

# A Fundus Image Myopia Diagnosis Model Based on Homogeneous Multimodal Feature Fusion

Peng-Ceng Wen<sup>1</sup>, Yu-Guan<sup>2</sup>, Jian-Qiang Li<sup>3</sup>, and Yin-Zheng Zhao<sup>4</sup>

<sup>1</sup> Faculty of Information Technology, Beijing University of Technology, Beijing, P.R. China  
wpc18581311918@outlook.com

<sup>2</sup> Faculty of Information Technology, Beijing University of Technology, Beijing, P.R. China  
Guanyu0010@126.com

<sup>3</sup> Faculty of Information Technology, Beijing University of Technology, Beijing, P.R. China  
lijianqiang\_bjut@126.com

<sup>4</sup> Beijing Children's Hospital, Capital Medical University, Beijing, P.R. China  
zhaoyinzheng7@163.com

**Abstract.** High myopia is one of the main causes of fundus diseases. If they can be found and treated in time, the risk of fundus lesions in the process of children's growth will be reduced, and the growth rate of patients with a visual disability will be effectively controlled. Therefore, the automatic detection of high myopia based on fundus image is of great significance in clinical practice. In addition, vascular, as one of the significant features in fundus images, are often used as auxiliary elements to help ophthalmologists diagnose. Therefore, in this paper, based on the idea of homogeneous multimodality, we design a neural network model with two branches of simultaneously processing the vascular feature image and original fundus image. Extensive comparative experiments were conducted between our method and other general classification models through a private retinal fundus data set. The results show that our method achieves the best performance of 93.4% in accuracy.

**Keywords:** Myopia, Vascular, Homogeneous multimodal, Classification.

## 1 Introduction

According to the urbanization and human development index in the "world vision report" [26] issued by the World Health Organization on October 10, 2019, the myopia population will increase from 1.95 billion in 2010 to 3.36 billion in 2030, and the number of high myopia related to specific complications [28] will double, including 3.12 billion people under the age of 19 and the number is still rising. Therefore, we should pay attention to the problem of myopia in teenagers and children. As one of the relatively intuitive media in medical images, the fundus image is not only a traditional means used by doctors to assist in diagnosis [18] but also one of the important basis for ophthalmologists to analysis eye diseases. It is a 2D image of the fundus captured by a monocular camera. A complete fundus image mainly includes the optic disc, optic cup, retina, central retinal artery, vein, and macula.

Children with early-onset myopia tend to develop relevant complications. However, some serious complications have no obvious symptoms in the early stage [17]. In addition, it is difficult for parents to find and pay attention to their vision problems on account of their young age and lack of accurate language expression ability. The traditional visual acuity detection and diagnosis process is complicated and involves privacy issues, resulting in a shortage of clinical information. Therefore, regular fundus detection of children is very important to both of the children's bright and medical field. For regions with different levels of development, medical resource imbalance is also a problem. Therefore, in recent years, many achievements in combining big data driven deep learning with medical imaging as auxiliary diagnostic tools have emerged in the industry [23, 14]. Compared with the traditional method of diagnosing fundus images by imaging doctors, the deep learning model can achieve close to doctors' performance through end-to-end optimal learning, to effectively alleviate the doctor deficiency, and make full use of the existing clinical information to speed up the process of fundus analysis.

In recent years, there has been a lot of research on fundus image-based assisted disease diagnosis, and the field of diagnosing high myopia through fundus image is currently in the stage of exploration and development. Due to the intrinsic black box characteristic and lack of interpretability of the deep learning model, even if there is only a single data source, the high diagnosis accuracy still does not have clinical reliability. Therefore, some people have studied the fusion training of multiple similar data sources or text combined images from the perspective of data sources to improve the robustness and accuracy of the model

[24, 27]. However, there are few research results from the perspective of multi-source data in the field of high myopia. For the research on the factors of high myopia, predecessors [5, 29] have carried out exploratory and clinical research on vascular factors, which enlightens us to classify vascular characteristics in the consideration of myopia diagnosis of fundus images. Therefore, using the non-invasive, easily available fundus color image data, from the perspective of multi-source, combined the original image with the blood vessel image based on its analysis. We can learn the multi-modal information through the homogeneous multi-modal neural network, so as to realize the effective diagnosis of the degree of myopia of fundus image. This study attempts to analyze the vascular of the fundus image, and learn the multimodal information through the homogeneous multimodal neural network with the original image, to realize an efficient diagnosis of the degree of myopia in the fundus image.

## 2 Related Work

Vascular is one of the research factors in the diagnosis of ocular diseases. Jonas et al. [22] evaluated the changes of intervertebral Disc Foveal Distance (DFD) in high myopia and found the straightening phenomenon of macular retinal vascular in a 10-year follow-up study based on high myopia. Shin et al. [15] found that in glaucoma patients with high myopia, Perfused Vessel Density (PVD) may be a useful parameter to monitor the progression of high myopia glaucoma. These findings suggest that vascular are a reliable and effective factor for the diagnosis of eye regional diseases, including high myopia.

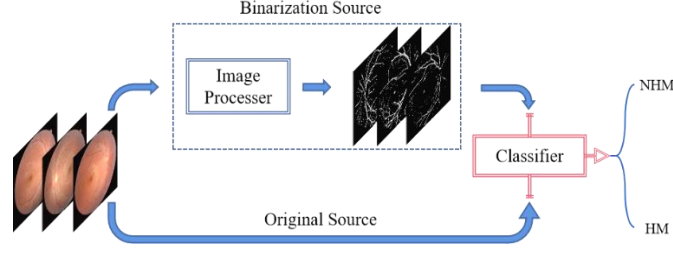
In recent years, the computer vision field of deep learning is mainly divided into two camps. One is the Convolution Neural Network (CNN) structure proposed by Krichevsky et al. [2]. The most representative structure is ConvNeXt proposed by Liu et al. [31] at present, the second is the vision transformer structure that originated by Google [4] and transferred by Dosovitskiy [1]. The research of Reid et al. [12] shows that artificial intelligence has great potential in the application of pediatric ophthalmology, and inspires us to explore this field in combination with computer vision.

Varadarajan et al. [3] proposed a deep learning model to predict the refractive error of retinal fundus images. However, the data source is only middle-aged or elderly patients, which is not suitable for children's ophthalmology. Shi et al. [33] proposed an automatic detection method of myopia in fundus images, but the network lacks the collaborative use of multimodal features. Ting et al. [7] creatively proposed an automatic detection method for pathological myopia and high myopia, which inspired us to adopt a two-classification strategy for children with high myopia. Yavuz et al. [34] proposed an unsupervised method of extracting retinal vascular and enhancing filtration, but it was not used in the diagnosis of retinal diseases. Hemalakshmi et al. [10] enhanced the retina by mixing CNN and Radial Basis Function (RBF) [19] and classified the extracted features into retinopathy, macular degeneration, and other diseases, but did not separate the enhanced features and integrate the multimodal idea, and the specular reflection removal in the preprocessing process will blur the cup features and lose the information of ocular myopia, so it is not suitable for the auxiliary diagnosis of myopia.

In multimodal fusion, according to the characteristics of the modal information to be fused, it can be divided into heterogeneous multimodal fusion (such as image and text) and homogeneous multimodal fusion (such as original image and edge detection image) [30], and homogeneous multimodal fusion has the advantage of sharing feature extraction. Based on this idea, we try to add vascular features into the deep learning model.

## 3 Methodology

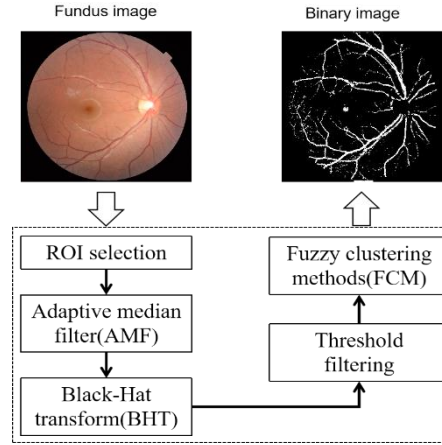
We investigate a new technical scheme for myopia fundus image classification: the homogeneous multimodal classification framework, which is composed of an image processor and a high myopia classifier. The image processor includes the extraction of the region of interest (ROI), the adaptive median filtering (AMF), the black hat transformation (BHT), a binarization, and a few other image processing methods. The classifier is a neural network model with the homogeneous multimodal data as input. The general situation is shown in Fig 1. We will introduce it in detail below.



**Fig. 1.** The homogeneous multimodal classification framework is mainly composed of a processor and a classifier. It input the original image and output the high myopia (HM) or non-high myopia (NHM) classification results.

### 3.1 The Image Processor

In order to explore the effectiveness of vascular skeleton features in myopia classification from the perspective of deep learning, it is necessary to establish processing steps to extract vascular features, including removing irrelevant region interference, reducing image noise and enhancing fundus vascular shape features. Then the shape characteristic curve of vascular is obtained by threshold denoising, morphological transformation, binarization, and extraction. The specific process is shown in Fig 2.

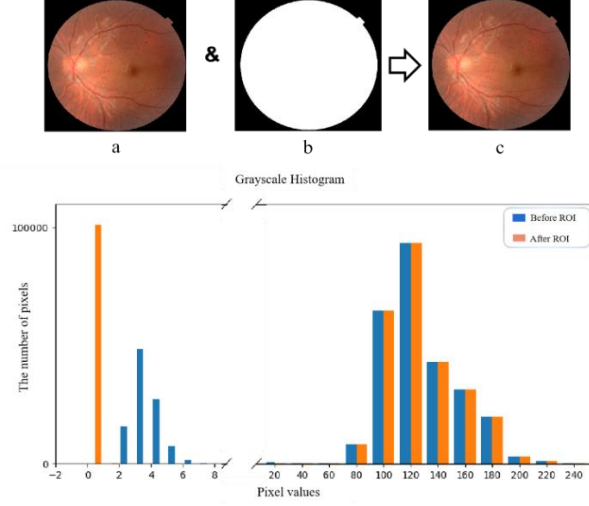


**Fig. 2.** The flow chart of the image preprocessing. It input the original fundus image and generate the binary vascular image.

**Selection of ROI and Noise Reduction.** In order to select the region of interest in fundus image and to eliminate the marginal possible interference, we adopt retinal region mask technology. Firstly, the RGB image is converted into a gray image, and the mask matrix of the ROI area is generated according to equation 1 and is shown in Fig. 3. b.

$$Mask(x, y) = \begin{cases} 0, & v \in (t_1, t_2) \\ 255, & v \in [0, t_1] \cup [t_2, 255] \end{cases} \quad (1)$$

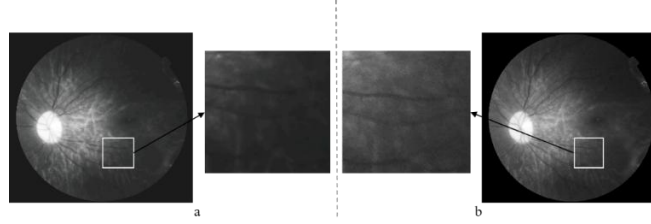
Where  $v$  denotes the gray value,  $t_1$  and  $t_2$  denote the upper and lower thresholds, respectively, and were set to 10 and 255 to achieve the best effectiveness.



**Fig. 3.** '&' represents the logical AND operation, 'a' is the original image, 'b' is the mask image segmented by the threshold, and 'c' is the image after ROI operation. The histogram shows the pixel values before and after ROI operation.

Then we perform logical AND operation on the mask matrix and the original image, so that the pixel value of the uncorrelated marginal region is set to zero and the ROI region remains the original value. As shown in Fig. 3, it can be found that there are a large number of pixels with gray values in the range of 0-10 in the non-ROI area of the image before this step, and the gray value of this area is set to zero after this step. For the ROI area, the pixel value does not change, so the histogram is consistent.

**AMF and BHT.** The most obvious feature of vascular is the fuzziness of the boundary, so it is necessary to protect the boundary information, thus we abandon the linear filter and use the nonlinear filter median filter instead. The probability of noise occurrence in fundus images is relatively high, and the effectiveness of the median filter algorithm is poor. Therefore, it is necessary to dynamically change the window size of the median filter according to the present conditions, to take into account of the effect of denoising and protecting details. As shown in Fig 4, it can be found that the algorithm can enhance the vascular features while protecting the vascular boundary, and has a certain noise reduction effect.

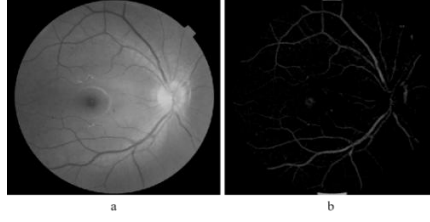


**Fig. 4.** Comparison images before and after AMF. 'a' is the image before AMF and 'b' is the image after processing. The vascular parts of the two sample images are enlarged for observation and comparison.

After that, we observe that the vascular contour area is darker than the surrounding area, that is, its gray value is lower, hence we adopt the BHT operation to highlight the dark vascular contour. To reduce the computational complexity, we change the image from  $3120 \times 2960$  to  $624 \times 624$  without distortion. The BHT is defined here:  $B_{Hat}$  is the output of the BHT,  $C(I_{ori}, E)$  is the morphological closing operation, where  $I_{ori}$  means the original input image and  $E$  is the mask matrix to assist in completing morphological operations. The calculation formula is as follows:

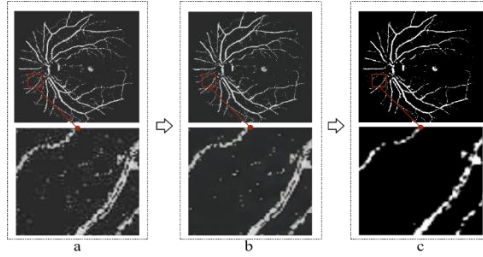
$$B_{Hat} = C(I_{ori}, E) - I_{ori} \quad (2)$$

Where  $C(I_{ori}, E)$  first carries the image dilation operation and then the erosion operation. The dilation operation expands the highlighted area of the image, and the erosion operation makes the highlighted area smaller. As shown in Fig. 5, this operation further enhances the vascular texture features, making the subsequent binary operation more accurate.



**Fig. 5.** Comparison images before and after BHT, in which 'a' is the result of the previous step and is used as the input, and 'b' is the result of BHT.

**Binary Image.** After the above operations, we found that BHT produces the secondary noise side effect. To enhance the texture characteristics of vascular features in the final image, we carry out a secondary noise reduction operation on the image after BHT, that is, a simple threshold filtering operation, by setting the threshold to take the mask matrix within the interval and calculate with the original image by logical AND operation. Finally, the vascular texture features are segmented through fuzzy clustering methods (FCM). The overall process is shown in Fig. 6.



**Fig. 6.** Comparison images during binarization. 'a' is the result of the previous process and the input of this step, 'b' is the effect after threshold filtering, and 'c' is the result of the binary vascular features.

FCM is a clustering method based on a fuzzy set. It uses membership to determine which center each data point belongs to. The specific algorithm process is as follows. We define the number of clustering categories  $\theta$ , initialize clustering center  $C$  and membership matrix  $M$ . Before the number of iterations reaches the threshold, there are three steps:

*Step 1.* Input sample  $x$ , update  $C$  according to the following formula:

$$C_i = \frac{\sum_{j=1}^N M_{ij}^\beta x_j}{\sum_{j=1}^N M_{ij}^\beta} \quad (3)$$

Where  $N$  depends on the image size,  $\beta$  is the membership factor.

*Step 2.* as for the sample  $x$ , update  $M$  according to the following formula:

$$M_{ij} = \left\{ \sum_{k=1}^{\theta} \left( \frac{\|x_j - C_i\|}{\|x_j - C_k\|} \right)^{\frac{1}{\beta-1}} \right\}^{-1} \quad (4)$$

Where  $\|*\|$  denotes the Euclidean distance.

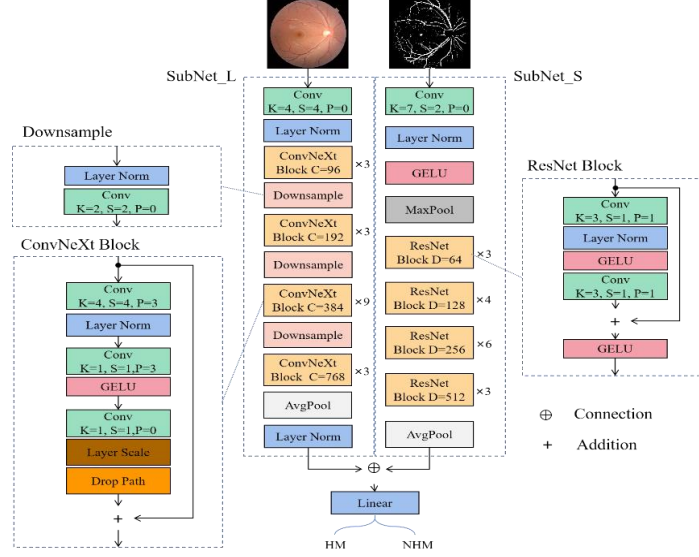
*Step 3.* then calculate  $J$  according to the following formula:

$$J = \sum_{i=1}^{\theta} \sum_{k=1}^N M_{ij}^\beta \|x_k - C_i\|^\beta \quad (5)$$

Finally, the image is clustered, that is, the maximum value of each column is selected from the membership matrix as the attribution domain of the corresponding point, and the clustering result is saved, and then the binary image is constructed according to the clustering result and the clustering center  $C$ .

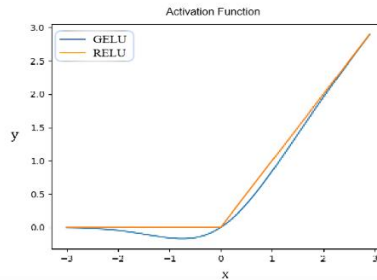
### 3.2 The Classifier – ComboNet

At this stage, we combine the valuable binarized vascular image and the original fundus image as parallel inputs for our classifier, the ComboNet, and since both inputs are image mode, the ComboNet is therefore a homogeneous multimodal convolution neural network. Its output includes two categories, the high myopia (HM) and the non-high myopia (NHM). Two branches of the ComboNet are separately named as Subnet\_L and Subnet\_S, where L represents for large and S for small by the size of the model that will be described later. The detailed structure of ComboNet is shown in Fig 6.



**Fig. 7.** ComboNet consists of a SubNet\_L, a Subnet\_S, and a full connection layer (the Linear), in which 'Conv' means convolution operation and 'K' denotes the size of convolution kernel, 's' denotes the sliding step size of convolution kernel, 'P' is the padding size, 'C' and 'D' are the channels' number of the output of the two subnets respectively, 'GELU' means the activation function, 'MaxPool' and 'AvgPool' correspond to max-pooling and mean pooling respectively, 'Layer Norm' is a kind of the normalization algorithm, 'Layer Scale' is a kind of standardization algorithm[11], 'Drop Path' is a kind of regularization algorithm[8]. The HM and NHM are the final output.

**SubNet\_S.** For the neural network structure of the vascular image path, we refer to the ResNet model studied by He et al. [16]. To align the indicators, we update the normalization algorithm from BatchNorm to LayerNorm [13], which calculates a mean and variance for each batch of data to achieve normalization in the feature dimension, so it does not depend on the batch size and overcomes the shortcomings of the former. The core structure is the ResNetBlock, which has a jump residual connection to transfer the information of the lower layer to avoid the gradient disappearance. As shown in Fig 7, the macro feature analysis process is constructed with the multi-layer design, and the ratio of modules at each layer is 3:4:6:3. In addition, the activation function is updated to GELU [6] considering the pixel characteristics of the binary vascular image.



**Fig. 8.** Comparison diagram of two activation functions ( $x \in [-3, +3]$ )

As can be seen from Fig 8, GELU has a faster zeroing speed than RELU, which means that the zeroing probability of weight value is higher, that is, the distribution of network weight is more biased towards vascular features. Therefore, it helps the model fit the vascular feature information.

**SubNet\_L.** For the neural network structure of the original fundus image path, we still refer to the multi-layer designed ResNet. Its core structure is ConvNeXt Block, which has different feature resolutions in each layer, and the ratio is adjusted to 3:3:9:3 according to Liu et al. [31]. Considering the information redundancy of the image, we downsample the input image and aggregate them to appropriate feature size. The Sampleblock, which can maintain the stability of the model through normalization, and realize the uncovered convolution operation with the same step size and kernel size, is adopted for the downsampling.

Compared with ResNet, ConvNeXtBlock takes use of a convolution kernel with a larger receptive field and has fewer activation functions, in addition, it also replaces BatchNorm with the superior LayerNorm. The LayerScale is for optimization, which is a standardized method to maintain the consistency of layer values and magnitude of the gradient. Specifically, we add a learnable scaling parameter matrix to solve the problem of increasing variance caused by residual connection, defined  $W$  as the learnable scaling parameter matrix,  $LS$  as the output, and  $X$  as the input, then LayerScale formula is as follows:

$$LS = W \cdot \frac{(X - \bar{X})}{\sqrt{Var(X) + \varepsilon}} + b \quad (6)$$

Where  $\bar{X}$  represents the mean value and  $Var$  represents the calculated variance,  $\varepsilon$  is an infinitesimal parameter to prevent the denominator from being zero, and  $b$  is an offset parameter.

ConvNeXtBlock also introduces a regularization method called DropPath. The commonly used regularization method, Dropout, is to multiply each hidden activation layer by an independent Bernoulli random variable, yet it will lose its effectiveness when combined with batch normalization [25]. While DropPath can automatically search dropout mode by adding a learnable hyper-parameter named the survival rate. We define  $\rho$  as the survival rate that starting from  $\rho = 1$ , and end with  $\rho_l$  representing the last survival rate of ConvNeXtBlock. Through a simple linear attenuation rule, the survival rate in the process is calculated as follows,

$$\rho_i = \frac{i}{L}(1 - \rho_l) \quad (7)$$

Where  $i$  represented the order of calculation, and  $L$  means the total number of calculations. The robustness, the generalization ability, as well as the training speed of the neural network, together can be improved through this method.

## 4 Experiment

### 4.1 Dataset

We conduct experiments on a private dataset provided by a regional children's hospital, including 994 labeled fundus images obtained from 561 patients (including rechecked patients). We combine the corresponding text vision screening data, eye axis number, corneal curvature, and other gold standards to label the data, which can ensure the reliability of the data, and finally divide them into two categories of high and not high myopia. For all experiments, we divide the data into a training set and validation set, with a ratio of 8:2. And we take accuracy as a reference index to verify the superiority of ComboNet. As for the classification model  $M$  and the test set  $T$  with size  $n$ , input  $x$ , and label  $y$ , the results of accuracy are as follows :

$$\text{Accuracy}(M; T) = \frac{1}{n} \sum (M(x_i) = y_i) \quad (8)$$

Generally, the gradient descent of parameters is realized through a multivariate loss function to complete the fitting of the model to multi-source data, so as to generate a multi-input and multi-output model framework, however, the network structures of different data are independent of each other. Therefore, we give the task of multivariate loss function to the full connection layer of the model, and form a coupled multi-input and single output model, so as to increase the multi-source model, and then use the univariate loss function. The advantage of this layout is that it does not need to consider the design of appropriate multivariate loss function, but hand over the process of multi-information interaction to the powerful network model. Therefore, for the training strategy with multi-source data, the verification process adopts the verification set of a single data source, and we also set up some ablation experiments for the verification set.

## 4.2 Implementation Details

**Data enhancement.** For the homogeneous multimodal input of our ComboNet, the original image is represented by O, and the binary image is represented by B. The dimension of all initial input images is  $624 \times 624$ , then the image is randomly cropped into different sizes and aspect ratios, and the cropped image is uniformly scaled to  $224 \times 224$ , rotate the image horizontally at random angle, and finally normalize and send to the neural network as input.

**Optimization Strategy.** We introduce the idea of transfer learning to better train the model and load the corresponding pre-trained weights generated from the public dataset of ImageNet for each experiment for optimization. The Imagenet project is a large visual database for visual object recognition software research, which contains more than 14 million labeled images. For all experiments, we take the same optimizer and training iteration number, and the learning rate varies according to the characteristics of the model. Experiments are carried out on Intel CPU and NVIDIA GPU. Since the memory size of our GPU is 8GB, we use the batch method for training. According to the model size, we set the batch size to 16 or 8, and set it to 4 on some large models.

**Training strategy.** The primary problem in a multi-source neural network is that the fitting speed of different data sources is different, but the training process is synchronous. Therefore, we freeze the training parameters after one data source is saturated. We also consider the uncertainty caused by different combinations between data sources and neural networks, so we carry out comparison experiments on all combinations to select the optimal one. We also visualize the last layer of each network to add interpretability and to understand the hidden information as much as possible.

## 5 Results

### 5.1 Single data source

We trained some traditional CNN based neural networks with O and B datasets with 50 iterations under the condition of single-source input such as ResNet, DenseNet [9], ShuffleNet [21], and the outstanding network structures based on CNN in recent two years such as RegNet [12], EfficientNet [20] and ConvNeXt [31]. In addition, the ViTransformer [1] and SwinTransformer [32] that based on the transformer are also included for the comparison experiments.

**Table 1.** Results of the model with single-source

Base	Model	Source & Index			
		O		B	
		Accuracy	Iteration	Accuracy	Iteration
Transformer	Vision Transformer	0.833	32	0.833	18
	Swin Transformer	0.915	40	0.889	27
CNN	DenseNet	0.914	33	0.873	15
	EfficientNet	0.899	46	0.854	29
	Shufflenet	0.914	17	0.838	11
	RegNet	0.884	35	0.889	19
	ResNet	0.924	14	<b>0.899</b>	6
	ConvNeXt	<b>0.925</b>	26	0.893	25

It can be seen from Table 1 that in the application scenario of fundus images and mode of single-source input, the network structure based on CNN is better than the Transformer. In terms of accuracy, 0.925 of ConvNeXt is the best fitting result for the dataset O, and 0.899 of traditional ResNet is the best fitting result for dataset B, so our homogeneous multi-source neural network is constructed based on these results. In terms of iteration, it can be found that the rate of optimal fitting of the dataset is different in the training processes, that is, the fitting speed of dataset B is generally faster than that of O. Therefore, we freeze the network parameters of training path of dataset B to prevent from overfitting. After experimental validation, we finally choose to freeze the parameters in the 15th iteration for SubNet\_S and the 20th iteration for SubNet\_L of dataset B with a total number of 50 iterations. Timings of frozen iterations vary with the data.



## 5.2 Multiple Data sources

As mentioned before, we validated the results with different settings of single-source. To further verify the structural rationality and effectiveness of ComboNet, we set up the ablation experiments through multi-source input. The specific results are shown in Table 2.

**Table 2.** Multi-source ablation experiments

Setting No.	Path	Source	Validation / Data	Accuracy
1	SubNet_L	O	SubNet_L / O	<b>0.934</b>
	SubNet_S	B		
2	SubNet_L	B	SubNet_L / B	0.875
	SubNet_S	O		
3	SubNet_L	O	SubNet_S / B	0.899
	SubNet_S	B		
4	SubNet_L	B	SubNet_S / O	0.907
	SubNet_S	O		


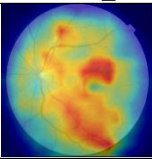
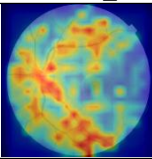

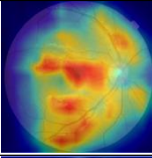
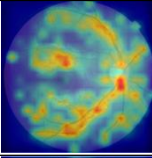

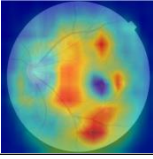
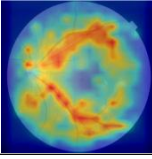
By observing setting No.1 and No.3 we can find that on the premise of training SubNet\_L with dataset O, the accuracy of the validation set from O is higher (0.934:0.899), which is consistent with the case of the generally low accuracy of B in Table 1, the single-source experiments. In addition, by observing No.2 and No.4 we can find that accuracy of a validation set from O is still higher (0.907:0.875) on the premise of training SubNet\_L with B.

Furthermore, observing No.1 and No.4, the accuracy of the validation set from O is higher on the premise of training SubNet\_L with O (0.934:0.907). Besides, observing No.2 and No.3, the accuracy is also higher (0.875:0.899) on-premise of using the validation set from B and training SubNet\_L with O, although the difference is very low. Combining Table I and Table II, we can find that SubNet\_L is more suitable for training O and SubNet\_S is more suitable for B.

## 5.3 Interpretability analysis

From the perspective of interpretability, on the premise of training SubNet\_L with O and training SubNet\_S with B, we visualize their last results of convolution layer respectively. As shown in Table III, the wavelength of color from small (blue) to large (red) indicates greater weight.

**Table 3.** Visualization heat map

	Origin	SubNet_L	SubNet_S
Exp 1			
Exp 2			
Exp 3			

It can be found that the weight of the SubNet\_S trained by dataset B is mainly gathered around the vascular features, and the weight of SubNet\_L trained by the O dataset is mainly concentrated in the area between vascular. This confirms that our model can learn more information than a single-source, that is, our homogeneous multi-source neural network has more advantages.

## 6 Conclusion

In the application of myopia classification of children's fundus images, we propose a new homogeneous multimodal method, which allows the model to learn the information of the original image and the information of vascular features at the same time, to distinguish NHM and HM. The results of different experiments show that our method achieves the best performance. Since there is no homogeneous multimodal scheme of fundus image combined with vascular features, our method is of great significance. This good result provides a good prospect for the clinical automatic detection of myopia in children.

**Acknowledgements.** This study is supported by the National Natural Science Foundation of China with the project no. 81970844 and the National Key R&D Program of China with project no. 2020YFB2104402

## References

1. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and Others, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
2. A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
3. A. V. Varadarajan, R. Poplin, K. Blumer, C. Angermueller, J. Ledsam, R. Chopra, P. A. Keane, G. S. Corrado, L. Peng, and D. R. Webster, "Deep learning for predicting refractive error from retinal fundus images," *Investigative ophthalmology & visual science*, vol. 59, pp. 2861--2868, 2018.
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
5. D. Cheng, Q. Chen, Y. Wu, X. Yu, M. Shen, X. Zhuang, Z. Tian, Y. Yang, J. Wang, F. Lu, and Others, "Deep perifoveal vessel density as an indicator of capillary loss in high myopia," *Eye*, vol. 33, pp. 1961--1968, 2019.
6. D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
7. D. S. J. Ting, V. H. Foo, L. W. Y. Yang, J. T. Sia, M. Ang, H. Lin, J. Chodosh, J. S. Mehta, and D. S. W. Ting, "Artificial intelligence for anterior segment diseases: Emerging applications in ophthalmology," *British Journal of Ophthalmology*, vol. 105, pp. 158--168, 2021.
8. G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," Springer, 2016, pp. 646--661.
9. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2017, pp. 4700--4708.
10. G. R. Hemalakshmi, D. Santhi, V. Mani, A. Geetha, and N. B. Prakash, "Classification of retinal fundus image using MS-DRLBP features and CNN-RBF classifier," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8747--8762, 2021.
11. H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. E. J E Gou, "Going deeper with image transformers," 2021, pp. 32--42.
12. J. E. Reid and E. Eaton, "Artificial intelligence for pediatric ophthalmology," *Current opinion in ophthalmology*, vol. 30, pp. 337--346, 2019.
13. J. L. Ba, J. R. Kiros and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
14. J. Lee, S. Jun, Y. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: general overview," *Korean journal of radiology*, vol. 18, pp. 570--584, 2017.
15. J. W. Shin, J. Kwon, J. Lee, and M. S. Kook, "Relationship between vessel density and visual field sensitivity in glaucomatous eyes with high myopia," *British Journal of Ophthalmology*, vol. 103, pp. 585--591, 2019.
16. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016, pp. 770--778.
17. L. Liu, R. Li, D. Huang, X. Lin, H. Zhu, Y. Wang, X. Zhao, X. Zhang, and H. Liu, "Prediction of premyopia and myopia in Chinese preschool children: a longitudinal cohort," *BMC ophthalmology*, vol. 21, pp. 1--10, 2021.
18. M. D. Abr A Moff, M. K. Garvin and M. Sonka, "Retinal imaging and image analysis," *IEEE reviews in biomedical engineering*, vol. 3, pp. 169--208, 2010.
19. M. T. Musavi, W. Ahmed, K. H. Chan, K. B. Faris, and D. M. Hummels, "On the training of radial basis function classifiers," *Neural networks*, vol. 5, pp. 595--603, 1992.
20. M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," PMLR, 2021, pp. 10096--10106.
21. N. Ma, X. Zhang, H. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," 2018, pp. 116--131.
22. R. A. Jonas, Y. N. Yan, Q. Zhang, Y. X. Wang, and J. B. Jonas, "Elongation of the disc-fovea distance and retinal vessel straightening in high myopia in a 10-year follow-up of the Beijing eye study," *Scientific Reports*, vol. 11, pp. 1--8, 2021.
23. R. Kapoor, S. P. Walters and L. A. Al-Aswad, "The current state of artificial intelligence in ophthalmology," *Survey of ophthalmology*, vol. 64, pp. 233--240, 2019.

24. S. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ digital medicine*, vol. 3, pp. 1--9, 2020.
25. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," PMLR, 2015, pp. 448--456.
26. W. H. Organization and Others, "World report on vision," 2019.
27. X. Wang, L. Dai, S. Li, H. Kong, B. Sheng, and Q. Wu, "Automatic grading system for diabetic retinopathy diagnosis using deep learning artificial intelligence software," *Current Eye Research*, vol. 45, pp. 1550--1555, 2020.
28. Y. Ikuno, "Overview of the complications of high myopia," *Retina*, vol. 37, pp. 2347--2351, 2017.
29. Y. Shi, L. Ye, Q. Chen, G. Hu, Y. Yin, Y. Fan, J. Zhu, J. He, Z. Zheng, H. Zou, and Others, "Macular vessel density changes in young adults with high myopia: A longitudinal study," *Frontiers in medicine*, vol. 8, 2021.
30. Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4835--4845, 2020.
31. Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.
32. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, pp. 10012--10022.
33. Z. Shi, T. Wang, Z. Huang, F. Xie, and G. Song, "A method for the automatic detection of myopia in Optos fundus images based on deep learning," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 37, p. e3460, 2021.
34. Z. Yavuz and C. K O Se, "Blood vessel extraction in color retinal fundus images with enhancement filtering and unsupervised classification," *Journal of healthcare engineering*, vol. 2017, 2017.