher model M_t.

10: $L_{ce} = CE(O_s^l, Y_l)$

11: $L_{focal} = Focal(O_s^l, Y_l)$

12: $L_{sup} = 0.3 \cdot L_{ce} + 0.7 \cdot L_{focal}$

13: $L_{KD} = KD(O_s^l, O_t^l, T)$

14: **If train_epoch >= N:**

15: $L_{cons} = \|softmax(O_s^u) - softmax(O_e)\|_2^2$.

16: **else**

17: $L_{cons} = 0$

18: $L = (1-\alpha) \cdot L_{sup} + \alpha \cdot L_{KD} + consistency\_wight \cdot L_{cons}$

19: **end**

20: Update the weight of M_s

21: Update the weight of M_e: $\theta_e \leftarrow ema\_decay \cdot \theta_e + (1 - ema\_decay) \cdot \theta_s$

22: Adjusting the learning rate:$\mathrm{lr} = \mathrm{base\_lr} \cdot \left(1 - \frac{\mathrm{iter\_num}}{\mathrm{max\_iterations}}\right)^{0.9}$

23: **end**

4. Experimental process and result analysis

This section validates and evaluates the proposed model and strategy through a series of experiments. It provides detailed descriptions of the experimental data processing, experimental setup, and in-depth analysis of the results to demonstrate the performance of the model in various aspects and verify the effectiveness of the research.

4.1. Experimental data

ISIC 2018: This dataset is a large-scale dermoscopic image classification dataset released by the International Skin Imaging Collaboration (ISIC) (Codella, 2017). It consists of 7 different categories of skin diseases, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, cutaneous fibrocarcinoma, and vascular lesions, with a total of 10,015 images. It is divided into official training set, validation set, and test set.

BUSI: This dataset was obtained from 600 female patients (aged between 25 and 75 years old) at Baheya Hospital in Cairo, Egypt (Al-Dhabyani et al., 2020). The dataset consists of 780 images, including 437 benign cases, 210 malignant cases, and 133 normal cases. The proportion of training dataset, test set and validation set is 60%, 20%, 20%.

Dermnet: This dataset contains images of 23 categories of skin diseases. The categories include acne, melanoma, Eczema, Seborrheic Keratoses, Tinea Ringworm, Bullous disease, Poison Ivy, Psoriasis, Vascular Tumors, etc. It consists of about 19,500 images of different pixel sizes in JPEG format and composed of 3 channels (Kumar et al., 2024). We selected 5 types of representative data for the experiment, among which the proportion of training dataset, test set and validation set is 60%, 20%, 20%.

We resized the images of the three datasets to 224*224 format and divided the training set into two parts in a ratio of 4:6. 40% of the data was used as HQD and 60% of the data was used to simulate natural data on the Internet. LQD was obtained by performing random image processing on the data, including adding motion blur, median filtering, Gaussian blur, Gaussian noise, and brightness transformation. Regarding augmentation techniques, noise injection was a crucial step. Gaussian noise was added to the images to simulate real − world noisy environments. In real − world medical imaging, especially in point − of − care scenarios, images are often affected by various types of noise. For example, in ultrasound imaging, electrical interference can introduce Gaussian − like noise, which distorts the original image features. By adding Gaussian noise with a mean of 0 and a standard deviation randomly selected from the range of [0, 0.05] to the training images, we mimic this real − world noise situation. This helps the model learn to be robust against such noise during training. In particular, we performed hair enhancement on the ISIC 2018 dataset to simulate the hair noise in skin disease images, which can better restore the environment of real skin disease diagnosis (as shown in Fig. 4). In addition to noise injection, we also applied other common augmentation techniques such as flipping, rotating, and scaling. Flipping the images horizontally or vertically helps the model learn to be invariant to the orientation of the objects in the image. Rotating the images by angles within the range of [-15, 15] degrees simulates different acquisition angles in real − world imaging. Scaling the images by factors between 0.8 and 1.2 mimics the variability in object distances from the imaging device. These combined preprocessing and augmentation steps not only enrich the dataset but also make the model more adaptable to the diverse and noisy conditions in real − world medical imaging.

4.2. Experimental setup

We use Accuracy, AUC, F1 Score and AUC-PR to evaluate the performance of the model. These parameters can be calculated using a confusion matrix. The formulas for these classification metrics are shown in Table 1 in order.

We also use the number of Parameters (Par), Giga Floating Point Operations per Second (GFLOPs), and the Average Inference Time (AIT) on the CPU to evaluate the size and efficiency of the model. We use the processed dataset images mentioned above for model training, set the batch size to 32, the initial learning rate to 0.001, perform 100 rounds of iterations, and adopt the learning rate decay strategy. The training will decay the learning rate if there is no performance improvement after 50 consecutive epochs. Other hyperparameters are set as follows:

The global teacher model distillation loss weight $\lambda = 0.7$. For the global teacher model distillation loss weight $\lambda = 0.7$, we tested different values in the range of [0.5, 0.9] with an increment of 0.1. A larger λ gives more importance to the knowledge transferred from the global teacher model. We found that when $\lambda = 0.7$, the student model achieved a good balance between learning from the global teacher and its own self-learning. Values smaller than 0.7 led to slower convergence in accuracy, while values larger than 0.7 made the model overly rely on the global teacher, resulting in overfitting on the training data.

The distillation temperature $T = 2$. The distillation temperature $T = 2$ was chosen after testing values from 1 to 4. A lower T value makes the soft-label distribution closer to the hard-label, while a higher T value makes it more spread out. We determined that $T = 2$ provided an optimal trade-off. It allowed the student model to learn the relative probabilities of different classes from the global teacher model effectively, enhancing its generalization ability.

The number of iterations $N = 10$ when the auxiliary teacher starts to

Fig. 4. Hair enhancement results. (a) Original image. (b) Result of adding hair and noise.

Table 1
Evaluation metrics used in experiments.

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Specificity	$\frac{TN}{TN+FN}$
F1 Score	$2\times\frac{\text{Precision}\times\text{Specificity}}{\text{Precision}+\text{Specificity}}$
AUC	Area under of the curve
AUC-PR	Area Under the Precision-Recall Curve

participate in training. The number of iterations $N = 10$ for the auxiliary teacher to start participating in training was set based on preliminary experiments. Starting the auxiliary teacher too early might introduce noisy knowledge as it has not "warmed up" enough. Starting it too late, on the other hand, would delay the model's learning from the unlabeled and noisy data. We determined that $N = 10$ was an appropriate starting point for the auxiliary teacher to contribute effectively to the training process.

The exponential decay rate $\alpha = 0.99$ of the auxiliary teacher model. The exponential decay rate $\alpha = 0.99$ for the auxiliary teacher model was selected through trial and error. A higher α means the auxiliary teacher's weights are more influenced by its previous state, providing stability. However, if α is too high, the model may not adapt well to changes in the student model. We found that $\alpha = 0.99$ maintained a good balance between stability and adaptability.

It is worth noting that we compared different hyperparameters under small batches of data to obtain the optimal and stable experimental standard deviation for cross-validation, and then conducted experiments on large samples.

The consistency loss weight μ of the auxiliary teacher is adjusted using the ramp-up strategy (Laine & Aila, 2017), which increases with the number of iterations.

During the experiment, we used Python and Pytorch libraries, used an NVIDIA 4090 GPU to train the model, calculated AIT on an AMD RYZEN 5800H CPU, and trained the network using the SGD optimizer and default parameter settings.

4.3. Experimental results

This section presents and analyzes the results of various experiments. Through ablation studies, model comparison experiments, and other approaches, the effectiveness of the model structure, teacher models, and the overall strategy are validated from different perspectives. The results intuitively demonstrate the achievements of this research in enhancing model performance.

4.3.1. Global teacher structure ablation experiment

Our global teacher model consists of two parts: a shallow network with a selective kernel attention mechanism and a deep network with a Transformer structure. To prove the effectiveness of each part, we conducted ablation experiments on the attention mechanism of the shallow network and the network structure of the deep network on the ISIC 2018 dataset and the BUSI dataset to prove the effectiveness of each part. Table 2 shows the performance comparison of the global teacher model with only the shallow network with different attention changes.

From Table 2, we can see that the global teacher model using the selected kernel attention performs better. We believe that this is because SKAttention can better capture the multi-scale features of complex image spaces, allowing the network to better focus on the attention information at each layer. At the same time, in order to further prove the effectiveness of the deep network structure of the global teacher model,

Table 2
Ablation experiment of attention mechanism of global teacher shallow network.

	ISIC 2018		BUSI	
	AUC %	ACC %	AUC %	ACC %
Global Teacher (w/o Attention)	91.80	90.01	90.63	89.14
Global Teacher (w/SENet (Hu et al., 2018)	92.43	90.51	91.15	89.41
Global Teacher (w/GCNet (Cao et al., 2019)	93.00	91.01	91.67	89.68
Global Teacher (w/CCNet (Huang et al., 2019)	93.65	91.52	92.19	89.95
Global Teacher (w/GENet (Hu et al., 2018)	94.21	92.12	92.71	90.22
Global Teacher (w/DANet (Fu et al., 2019)	94.81	92.54	93.23	90.49
Global Teacher (w/CBAM (Woo et al., 2018)	95.40	93.09	93.75	90.76
Global Teacher (ours)	**96.03**	**93.84**	**94.27**	**91.03**

we also verified the structure of the deep network, using CNN, Transformer and Swin Transformer as the structure of the deep network, and the experimental results are shown in Fig. 5. The experimental results show that the deep structure using Swin Transformer (our method) achieves better results.

4.3.2. Dual-teacher ablation experiment

The student model we proposed mainly relies on the Global Teacher (GT) model and the Auxiliary Teacher (AT) model. In order to evaluate the teaching effectiveness of these two teacher models, we conducted ablation experiments on the ISIC 2018 dataset to demonstrate their contribution to the overall performance. We tested four training cases, Student-Net w/o GT & AT, Student-Net w/o GT, Student-Net w/o AT, and Student-Net, on noisy low-quality data and noise-free high-quality data, respectively. The results are shown in Table 3.

According to the results in Table 3, Under the teaching of GT, the ACC and AUC of the student model in the noise-free environment have been greatly improved compared with no teaching at all, proving the effectiveness of the distillation of the global teacher model and successfully transferring knowledge to the student model. Under the teaching of AT, the ACC and AUC of the student model in the noisy environment have been greatly improved, proving that the auxiliary teacher allows the student model to adapt to the noisy environment and improves the robustness of the student model. In addition, the student model combining GT and AT teaching performed best in both environments, illustrating the complementarity of the teaching of the two teacher models, allowing the student to be improved in all aspects.

At the same time, we also observed that our proposed method achieved a relatively high accuracy rate of 60.19% on low-quality data (LQD) compared to some ablation cases, but this value was still lower than the performance on high-quality data (HQD). This discrepancy can be attributed to several factors. Firstly, the model's sensitivity to noise is a significant issue. Low-quality data is often contaminated with various types of noise, such as Gaussian noise, motion blur, and median filtering artifacts introduced during data processing. These noise sources can distort the original features of medical images. Although our model employs an adaptive learning mechanism using an auxiliary teacher, it still struggles to fully extract discriminative features in the presence of complex noise. Secondly, the presence of outliers in LQD also affects performance. Outliers in medical images could be images with extreme lighting conditions or mislabeled data. Since the model is trained on a mixture of data, these outliers can mislead the learning process. However, compared to other model variants in the ablation process, our method achieved a significant performance improvement, which further confirms the effectiveness of our method.

4.3.3. Shift MLP structural validity test

As the main contribution of the student model, the Shift MLP structure ensures that the model is lightweight while maintaining good performance. Therefore, we experimented with the effectiveness of the Shift MLP structure. In order to evaluate the impact of this structure on the

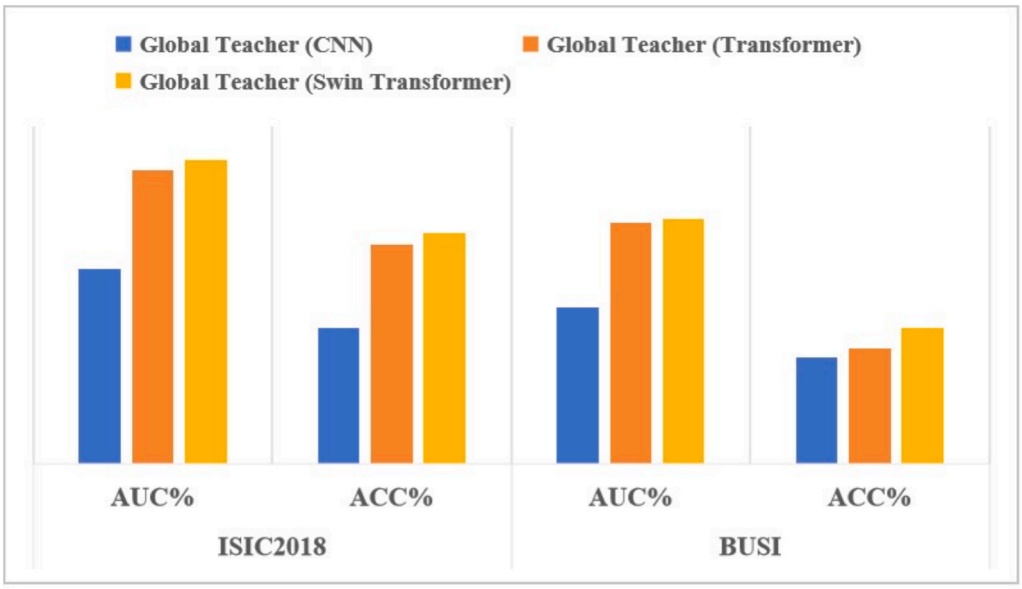


Fig. 5. Experimental results of the deep network structure of the global teacher model.

Table 3
Ablation experiment results on the ISIC 2018 dataset.

	HQD		LQD	
	AUC%	ACC%	AUC%	ACC%
Student-Net w/o GT&AT	89.75	78.76	74.82	36.44
Student-Net w/AT	90.78	79.27	84.97	49.67
Student-Net w/GT	91.11	88.11	81.58	46.97
Student-Net (ours)	**91.62**	**88.74**	**84.91**	**60.19**

student model, we applied the Shift MLP structure on different numbers of residual branches and conducted comparative experiments with the same other experimental settings. Table 4 shows the performance of the model on the test set when the number of branches is 0, 1, 2, 3, and 4. The number of branches 0 means that we do not use the Shift MLP structure.

The results show that the performance of the student model using the Shift MLP structure will be improved, and the improvement is related to the number of Shift MLP structures. The model with the Shift MLP structure added to the 4-layer residual branch has the best performance, but the improvement is not obvious. In order to avoid adding too many Shift MLP structures and introducing more complex parameters, which will make the model not lightweight enough, we chose to apply the Shift MLP structure on the 3-layer residual branch of the student model for subsequent experiments.

4.3.4. Model comparison experiment

In order to prove the advancedness of our proposed method, we conducted comparative experiments on the test sets of three datasets with existing advanced methods. The experimental results of classification accuracy are shown in Table 5.

Table 4
Performance comparison of the Shift MLP structure applied on different numbers of residual branches.

	ISIC 2018		BUSI	
	AUC%	ACC%	AUC%	ACC%
Student-Net + Shift MLP (0)	85.77	83.29	81.98	81.97
Student-Net + Shift MLP (1)	88.11	85.47	84.62	84.58
Student-Net + Shift MLP (2)	90.45	87.65	86.38	86.32
Student-Net + Shift MLP (3)	91.62	88.74	87.26	87.19
Student-Net + Shift MLP (4)	91.89	88.92	87.62	87.68

Table 5
Comparative experimental results of classification accuracy on the test sets of three datasets.

	ISIC 2018		BUSI		Dermnet	
	AUC %	ACC %	AUC %	ACC %	AUC %	ACC %
ResNet50	92.86	81.87	78.42	71.79	83.22	64.02
VIT-B-16	87.16	78.76	58.91	60.26	91.08	74.39
DenseNet121	93.47	83.42	92.45	82.05	94.60	82.01
EfficientNet-b2	93.86	83.94	**94.89**	85.90	89.32	80.79
BiFormer-B (Zhu et al., 2023)	89.64	82.66	61.27	56.41	77.81	56.40
Eff2Net (Karthik et al., 2022)	91.53	85.78	90.11	86.77	90.23	81.75
Hifuse (Huo, 2024)	90.40	85.85	76.55	70.51	84.76	62.80
DGLA-ResNet50 (Tan et al., 2024)	93.11	90.71	91.34	86.97	95.32	83.31
SRC-MT (Liu et al., 2020)	93.58	92.54	92.83	87.18	95.56	83.23
Global Teacher	**96.03**	**93.84**	94.27	**91.03**	**97.59**	**84.45**
Student-Net	91.62	88.74	87.26	87.19	87.58	79.04

We also observed differences in the model's performance across the ISIC, BUSI, and Dermnet datasets based on the results. We analyzed the primary reasons for these differences. First, the diversity of the data varies among the datasets. The ISIC 2018 dataset focuses on specific types of skin diseases, resulting in relatively homogeneous data that allows the model to easily learn discriminative features and achieve better performance. The BUSI dataset, which centers on breast ultrasound images, has unique characteristics in terms of feature extraction and classification. The Dermnet dataset, covering a wide range of skin diseases, presents more complex features and increases the difficulty of classification for the model. Second, the number and distribution of samples differ across the datasets, which affects the comprehensiveness of the features learned by the model and, in turn, influences its performance on different datasets. In order to more fully evaluate the generalization ability of our method on imbalanced datasets, we compare the F1 score and the AUC-PR evaluation metrics.

From Table 6, we can see that our method also achieved excellent performance on imbalanced datasets, thanks to the Focal Loss we used. We also visualized the comparison results, as shown in Fig. 6.

Table 6
Comparative experimental results of three test sets with classification imbalance generalization evaluation indicators.

	ISIC 2018 F1 score	AUC-PR	BUSI F1 score	AUC-PR	Dermnet F1 score	AUC-PR
ResNet50	82.31	80.12	70.12	68.21	62.12	60.01
VIT-B-16	76.42	73.52	57.42	55.32	72.12	70.01
DenseNet121	83.92	82.01	80.12	78.01	80.12	78.01
EfficientNet-b2	84.32	83.02	83.42	81.32	78.12	76.01
BiFormer-B (Zhu et al., 2023)	81.27	79.12	54.12	52.01	54.12	52.01
Eff2Net (Karthik et al., 2022)	85.23	84.12	84.23	82.12	79.12	77.01
Hifuse (Huo, 2024)	84.75	83.62	68.12	66.01	60.12	58.01
DGLA-ResNet50 (Tan et al., 2024)	89.12	87.91	84.32	82.21	81.12	79.01
SRC-MT (Liu et al., 2020)	91.23	89.32	85.12	83.01	81.12	79.01
Global Teacher	**92.12**	**90.84**	**88.12**	**86.01**	**82.12**	**80.01**
Student-Net	87.34	86.12	85.12	83.01	76.12	74.01

4.3.5. Comparative experiment in noisy environment

We processed the test set images of the ISIC and BUSI datasets with noise to simulate the noisy environment, and used the processed test sets to conduct experiments to test the robustness of each model. The test results are shown in Table 7.

From Table 7, we can see that the ACC and AUC of the compared advanced models are affected to varying degrees. The teacher model we designed performed well overall on the dataset. The average ACC of the lightweight student model we proposed in a noisy environment reached 56.85%, and the average AUC reached 83.00%. Not only did the accuracy not affect the reduction of the number of parameters, but the results were better than most advanced comparison methods. This shows that the robustness of the student model we proposed is very strong and can achieve a high accuracy in a noisy environment. This is attributed to the fact that the auxiliary teacher made full use of Low-Quality Data (LQD) with no labels and poor image quality to assist in the training of the model during the training process, effectively helping the model to adapt to the image data in various imaging environments in advance, thereby improving the robustness of the model.

4.3.6. Inference time and GPU usage comparison experiment

At the same time, in order to show the advantage of the student model in terms of parameter quantity, we compared the parameters (M), computational complexity (GFLOPs), and average inference time (ms) on the CPU with the comparison method. The comparison results are shown in Table 8.

From the results in Table 5 and Table 7, we can see that the global teacher model achieves ACC of 93.84% and AUC of 96.03% on the ISIC

Table 7
Comparative experiments under noisy datasets.

	ISIC 2018 AUC %	ACC %	BUSI AUC %	ACC %	AUC-Avg%	ACC-Avg%
ResNet50	82.58	53.15	63.19	32.11	72.89	42.63
VIT-B-16	76.60	41.11	62.93	41.82	69.77	41.47
DenseNet121	87.11	59.14	81.88	48.50	84.50	53.82
EfficientNet-b2	79.11	53.94	78.45	48.79	78.78	51.37
BiFormer-B (Zhu et al., 2023)	77.19	28.02	58.91	33.22	68.05	30.62
Eff2Net (Karthik et al., 2022)	78.26	46.17	79.38	48.55	78.82	47.36
Hifuse (Huo, 2024)	75.41	35.78	64.52	40.72	69.97	38.25
DGLA-ResNet50 (Tan et al., 2024)	80.46	56.33	81.32	49.86	80.89	53.10
SRC-MT (Liu et al., 2020)	83.55	55.46	80.78	50.86	82.17	53.16
Student-Net	84.91	60.19	81.08	53.50	**83.00**	**56.85**

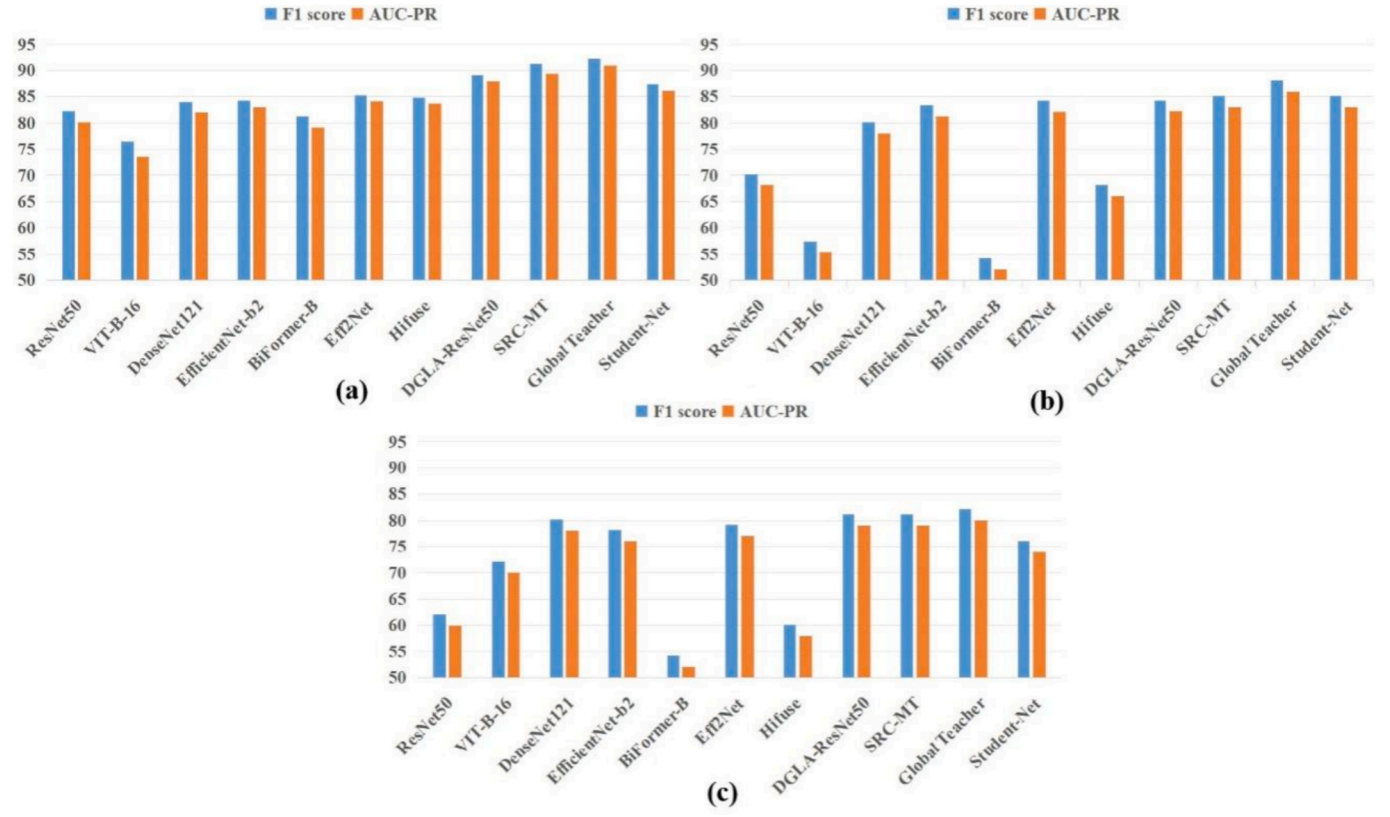


Fig. 6. Visual comparison results of our model with existing methods on F1 scores and AUC-PR.

Table 8
Comparison results of parameter quantity (M), computational complexity (GFLOPs), and average inference time (ms) on the CPU.

	Par	GFLOPs	AIT
ResNet50	23.52	4.13	77.33
VIT-B-16	85.80	16.86	295.63
DenseNet121	6.96	2.89	79.56
EfficientNet-b2	7.71	0.70	57.41
BiFormer-B (Zhu et al., 2023)	56.04	9.80	406.85
Eff2Net (Karthik et al., 2022)	96.40	13.60	236.21
Hifuse (Huo, 2024)	127.80	10.97	361.32
DGLA-ResNet50 (Tan et al., 2024)	104.20	15.60	287.23
SRC-MT (Liu et al., 2020)	6.96	2.89	81.15
Teacher-Net	176.37	13.66	501.65
Student-Net	**0.18**	**0.25**	**18.94**

2018 dataset, ACC of 91.03% and AUC of 94.27% on the BUSI dataset, and ACC of 84.45% and AUC of 97.59% on the Dermnet dataset, which is significantly better than the existing advanced models. This shows that the global teacher model we proposed can better capture the distribution characteristics of these data in the high-order space and accurately classify them. At the same time, the accuracy of our proposed student model is also close to the existing advanced models, but compared with the best existing methods, its parameter volume is only 0.18M,which is 38 times lower, GFLOPs is only 0.25, which is 11 times lower, and AIT is only 18.94ms, which greatly improves the efficiency of the model. In terms of robustness, the model demonstrates a strong capability for handling noisy data, which is crucial in real-world scenarios such as busy clinical environments. Regarding efficiency, the low number of parameters and computational complexity of the model enable fast inference times, making it suitable for real-time applications on devices with limited computational resources. Nonetheless, power consumption remains an issue in continuous usage scenarios. Future research should focus on optimizing the model to reduce power consumption while maintaining accuracy, thereby enhancing its overall applicability to point-of-care medical applications. In order to more intuitively demonstrate the advantages of the student model, we have visually compared the results in Table 6, as shown in Fig. 7.

5. Discussion

Medical image analysis plays a vital role in disease diagnosis and treatment planning. With the development of medical technology, instant imaging equipment is increasingly used in clinical settings, such as Point-Of-Care Ultrasound (POCUS) and smartphone cameras. While these devices provide instant diagnostic information, they also face the challenges of poor imaging quality and limited computing resources. Although traditional deep learning models perform well in image analysis, their high computational complexity and parameter volume make them difficult to deploy on these devices. Therefore, developing lightweight and efficient medical image classification models has become an urgent need. In this study, we propose a lightweight student model that optimizes residual information and significantly improves its accuracy and noise resistance in medical image classification tasks through a multi-teacher distillation strategy. First, the lightweight student model we designed significantly improves the model's ability to capture multiscale spatial features by introducing the Shift MLP structure on the residual branch. This design not only reduces the number of parameters of the model, but also reduces the computational complexity. At the same time, we designed validation experiments to evaluate the impact of this structure on model performance under different configurations. Overall, adding the Shift MLP structure to residual branches at different depths improves the model's classification performance. However, as the model depth increases, incorporating Shift MLP into the deepest residual branch does not significantly enhance the overall performance of the model. Therefore, in order to balance the increase in the number of layers of the Shift MLP structure and the lightweight of the model, we chose to use this structure on the 3-layer residual branch to form our student model. Secondly, we propose a multi-teacher distillation strategy, including auxiliary teachers and global teacher models. The auxiliary teacher model enhances the robustness of the student model by extracting adaptive knowledge from low-quality data through semi-supervised learning. The global teacher model transfers deep feature knowledge to the student model through soft label distillation, improving its classification accuracy. This strategy not only improves the performance of the student model, but also indirectly enhances the teaching ability of the assistant teacher and achieves knowledge transfer at the global level.

The success of our proposed architecture can be attributed to several underlying physical interpretations. From the perspective of feature extraction, the Shift MLP structure in the lightweight student model effectively captures multi-scale spatial features. In medical images, different pathological features may vary in scale. The Shift MLP structure, through its operations such as patch embedding and feature shift along width and height axes, is able to adaptively focus on these

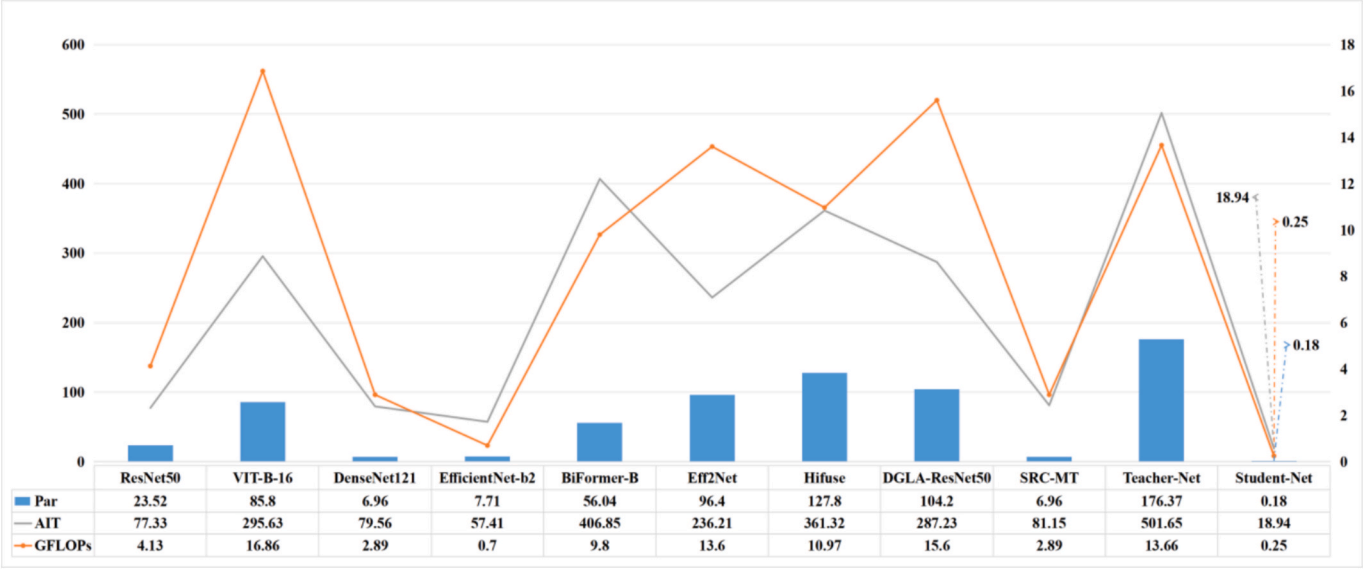


Fig. 7. The results of visual comparisons between student models in terms of number of parameters, computational complexity, and average inference time.

different – scale features. This is similar to how the human visual system processes images; it can quickly detect both global and local details. By extracting multi-scale features, the model can comprehensively analyze the image content, which is crucial for accurate disease diagnosis. Regarding the multi – teacher distillation strategy, the auxiliary teacher model plays a significant role in enhancing the model's robustness. In real – world medical imaging, the data is often contaminated by noise, which can be considered as random fluctuations in the image signal. The auxiliary teacher model uses unlabeled and noisy data for adaptive learning. This is equivalent to training the model to be resilient to these signal fluctuations. It learns to identify the underlying true patterns in the data, much like how a doctor with experience can still make accurate diagnoses even when facing slightly blurry or noisy images. The global teacher model, on the other hand, transfers deep – feature knowledge. In medical images, these deep – level features represent complex semantic information related to diseases. By imitating the probability distribution of the global teacher model's prediction results, the student model can learn to better distinguish different disease categories, similar to a novice doctor learning from an experienced expert to improve diagnostic accuracy.

Although the lightweight student model and multi-teacher distillation strategy proposed in this study have shown significant performance improvements in medical image classification tasks, we recognize that there is still room for further optimization and expansion. When compared with similar schemes, our proposed approach exhibits distinct advantages. In terms of model structure, many existing methods rely on complex convolutional neural network architectures with numerous layers and parameters (such as ResNet, DenseNet), which not only increase the computational complexity but also make them unsuitable for deployment on resource – constrained point – of – care devices. Our lightweight student model, on the other hand, incorporates a Shift MLP – based structure on the residual branch. This design significantly reduces the number of parameters (only 0.18M, 38 times lower than some of the best – performing existing models) and computational complexity (GFLOPs of 0.25, 11 times lower), while maintaining high classification accuracy. Regarding the learning mechanism, traditional methods often struggle to handle noisy data and lack effective utilization of unlabeled data. Our adaptive learning method based on auxiliary teachers, however, can leverage unlabeled and noisy data for training. This enables the model to adapt to various imaging environments in advance, enhancing its robustness. As demonstrated in the comparative experiment in a noisy environment, the average accuracy of our student model reached 56.85% and the average AUC reached 83.00%, outperforming most of the advanced comparison methods. In terms of the knowledge distillation strategy, our multi – teacher distillation strategy, which combines an auxiliary teacher and a global teacher model, achieves more comprehensive knowledge transfer. The global teacher model transfers deep – feature knowledge to the student model, improving its classification accuracy, while the auxiliary teacher model enhances the student model's robustness. This is in contrast to some single – teacher or less – sophisticated distillation strategies in similar studies, which may not fully exploit the potential of knowledge transfer.

Our proposed model shows great potential for clinical deployment, thanks to its lightweight architecture (0.18M parameters, 0.25 GFLOPs) and fast CPU inference time (18.94ms). However, several limitations exist. Data variability in clinical settings, caused by different imaging equipment, patient conditions, and protocols, can still affect the model's performance. For example, images from different ultrasound machines may lead to misclassification. In real – world implementation, while computational requirements are low compared to traditional models, they can still be a problem for low – end devices. Also, obtaining large amounts of high – quality labeled data for training is time – consuming, costly, and may raise privacy issues. Compared to similar schemes, our approach relies heavily on complex teacher models, which require substantial resources and time to train, increasing development costs. Additionally, our model's ability to extract image information at different resolutions is limited compared to methods focusing on dynamic resolution adaptation. To address these limitations, we considered exploring a new design from domain generalization and knowledge distillation that would not only improve the performance of the model, but also make it better able to handle various challenges in clinical data variability and practical applications.

6. Conclusion

The lightweight student model we proposed successfully optimizes the processing of residual information by introducing the Shift MLP structure, thereby significantly enhancing the model's ability to capture multi-scale spatial features. This innovative design not only reduces the number of model parameters and computational complexity but also maintains good performance, representing an important improvement over traditional model structures. Meanwhile, the multi-teacher distillation strategy we designed, which combines auxiliary and global teacher models, achieves efficient knowledge transfer and significantly improves the accuracy and noise resistance of the student model. This strategy demonstrates unique advantages in medical image classification tasks and provides an effective solution to the practical challenges faced by point-of-care imaging devices. Our experimental results show that the student model after multi-teacher distillation occupies less computing resources and has a shorter inference time with almost no loss in accuracy, which is conducive to deployment on point-of-care imaging devices. Our research findings bring new ideas and methods to the field of medical image analysis and are expected to promote further advancements in clinical diagnostic technologies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by National Natural Science Foundation of China, Regional Science Fund Project, No:72264037.

Data availability

Data will be made available on request.

References

Abbasi, S., et al. (Nov. 2021). Classification of diabetic retinopathy using unlabeled data and knowledge distillation. *Artificial Intelligence in Medicine, 121*, Article 102176. https://doi.org/10.1016/j.artmed.2021.102176

W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, no. 2, 2020, Accessed: Jul. 23, 2024. [Online]. Available: https://www.zhangqiaokeyan.com/academic-journal-foreign_detail_thesis/0204121164197.html.

Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0. Accessed: Jul. 26, 2024. [Online]. Available: https://openaccess.thecvf.com/content_ICCVW_2019/html/NeurArch/Cao_GCNet_Non-Local_Networks_Meet_Squeeze-Excitation_Networks_and_Beyond_ICCVW_2019_paper.html.

N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC: IEEE, Apr. 2018, pp. 168–172. doi: 10.1109/ISBI.2018.8363547.

Dutta, S. (Jan. 2019). Point of care sensing and biosensing using ambient light sensor of smartphone: critical review. *TrAC Trends in Analytical Chemistry, 110*, 393–400. https://doi.org/10.1016/j.trac.2018.11.014

J. Fu *et al.*, "Dual Attention Network for Scene Segmentation," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154. Accessed: Jul. 26, 2024. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Fu_Dual_Attention_Network_for_Scene_Segmentation_CVPR_2019_paper.html.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (Jun. 2021). Knowledge distillation: a survey. *International Journal of Computer Vision, 129*(6), 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Grotberg, J. C., McDonald, R. K., & Co, I. N. (Jan. 2024). Point-of-care echocardiography in the difficult-to-image patient in the ICU: a narrative review. *Critical Care Explorations, 6*(1), e1035.

Gu, Y., et al. (Jan. 2025). Dual structure-aware image filterings for semi-supervised medical image segmentation. *Medical Image Analysis, 99*, Article 103364. https://doi.org/10.1016/j.media.2024.103364

G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," Mar. 09, 2015, *arXiv*: arXiv:1503.02531. doi: 10.48550/arXiv.1503.02531.

A. Howard *et al.*, "Searching for MobileNetV3," Nov. 20, 2019, *arXiv*: arXiv:1905.02244. doi: 10.48550/arXiv.1905.02244.

J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018. Accessed: Jul. 26, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/dc363817786ff182b7bc59565d864523-Abstract.html.

J. Hu, L. Shen, G. Sun, "Squeeze-and-Excitation Networks," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141. Accessed: Jul. 26, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612. Accessed: Jul. 26, 2024. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Huang_CCNet_Criss-Cross_Attention_for_Semantic_Segmentation_ICCV_2019_paper.html.

Huang, O., & Palmeri, M. L. (2024). TPU based deep learning image enhancement for real-time point-of-care ultrasound. *IEEE Transactions on Computational Imaging, 10*, 461–468. https://doi.org/10.1109/TCI.2024.3372445

Huo, X., et al. (Jan. 2024). HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control, 87*, Article 105534. https://doi.org/10.1016/j.bspc.2023.105534

Karthik, R., Vaichole, T. S., Kulkarni, S. K., Yadav, O., & Khan, F. (Mar. 2022). Eff2Net: An efficient channel attention-based convolutional neural network for skin disease classification. *Biomedical Signal Processing and Control, 73*, Article 103406. https://doi.org/10.1016/j.bspc.2021.103406

M. Khan, J. Ahmad, A. E. Saddik, and W. Gueaieb, "Skin-Former: Mobile-Friendly Transformer for Skin Lesion Diagnosis," in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2024, pp. 1–6. doi: 10.1109/ICCE59016.2024.10444175.

Kumar, A., Vishwakarma, A., Bajaj, V., & Mishra, S. (2024). Novel mixed domain hand-crafted features for skin disease recognition using multiheaded CNN. *IEEE Transactions on Instrumentation and Measurement, 73*, 1–13. https://doi.org/10.1109/TIM.2024.3370772

S. Laine, T. Aila, "Temporal Ensembling for Semi-Supervised Learning," Mar. 15, 2017, *arXiv*: arXiv:1610.02242. doi: 10.48550/arXiv.1610.02242.

Y. Li, X. Liu, J. Yu, Y. Li. (2023). A full-set tooth segmentation model based on improved PointNet++. *Visual Intelligence* 1, Article no. 21.

Z. Liao, Q. Dong, Y. Ge, W. Liu, H. Chen, and Y. Song, "Knowledge Distillation of Attention and Residual U-Net: Transfer from Deep to Shallow Models for Medical Image Classification," in *Pattern Recognition and Computer Vision - 6th Chinese Conference, PRCV 2023, Xiamen, China, October 13-15, 2023, Proceedings, Part XIII*, Q. Liu, H. Wang, Z. Ma, W. Zheng, H. Zha, X. Chen, L. Wang, and R. Ji, Eds., in Lecture Notes in Computer Science, vol. 14437. Springer, 2023, pp. 162–173. doi: 10.1007/978-981-99-8558-6_14.

Liu, H., Ren, P., Song, C., Yuan, Y., & Luo, F. (2025). Stochastic augmented-based dual-teaching for semi-supervised medical image segmentation. *CMC, 82*(1), 543–560. https://doi.org/10.32604/cmc.2024.056478

Liu, Q., Yu, L., Luo, L., Dou, Q., & Heng, P. A. (Nov. 2020). Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging, 39*(11), 3429–3440. https://doi.org/10.1109/TMI.2020.2995518

Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Aug. 17, 2021, *arXiv*: arXiv:2103.14030. doi: 10.48550/arXiv.2103.14030.

H. Mazumdar, C. Chakraborty, M. Sathvik, sabyasachi Mukhopadhyay, P. K. Panigrahi, "GPTFX: A Novel GPT-3 Based Framework for Mental Health Detection and Explanations," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–8, 2023, doi: 10.1109/JBHI.2023.3328350.

S. Mehta, M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," Mar. 04, 2022, *arXiv*: arXiv:2110.02178. doi: 10.48550/arXiv.2110.02178.

K. A. Mekonnen, N. Dall'Asen, P. Rota, "Adv-KD: Adversarial Knowledge Distillation for Faster Diffusion Sampling," May 31, 2024, *arXiv*: arXiv:2405.20675. doi: 10.48550/arXiv.2405.20675.

Qiu H., Shi C., "LM-UNet: Lateral MLP Augmented U-Net for Medical Image Segmentation," Jun. 23, 2023. doi: 10.21203/rs.3.rs-3082767/v1.

Rasheed, S., Kanwal, T., Ahmad, N., Fatima, B., Najam-ul-Haq, M., & Hussain, D. (Apr. 2024). Advances and challenges in portable optical biosensors for onsite detection and point-of-care diagnostics. *TrAC Trends in Analytical Chemistry, 173*, Article 117640. https://doi.org/10.1016/j.trac.2024.117640

A. Samardzija *et al.*, "Low-Field, Low-Cost, Point-of-Care Magnetic Resonance Imaging," *Annual Review of Biomedical Engineering*, vol. 26, no. Volume 26, 2024, pp. 67–91, Jul. 2024, doi: 10.1146/annurev-bioeng-110122-022903.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

M. B. Sariyildiz, P. Weinzaepfel, T. Lucas, D. Larlus, Y. Kalantidis, "UNIC: Universal Classification Models via Multi-teacher Distillation," Aug. 09, 2024, *arXiv*: arXiv:2408.05088. doi: 10.48550/arXiv.2408.05088.

G. Sivapriya, R. Manjula Devi, P. Keerthika, and V. Praveen, "Automated diagnostic classification of diabetic retinopathy with microvascular structure of fundus images using deep learning method," *Biomedical Signal Processing and Control*, vol. 88, p. 105616, Feb. 2024, doi: 10.1016/j.bspc.2023.105616.

Tan, L., Wu, H., Xia, J., Liang, Y., & Zhu, J. (Jan. 2024). Skin lesion recognition via global-local attention and dual-branch input network. *Engineering Applications of Artificial Intelligence, 127*, Article 107385. https://doi.org/10.1016/j.engappai.2023.107385

Tan, Z., Yu, Y., Meng, J., Liu, S., & Li, W. (Jan. 2024). Self-supervised learning with self-distillation on COVID-19 medical image classification. *Computer Methods and Programs in Biomedicine, 243*, Article 107876. https://doi.org/10.1016/j.cmpb.2023.107876

Y. Tian, D. Krishnan, P. Isola, "Contrastive Representation Distillation," Jan. 24, 2022, *arXiv*: arXiv:1910.10699. doi: 10.48550/arXiv.1910.10699.

S. K. Vashist, "Point-of-Care Diagnostics: Recent Advances and Trends," *Biosensors*, vol. 7, no. 4, Art. no. 4, Dec. 2017, doi: 10.3390/bios7040062.

Veeramani, N., Jayaraman, P., Krishankumar, R., Ravichandran, K. S., & Gandomi, A. H. (Jan. 2024). DDCNN-F: Double decker convolutional neural network 'F' feature fusion as a medical image classification framework. *Scientific Reports, 14*(1), 676. https://doi.org/10.1038/s41598-023-49721-x

S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19. Accessed: Jul. 26, 2024. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.html.

X. Li, W. Wang, X. Hu, J. Yang, "Selective Kernel Networks," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519. Accessed: Jul. 24, 2024. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Li_Selective_Kernel_Networks_CVPR_2019_paper.html.

Ye, M., Ji, C., Chen, H., Lei, L., Lu, H., & Qian, Y. (Sep. 2020). Residual deep PCA-based feature extraction for hyperspectral image classification. *Neural Computing and Applications, 32*(18), 14287–14300. https://doi.org/10.1007/s00521-019-04503-3

Yuan, L., Liu, X., Yu, J., et al. (2023). A full-set tooth segmentation model based on improved PointNet++. *Vis. Intell, 1*, 21. https://doi.org/10.1007/s44267-023-00026-7

L.-L. Zeng *et al.*, "Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI," *eBioMedicine*, vol. 30, pp. 74–85, Apr. 2018, doi: 10.1016/j.ebiom.2018.03.017.

Q. Zeng, Y. Xie, Z. Lu, Y. Xia. (2024). A human-in-the-loop method for pulmonary nodule detection in CT scans. *Visual Intelligence* 2, Article no. 19.

L. Zhu, X. Wang, Z. Ke, W. Zhang, R. W. H. Lau, "BiFormer: Vision Transformer With Bi-Level Routing Attention," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10323–10333. Accessed: Jul. 24, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Zhu_BiFormer_Vision_Transformer_With_Bi-Level_Routing_Attention_CVPR_2023_paper.html.