J Inf Process Syst, Vol.21, No.5, pp.542~554, October 2025
https://doi.org/10.3745/JIPS.02.0229 ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Classification Model of Diabetic Retinopathy Based on a Lightweight Feature-Enhanced Residual Swin Transformer

Na Li* and Kai Ren

Abstract

The automated detection of diabetic retinopathy (DR) relies heavily on retinal image analysis. While artificial intelligence models have shown promise in DR management, they often face challenges such as high computational complexity, reduced accuracy due to class imbalance and small inter-class gaps, and increased processing times. Addressing these limitations, this study introduces the lightweight feature-enhanced residual Swin (LFRS) Transformer, a model that maintains high accuracy despite significantly lowered computational demands. Our approach begins by converting color fundus images to grayscale, followed by local feature extraction using a depthwise separable convolution module. These features are subsequently subjected to processing by a lightweight Swin Transformer enhanced with residual connections, improving both global feature extraction and computational efficiency. Evaluated on the DR classification dataset released by APTOS 2019, the LFRS Transformer achieves an accuracy of 0.928, a recall of 0.965, and a weighted kappa score of 0.957. Compared to the baseline Swin Transformer, our model reduces computational load by 22.2 GFLOPs and decreases model parameters by 7.2M, demonstrating a substantial improvement in efficiency. These results underscore the LFRS Transformer as a highly efficient and reliable DR screening tool, positioning it as well-suited for large-scale clinical screening programs.

Keywords

Classification of Diabetic Retinopathy, Depthwise Separable Convolution, Residual Connection, Swin Transformer

1. Introduction

Diabetic retinopathy (DR) is a common and serious complication associated with diabetes mellitus, characterized by a series of chronic progressive fundus diseases that lead to retinal microvascular leakage and occlusion. These conditions include microaneurysms, hard exudates, hemorrhages, neovascularization, vitreous proliferation, macular edema, and retinal detachment, collectively making DR a leading cause of irreversible blindness [1]. Early detection and timely intervention can prevent over 90% of blindness cases caused by DR. DR is broadly classified into two principal types: non-proliferative (NPDR) and proliferative (PDR). NPDR is further stratified into three stages [2], resulting in a total of five severity levels for DR classification [3]. Accurate differentiation among these stages is critical, as each stage necessitates distinct treatment strategies. Fundus imaging serves as the primary diagnostic tool

Manuscript received October 28, 2024; first revision March 21, 2025; accepted May 26, 2025.

***Corresponding Author:** Na Li (lina@scuec.edu.cn)
School of Computer Science (School of Artificial Intelligence), South-Central Minzu University, Wuhan Hubei, China (lina@scuec.edu.cn, 81681767@qq.com)

for distinguishing these stages. However, defining DR lesions is a complex and labor-intensive task due to the subtle morphological differences between lesions at each stage.

Recently, deep learning techniques gained significant attention for their potential in automating DR classification [4]. Among deep learning methods, convolutional neural networks (CNNs) [5-7] dominate current implementations. For instance, Jiang et al. [8] utilized a pretrained ResNet as the backbone architecture and enhanced it with a cross-attention mechanism. The effectiveness of this improved network was validated using both the internal DFiD dataset and the publicly available DeepDRiD dataset. Similarly, Liu et al. [9] employed a pretrained EfficientNet-B3 as the backbone network, fine-tuning it through oversampling and class weight adjustments. Their method demonstrated superior performance on the Kaggle OCT2017 dataset. Zheng et al. [10] introduced a modified ResNet50 network, replacing traditional convolution with dilated convolution to expand the receptive field while maintaining computational efficiency. The incorporation of an attention mechanism further enhanced the network's ability to extract lesion features, yielding superior classification performance relative to the baseline model. Despite their success, most CNN-based models tend to overlook global image information. This limitation has prompted the integration of attention mechanisms to improve global feature extraction.

The Vision Transformer (ViT) model offers significant advantages over traditional CNNs. By leveraging a cross-domain attention mechanism, the ViT excels at capturing global image information, overcoming the inherent limitation of CNNs, which are often constrained by local receptive fields. This capability enables the ViT to address a critical shortcoming of CNNs in tasks requiring comprehensive feature extraction. In [11, 12], the authors have explored the application of existing Transformer models, such as the ViT, for DR classification, reporting promising results. These findings collectively highlight the potential of Transformer-based architectures in advancing the accuracy and robustness of automated DR diagnosis. Further advancements have been made by Oulhadj et al. [13], who proposed a hybrid approach combining the ViT model with an enhanced capsule network to achieve comprehensive feature extraction. Following this approach, the CoT-XNet framework [14] integrates transformer-based context modeling with Xception blocks through parallel pathways, effectively boosting DR diagnosis performance. The Swin Transformer has emerged as a prominent Transformer-based framework in computer vision, demonstrating exceptional performance in DR grading tasks. Wang et al. [15] adapted a pretrained Swin Transformer Tiny model for DR classification and compared its performance against three classic CNN models: EfficientNetV2, ResNet-50, and GoogLeNet. The Swin Transformer achieved Top-1 accuracy improvements of 2.3%, 5.4%, and 7.1% over these models, highlighting its superior capability for DR classification tasks. While Transformer-based models have shown promise, their substantial parameter size remains a significant limitation, posing challenges for real-time medical applications. To address this, the development of lightweight modifications to these models is essential to ensure computational efficiency and practical applicability in clinical settings.

We propose a lightweight feature-enhanced residual Swin Transformer (LFRS Transformer), a novel lightweight Transformer model specifically designed to address the dual challenges of computational efficiency and classification accuracy in DR diagnosis. The lightweight model's performance is boosted through the novel local feature (LF) extraction module that exploits CNN-native inductive biases. By integrating this module into the model architecture, we effectively combine the complementary advantages of Swin Transformer and CNN, enabling the model to capture global contextual information through the Transformer and fine-grained local features through the CNN-based module. In our approach, fundus images are first processed through the LF extraction module, which generates detailed feature maps. The resulting feature representations are then processed by a lightweight Swin Transformer for final classification. To further enhance the model's performance and stability, we incorporate residual

connections into the Swin Transformer blocks (STBs), facilitating smoother gradient flow and improving training efficiency. Extensive experimental evaluations demonstrate that the proposed LFRS Transformer attains an optimal trade-off in classification performance versus computational demands relative to current best-performing approaches. The key contributions of this work are as follows:

- Lightweight architecture: We propose a lightweight Swin Transformer model optimized for real-time medical image classification. This architecture substantially lowers computational demands without compromising accuracy, enabling practical implementation in medical settings with limited resources.
- Local feature extraction module: The LF extraction module is primarily composed of multi-layer depthwise separable convolutions, which introduce the inductive bias of convolutions to the model. This design alleviates the Transformer architecture's reliance on extensive training data and effectively integrates the advantages of CNN and Transformer models. The proposed module augments the model's LF representation capability, a crucial factor for reliable DR diagnosis.
- Residual connections in STBs: We incorporate residual connections into the STBs to improve gradient flow and stabilize training. This modification enhances the model's ability to learn complex feature representations while maintaining computational efficiency.
- Comprehensive evaluation: Our rigorous experimental evaluation shows the model's efficacy. The results demonstrate that the LFRS Transformer achieves an optimal trade-off between accuracy and processing efficiency relative to current approaches, making it a robust solution for real-time DR diagnosis.

This work provides a computationally efficient and accurate solution for diabetic retinopathy classification, with potential applications in real-time medical diagnostics and other resource-constrained environments.

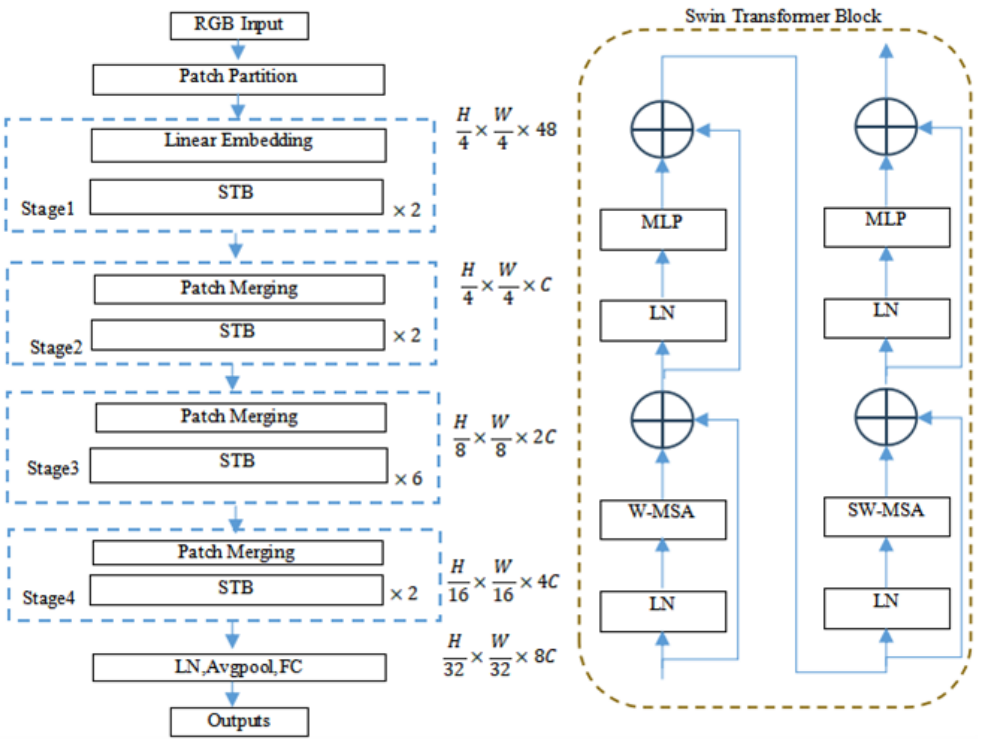


Fig. 1. Swin Transformer Tiny architecture.

2. Swin Transformer Model

The Swin Transformer model, designed by Microsoft Research at ICCV 2021 [16], builds on the ViT [17] and employs a hierarchical CNN-like framework for multi-scale image detection. It utilizes Windows multi-head self-attention (W-MSA) and shifted Windows multi-head self-attention (SW-MSA) modules to compute self-attention within and between windows, respectively. The model comes in four configurations—Tiny, Small, Base, and Large—scaled by parameter count. Fig. 1 depicts the model architecture of Swin Transformer Tiny. Its architecture consists of four stages (Stage1–Stage4), each comprising stacked STBs with paired W-MSA and SW-MSA modules. While Swin Transformer has demonstrated strong capabilities in image classification and related tasks like segmentation and object detection, its high computational complexity and large parameter size pose challenges for deployment in medical image processing. To ensure operational efficiency and compatibility with medical equipment, the model requires lightweight adaptations.

3. Local Feature Extraction Module

Transformer-based models typically require training on large-scale datasets to achieve performance superior to models based on CNNs. This is primarily attributed to the lack of inductive bias in the Transformer architecture, which is inherently embedded in CNNs. CNNs utilize sliding-window convolutions to aggregate local spatial features, whereas Transformer models rely on the self-attention mechanism to establish non-local connectivity patterns in the input sequence, thereby capturing long-range dependencies. As a result, Transformer models require significantly larger amounts of training data to achieve performance comparable to CNNs. Transformers incur heavier computational burdens, a direct consequence of their position-agnostic attention and lack of built-in spatial hierarchies. Bridging these efficiency gaps, we architect a lightweight Swin Transformer variant. However, the lightweight nature of the model may lead to a reduction in accuracy. To compensate for this, we introduce a LF extraction module (hereafter, LF module). This module is designed as a multi-layer architecture based on depthwise separable convolutions, which are computationally more efficient than traditional convolutions due to their reduced parameter count and lower computational requirements [18,19]. The LF module is positioned before the lightweight Swin Transformer to incorporate the inductive bias of convolution into the overall model, thereby enhancing its ability to capture local features while maintaining computational efficiency.

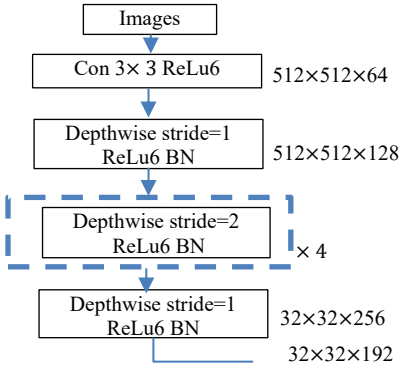


Fig. 2. Local feature extraction module.

Fig. 2 presents the system-level schematic of our proposed solution. The LF module consists of six layers. Feature extraction begins with a 3×3 convolutional operation (stride 1, 64 channels). A depthwise separable convolution layer follows, maintaining the 3×3 receptive field and stride while expanding to 128 output channels. The next four layers are depthwise separable convolution layers, each employing a 3×3 convolution kernel with a stride of 2. This design systematically condenses the feature map's spatial extent while expanding channel depth, thereby balancing computational efficiency with feature representation capability. Through the LF module, a feature map of size 32×32×192 is generated, which is then passed to the subsequent stages of the network.

By integrating the LF extraction capabilities of CNNs with the global context modeling of Transformers, this architecture achieves a balance between computational efficiency and the ability to capture both local and global dependencies in the data. The proposed LF module not only compensates for the accuracy loss caused by the lightweight Swin Transformer but also enhances the model's overall performance by leveraging the strengths of both convolutional and self-attention mechanisms.

4. Proposed Method

In this paper, the LFRS Transformer model is proposed for DR classification. The model is composed of a LF module and a lightweight Swin Transformer. Our lightweight Swin Transformer comprises five key components: an image partition module, residual Swin Transformer blocks (RSTBs), a normalization module, average pooling, and fully connected layers. Initially, the preprocessed fundus images' local features are extracted by the LF module. Following this, the obtained LF map is divided into 64 non-overlapping patches, each measuring 4×4 in size. In terms of the design of the Swin Transformer Tiny, the model capacity and calculations are concentrated in four stages of STBs, thereby making them lightweight. Within the Swin Transformer, STBs are improved by skip connection. The architecture of the RSTB is shown on the right-hand side of Fig. 3. The main architecture of the STB remains unchanged, but residual connections have been incorporated, as visually annotated in Fig. 3. The parameters and computational cost of the RSTB have not increased significantly. The architectural specifications of our lightweight Swin Transformer are summarized in Table 1.

Table 1. The lightweight Swin Transformer model parameters

Patch size	Output size	Window size	*dim*	*head*	*h*
4×4	32×32	7×7	192	4	2
8×8	16×16	7×7	384	8	2
8×8	16×16	7×7	384	16	2
8×8	16×16	7×7	384	32	2

The patch size in the table is the dimensions of the image block after the patch merging module. The window size in the W-MSA and SW-MSA modules is 7×7. *dim* is the length of the patch sequence; *head* denotes the count of attention heads, while *h* quantifies RSTB instances. The model utilizes a pyramid-style attention head count configuration (4, 8, 16, 32), corresponding to STBs with layer depths of (2, 2, 2, 2). In the first stage, four attention heads are used to focus on the extraction of local details. In the second stage, the number of attention heads is extended to 8 to capture medium-scale feature correlations. In the third stage, 16 attention heads are added to enhance feature extraction through a deeper network layer. In the fourth stage, 32 attention heads are used to effectively extract global semantic features in the

deepest layer. This progressive expansion of attention heads allows the network to hierarchically construct feature representations, evolving from fine-grained local patterns to comprehensive global understanding.

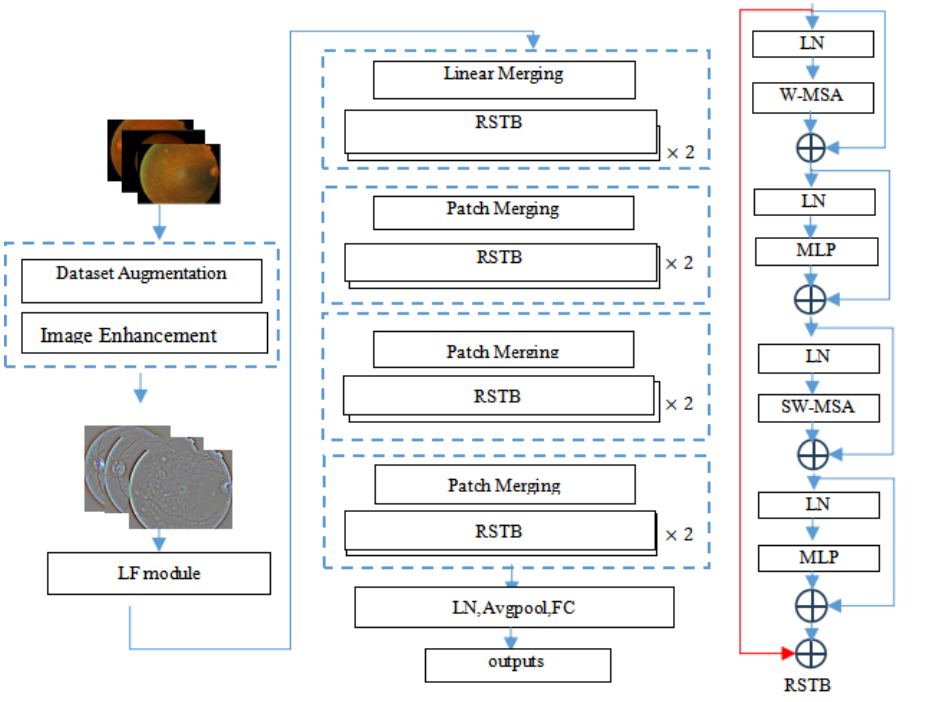


Fig. 3. LFRS Transformer architectural design.

5. Experimental Results

5.1 Dataset

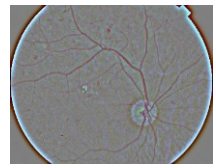
The APTOS 2019 dataset was selected to evaluate our approach, as it has been the focus of numerous studies and offers a comprehensive assessment. This dataset contains 3,662 training samples and 1,982 test instances. Since the test set labels are undisclosed, we adopted established protocols by dedicating 90% of the training data for model optimization, reserving the remaining 10% for performance evaluation. The following are the labels assigned to the dataset: 0 for normal, 1 for mild, 2 for moderate, 3 for severe, and 4 for proliferative. A representative sample of the dataset is shown in Fig. 4, illustrating examples from all five categories.

(a) (b) (c) (d) (e)

Fig. 4. Samples of diabetic retinal images: (a) No_DR, (b) Mild_DR, (c) Moderate_DR, (d) Severe_DR, and (e) Proliferative_DR.

Data preprocessing is crucial step in our algorithm's workflow. As shown in Fig. 5(a), the blood vessels, exudates, and cotton wool spots, which serve as key classification features, are not obvious. Additionally, we observed variations in size, contrast, and brightness among the images in the dataset. To address these issues, we conducted image preprocessing to achieve higher image quality. To this end, we implemented a series of preprocessing steps. Initially, we converted the original color images to grayscale images. To minimize unnecessary noise within the image, we subsequently utilized a Gaussian filter. This enhanced the robustness of subsequent analysis by providing clearer and more precise extraction of key features in fundus images, for instance, exudates and blood vessels. The comparison before and after image enhancement is shown in Fig. 5. Furthermore, the dataset suffered from an issue of data imbalance. To solve this problem, we adopted a series of data augmentation strategies, for example rotation, scaling, and flipping, thereby achieve a more balanced distribution of all five categories. Table 2 quantifies the sample distribution before and after augmentation.

(a)



(b)

Fig. 5. Retinal images before and after preprocessing: (a) original image and (b) processed image.

Table 2. Distribution of the training set before and after data image augmentation

Label	Amount of original data	Amount of data after image augmentation	Proportion of each category after data augmentation (%)
0	1,625	1,625	19.78
1	270	1,620	19.72
2	980	1,660	20.21
3	164	1,640	19.96
4	258	1,670	20.33

5.2 Experimental Environment

In this study, we utilized the PyTorch deep learning framework. It is an open-source deep learning software library, particularly suited for executing computer vision-related tasks. Our LFRS Transformer was trained on a single NVIDIA GeForce RTX4060 16 GB of VRAM. During model training, all training images were resized to 512×512. Additionally, a fixed batch size of 32 images per transmission was implemented. Model training adopted the cross-entropy criterion. The activation function in the model is the leaky ReLU, and the parameter α value is 0.1. The activation function was applied after each batch normalization layer. In order to alleviate the issue of overfitting, L2 weight attenuation decay (with a value of 0.001) was applied to all trainable variables. The model was trained with 8 threads concurrently.

5.3 Performance Metrics

For the systematic assessment of the proposed method, we employed four key metrics: accuracy, recall, precision, and the weighted kappa score (K). As the core evaluation criterion, accuracy measures the holistic correctness of predictions generated by trained models.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

TP signifies correctly predicted positive cases from the sample population. TN signifies the accurately predicted members of the negative class. FP enumerates the instances where negative-class samples are erroneously classified as positive. On the contrary, FN enumerates the instances where positive-class samples are erroneously classified as negative. When training models on imbalanced datasets, accuracy tends to give priority to the majority of categories with larger proportions in the dataset, giving more weight to their correct prediction, and may ignore a few categories [20]. In this situation, it is recommended to utilize the weighted kappa score for the purpose of evaluation metric. In such scenarios, the weighted kappa coefficient is frequently employed as an evaluation criterion. Weighted kappa is an index to measure the consistency of multi class classification problems. It is achieved by comparing the expected score and the predicted score. A higher score indicates better consistency and superior model performance [21].

$$K = 1 - \frac{\sum_{ij} W_{ij} O_{ij}}{\sum_{ij} W_{ij} E_{ij}} \tag{4}$$

$$E_{ij} = \left(\sum_{i} O_{ij}\right) \times \left(\sum_{j} O_{ij}\right) / \left(\sum O_{ij}\right) \tag{5}$$

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{6}$$

where N is the cardinality of the classification space, i and j are integers with values of $[0, N-1]$. O_{ij} enumerates the prediction instances where category i is classified as category j. W_{ij} is the weight coefficient, which is 0 when the prediction is correct. When the prediction is incorrect ($i \neq j$), the weight coefficient W_{ij} increases.

5.4 Ablation Experiment

To evaluate the effects of the LF module and residual connections on the model's performance, comparative experiments were conducted. Five groups of experiments were designed, and each experiment was carried out using the identical training and test datasets. The outcomes for each model are presented in Table 3. Swin-T is a pretrained Swin Transformer Tiny model. In all experiments, the lightweight Swin-T performed worst in accuracy. It indicates that the model's lightweight design significantly compromises its performance. Comparing the performance of Swin-T and the LFRS Transformer, the GPU floating point operations per second (GFLOPs) and the number of parameters of the LFRS Transformer decrease significantly, indicating that the speed of the LFRS Transformer is significantly higher. Although the Top-1 accuracy decreased by 0.4% and the weighted kappa score decreased by 1.1% compared with the baseline model, it was in line with expectations and the model was made more lightweight at the expense of some accuracy. By replacing depthwise separable convolution with conventional convolution in the same layers within the LF module and removing the residual

connection in the STB, the Top-1 accuracy for the conventional convolution model increased by 0.1%, although the number of GFLOPs also significantly increased. The accuracy of the LFRS Transformer with residual connections in the STBs was higher than when using depthwise separable convolution with the lightweight Swin-T, indicating that the residual connection improved the accuracy of the model. Considering speed and accuracy, our LFRS Transformer model demonstrated an improvement over other models, as shown in Table 3. Therefore, it more effectively fulfills the needs for immediate detection in medical applications.

Table 3. Results of ablation experiments

Model	Top-1 accuracy (%)	Weighted kappa (%)	GFLOPs	Parameters
Swin-T	93.2	97.6	24.76G	27.51M
Lightweight Swin-T	87.8	90.1	6.47G	20.59M
Depthwise separable convolution + lightweight Swin-T	89.8	91.2	6.55G	20.77M
Conventional convolution + lightweight Swin-T	89.9	90.4	334.98G	21.67M
LFRS Transformer	92.8	96.5	6.55G	20.77M

5.5 Comparative Experiment

Table 4 quantitatively establishes our method's advantage across three key metrics: accuracy, recall, and weighted kappa compared to most previous studies. To comprehensively evaluate the effectiveness of our model, we conducted a comparative analysis with three categories of models.

Table 4. Results of comparison experiments

Model	Top-1 accuracy (%)	Recall	Precision	Weighted kappa
CHBP [4]	89.4	92.6	88.6	90.5
Resnet-151 [22]	85.4	84.2	84.5	81.7
DenseNet169 [23]	85.3	87.5	83.4	83.4
EffecientNet-v2 [24]	87.6	84.2	85.8	85.8
ConvNext-T [25]	84.6	85.4	0.854	87.7
Refined ResNet18 [26]	88.3	88.1	0.846	88.3
HDeep [27]	93.5	92.4	0.904	92.6
ML-FEC [28]	92.1	76.3	0.895	92.8
CRA-Net [29]	90.1	86.6	0.866	93.2
STMF-DRNet [31]	87.9	81.6	0.806	89.0
SVPL [30]	93.5	93.1	0.931	91.7
Proposed model	**92.8**	**96.5**	**0.935**	**95.7**

The best metrics are highlighted in bold.

The first category of comparative methods comprises six classical CNN-based image classification models: CHBP [4], ResNet [22], DenseNet169 [23], EfficientNet2 [24], ConvNeXt [25], and refined ResNet18 [26], all of which are based on single-network architectures. These models exhibited lower accuracy compared to our proposed model. The significant performance gap between these models and the pretrained Swin Transformer underscores the latter's superiority in DR classification tasks.

The second category comprises integrated network models: HDeep [27] and ML-FEC [28]. The HDeep network integrates multiple models, including a fine-tuned DenseNet-121, two EfficientNet-B7 models,

and a ResNet50 model, each trained on different datasets. Specifically, the DenseNet-121 model was designed as a binary classifier to identify mild DR cases. The training dataset for this model utilized binary labels (0 and 1), where Label 0 represented mild DR images from the APTOS dataset, and Label 1 included images from the other four severity categories. The remaining three models also functioned as binary classifiers, each predicting a specific DR category. Although this approach achieved strong performance, with a kappa score of 92.4% and a Top-1 accuracy of 93.5% on the test dataset, it suffered from significantly slower detection speeds due to the need to process each image through the four separate models. In contrast, the ML-FEC [28] first condenses features using PCA and subsequently applies an ensemble of three classification models: a ResNet50, a ResNet152, and a SqueezeNet. This approach achieved a Top-1 accuracy exceeding 90%. While both [27] and [28] improved accuracy by integrating multiple neural networks, their increased complexity and computational demands render them less efficient compared to our proposed model.

The third category includes three Transformer-based models: CRA-Net [29], SVPL [30], and STMF-DRNet [31]. CRA-Net architecturally integrates CNNs' LF extraction with Transformers' global relational modeling through a cascaded processing paradigm. This method employs the ViT as the backbone network but does not address model lightweighting. As a result, the model incurs substantial computational overhead and exhibits lower accuracy compared to our proposed model. The semantic-oriented visual prompt learning (SVPL) also utilizes a ViT as the backbone network, appending prompt groups corresponding to different semantic categories to patch embeddings in the input of each Transformer encoder layer. This enhancement significantly improves the model's accuracy, achieving 93.5%. The model attains a parameter complexity of 99.7 million learnable weights, which is significantly larger than that of our proposed method. Rooted in Swin Transformer V2, STMF-DRNet implements a cascaded multi-path framework to amplify discriminative feature learning and boost classification performance. Although this model achieves robust performance, the absence of data preprocessing operations results in lower accuracy compared to our model. Notably, none of the three aforementioned models address the issue of model lightweighting.

In summary, our model achieves a better balance between accuracy and speed compared to the mainstream classification models used for DR classification tasks, making it more suitable for practical medical applications.

6. Conclusion

The approach to improving classification models generally falls into two categories: increasing accuracy and making models more lightweight. In recent years, most DR classification algorithms have focused on achieving high accuracy, yielding promising results. However, relatively little research has been conducted on making models more lightweight. The LFRS Transformer introduced in this study is a lightweight Transformer model. As part of its architecture, a local feature extraction module operates to strengthen the lightweight model's classification fidelity. This module incorporates convolutional inductive priors into the model's learning framework, reducing the significant demand for training data typically required by Transformer models. While the Swin Transformer model focuses on global feature extraction and CNN models focus on local feature extraction, our model combines the advantages of both approaches. To further enhance the model's accuracy, a residual connection is employed in the STB. Experiments demonstrate that the residual connection improves the model's precision by 3%. The design of the model fully considers the real-time requirements of medical tasks. The LFRS Transformer model,

with only 20.7M parameters, requires 6.55 GFLOPs of computation, whereas the Swin Transformer model, with 27.5M parameters, demands 24.76 GFLOPs. Although there is a slight decrease in accuracy, the increase in speed is substantial. Compared with current single backbone networks based on CNNs, our model performs well in DR classification, and the Top-1 accuracy is generally higher than that of CNN-based models. This further highlights the advantages of utilizing the Swin Transformer model in DR classification applications.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work is supported by the Fundamental Research Funds for the Central Universities of South-Central Minzu University (Grant No. CZZ25005).

References

[1] Y. Yin and B. Zhao, "Techniques and application of early diagnosis of diabetic retinopathy," *International Eye Science*, vol. 22, no. 3, pp. 438-442, 2022. https://doi.org/10.3980/j.issn.1672-5123.2022.3.18

[2] Z. Yang, T. E. Tan, Y. Shao, T. Y. Wong, and X. Li, "Classification of diabetic retinopathy: past, present and future," *Frontiers in Endocrinology*, vol. 13, article no. 1079217, 2022. https://doi.org/10.3389/fendo.2022.1079217

[3] J. Li, M. Chen, R. Yang, W. Ma, X. Lai, D. Huang, D. Liu, X. Ma, and Y. Shen, "Fundus Image Screening for Diabetic Retinopathy," *Chinese Journal of Lasers*, vol. 49, no. 11, article no. 1107001, 2022. http://dx.doi.org/10.3788/CJL202249.1107001

[4] K. Ashwini and R. Dash, "Grading diabetic retinopathy using multiresolution based CNN," *Biomedical Signal Processing and Control*, vol. 86, article no. 105210, 2023. https://doi.org/10.1016/j.bspc.2023.105210

[5] A. Ali, S. Qadri, W. Khan Mashwani, W. Kumam, P. Kumam, S. Naeem, et al., "Machine learning based automated segmentation and hybrid feature analysis for diabetic retinopathy classification using fundus image," *Entropy*, vol. 22, no. 5, article no. 567, 2020. https://doi.org/10.3390/e22050567

[6] W. Almattar, H. Luqman, and F. A. Khan, "Diabetic retinopathy grading review: current techniques and future directions," *Image and Vision Computing*, vol. 139, article no. 104821, 2023. https://doi.org/10.1016/j.imavis.2023.104821

[7] S. Karthika and M. Durgadevi, "Improved ResNet_101 assisted attentional global transformer network for automated detection and classification of diabetic retinopathy disease," *Biomedical Signal Processing and Control*, vol. 88, article no. 105674, 2024. https://doi.org/10.1016/j.bspc.2023.105674

[8] L. Jiang, S. Sun, H. Zou, L. Lu, and R. Feng, "Diabetic retinopathy grading based on dual-view image feature fusion," *Journal of East China Normal University (Natural Science)*, vol. 2023, no. 6, pp. 39-48, 2023. https://doi.org/10.3969/j.issn.1000-5641.2023.06.004

[9] Y. Liu, J. Chang, H. Zhang, and T. Hu, "Research on retinal OCT image classification method based on EfficientNet," *Journal of Xi'an University of Arts and Sciences (Natural Science Edition)*, vol. 26, no. 2, pp.

63-67, 2023. https://doi.org/10.3969/j.issn.1008-5564.2023.02.013

[10] W. Zheng, Q. Shen, and J. Ren, "Recognition and classification of diabetic retinopathy based on improved DR-Net algorithm," *Acta Optica Sinica*, vol. 41, no. 22, article no. 2210002, 2021. https://doi.org/10.3788/AOS202141.2210002

[11] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, "Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images," *IEEE Access*, vol. 11, pp. 117546-117561, 2023. https://doi.org/10.1109/ACCESS.2023.3326528

[12] Z. Zhou, H. Yu, J. Zhao, X. Wang, Q. Wu, and C. Dai, "Automatic diagnosis of diabetic retinopathy using vision transformer based on wide-field optical coherence tomography angiography," *Journal of Innovative Optical Health Sciences*, vol. 17, no. 2, article no. 2350019, 2024. https://doi.org/10.1142/S1793545823500190

[13] M. Oulhadj, J. Riffi, C. Khodriss, A. M. Mahraz, A. Yahyaouy, M. Abdellaoui, I. B. Andaloussi, and H. Tairi, "Diabetic retinopathy prediction based on vision transformer and modified capsule network," *Computers in Biology and Medicine*, vol. 175, article no. 108523, 2024. https://doi.org/10.1016/j.compbiomed.2024.108523

[14] S. Zhao, Y. Wu, M. Tong, Y. Yao, W. Qian, and S. Qi, "CoT-XNet: contextual transformer with Xception network for diabetic retinopathy grading," *Physics in Medicine & Biology*, vol. 67, no. 24, article no. 245003, 2022. https://doi.org/10.1088/1361-6560/ac9fa0

[15] G. Wang, H. Wang, S. Wang, Y. Zhu, and M. Liu, "Construction of a pre-trained Swin Transformer model and analysis of its diagnostic efficacy for diabetic retinopathy," *Chinese Journal of New Clinical Medicine*, vol. 16, no. 4, pp. 360-365, 2023.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, Z., ... & Guo, B. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, 2021, pp. 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021 [Online]. Available: https://arxiv.org/abs/2010.11929

[18] X. Bi, S. Chen, and L. Zhang, "Blueprint separable convolutional transformer network for lightweight image super-resolution," *Journal of Image and Graphics*, vol. 29, no. 4, pp. 875-889, 2024. https://doi.org/10.11834/jig.230225

[19] X. Xing, X. Li, C. Wei, Z. Zhang, O. Liu, S. Xie, et al., "DP-GAN+ B: a lightweight generative adversarial network based on depthwise separable convolutions for generating CT volumes," *Computers in Biology and Medicine*, vol. 174, article no. 108393, 2024. https://doi.org/10.1016/j.compbiomed.2024.108393

[20] J. Cohen, "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213-220, 1968. https://doi.org/10.1037/h0026256

[21] Y. Kim, Y. Lee, and M. Jeon, "Imbalanced image classification with complement cross entropy," *Pattern Recognition Letters*, vol. 151, pp. 33-40, 2021. https://doi.org/10.1016/j.patrec.2021.07.017

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2261-2269. https://doi.org/10.1109/CVPR.2017.243

[24] M. Tan and Q. Le, "EfficientNetv2: smaller models and faster training," *Proceedings of Machine Learning Research*, vol. 139, pp. 10096-10106, 2021.

[25] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022, pp. 11966-11976. https://doi.org/10.1109/CVPR52688.2022.01167

[26] S. Sunkari, A. Sangam, R. Raman, and R. Rajalakshmi, "A refined ResNet18 architecture with Swish activation function for diabetic retinopathy classification," *Biomedical Signal Processing and Control*, vol. 88, article no. 105630, 2024. https://doi.org/10.1016/j.bspc.2023.105630

[27] T. F. de Sousa and C. G. Camilo, "HDeep: hierarchical deep learning combination for detection of diabetic retinopathy," *Procedia Computer Science*, vol. 222, pp. 425-434, 2023. https://doi.org/10.1016/j.procs.2023.08.181

[28] T. M. Usman, Y. K. Saheed, D. Ignace, and A. Nsang, "Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 78-88, 2023. https://doi.org/10.1016/j.ijcce.2023.02.002

[29] F. Zang and H. Ma, "CRA-Net: transformer guided category-relation attention network for diabetic retinopathy grading," *Computers in Biology and Medicine*, vol. 170, article no. 107993, 2024. https://doi.org/10.1016/j.compbiomed.2024.107993

[30] Y. Zhang, X. Ma, K. Huang, M. Li, and P. A. Heng, "Semantic-oriented visual prompt learning for diabetic retinopathy grading on fundus images," *IEEE Transactions on Medical Imaging*, vol. 43, no. 8, pp. 2960-2969, 2024. https://doi.org/10.1109/TMI.2024.3383827

[31] Y. Liu, D. Yao, Y. Ma, H. Wang, J. Wang, X. Bai, G. Zeng, and Y. Liu, "STMF-DRNet: a multi-branch fine-grained classification model for diabetic retinopathy using Swin-TransformerV2," *Biomedical Signal Processing and Control*, vol. 103, article no. 107352, 2025. https://doi.org/10.1016/j.bspc.2024.107352

Na Li **https://orcid.org/0000-0001-9531-9003**

She is a lecturer in South-Central Minzu University. Na received master's degree from South-Central Minzu University and Ph.D. degree from Wuhan University. Her main research area is machine learning, medical image processing.

Kai Ren **https://orcid.org/0009-0000-4329-3972**

He holds a multidisciplinary academic trajectory, having earned his Master's degree from South-Central Minzu University, followed by a Ph.D. at Wuhan University. His main research area is Machine learning, natural language processing.