**World Scientific**
www.worldscientific.com

# Computer-aided diagnosis of retinopathy based on vision transformer

Zhencun Jiang*, Lingyang Wang*, Qixin Wu*, Yilei Shao†,
Meixiao Shen†, Wenping Jiang*,‡,¶ and Cuixia Dai*,§,¶
*Shanghai Institute of Technology
100 Haiquan Road, Shanghai 201418, China

†School of Ophthalmology and Optometry
Wenzhou Medical University
Xueyuan Road 270, Wenzhou, Zhejiang 325027, China
‡jiangwenping@sit.edu.cn
§sdadai7412@163.com

Age-related Macular Degeneration (AMD) and Diabetic Macular Edema (DME) are two common retinal diseases for elder people that may ultimately cause irreversible blindness. Timely and accurate diagnosis is essential for the treatment of these diseases. In recent years, computer-aided diagnosis (CAD) has been deeply investigated and effectively used for rapid and early diagnosis. In this paper, we proposed a method of CAD using vision transformer to analyze optical coherence tomography (OCT) images and to automatically discriminate AMD, DME, and normal eyes. A classification accuracy of 99.69% was achieved. After the model pruning, the recognition time reached 0.010 s and the classification accuracy did not drop. Compared with the Convolutional Neural Network (CNN) image classification models (VGG16, Resnet50, Densenet121, and EfficientNet), vision transformer after pruning exhibited better recognition ability. Results show that vision transformer is an improved alternative to diagnose retinal diseases more accurately.

Keywords: Vision transformer; OCT; image classification; retinopathy; computer-aided diagnosis; model pruning.

## 1. Introduction

Age-related macular degeneration (AMD) and diabetic macular edema (DME) are two common retinal diseases that occur among the elders. AMD mostly emerges in people over 45 years old and the prevalence of AMD disease increases with age gradually.[1-3] DME, which is the direct cause of vision impairment in most diabetic patients, has a

¶Corresponding authors.

high incidence all over the world.[4,5] Timely and accurate diagnosis of AMD and DME is an important premise for effective treatment *in lieu* of irreversible blindness (to prevent irreversible blindness). The noninvasive imaging through the three-dimensional cellular cytolytic tissue, synchronous medical imaging technology, can improve early medical diagnosis and contribute to the development of novel therapies. Optical coherence tomography (OCT) is a noninvasive, high-resolution optical medical diagnostic imaging technique, which enables *in vivo* cross-sectional tomographic visualization of the internal microstructure of eyes. In recent years, it has become a golden standard for eye disease diagnosis.[6–12]

Diagnosing retinal diseases using OCT images is currently the most effective CAD of retinopathy. In recent years, deep learning has developed rapidly, and it has provided many solutions in the field of medical diagnosis. Especially, Convolutional Neural Network (CNN) has achieved great success in medical image classification. A large number of researchers have made great achievements in the field of using CNN to classify OCT images for disease diagnosis. Shih *et al.* proposed to classify four types of OCT images using the VGG16 model, which involved three retinal diseases and normal retina. The accuracy of the model on the data test set reached 99%. The results demonstrate the effectiveness of the deep learning algorithm in the diagnosis of retinal diseases.[13] Zhang Quan *et al.* proposed a multi-scale deep learning model using OCT images to identify diabetic edema macula, and the model achieved an accuracy rate of 94.5%. By comparison with other models, the author proved that this model provided a better ability to recognize low-quality medical images.[14] Saratxaga Cristina *et al.* used deep learning algorithm to classify OCT images of mouse colons, and the recognition accuracy of their model reached 96.65%, which effectively proved the excellent performance of deep learning in computer-aided diagnosis (CAD) of colonic lesions.[15] Luo Yuemei used OCT images combined with deep learning to assist in the diagnosis of sebaceous glands. The classification accuracy of the model proposed in this paper reached 97.9%, which is of great help to medical personnel in the diagnosis and treatment of serum problems.[16] Potapenko Ivan *et al.* used noisy OCT image data sets for deep learning training to complete the identification task of AMD disease. The recognition

accuracy of the model reached 90.9%, which proved that the model could assist medical personnel in the clinical diagnosis of AMD.[17]

It should be noticed that most of the above-mentioned studies are based on CNN. CNN has the advantage of being able to extract image features very well, which has been verified by a large number of scholars. However, there are still little researches on vision transformer CAD. Vision transformer is a new image classification model that has been proposed in 2020.[18] It does not rely on any CNN, but it is completely performed based on the transformer structure, which has a different feature extraction method from CNN. Yakoub Bazi *et al.* have proved that vision transformer has better classification capabilities than CNN to solve image classification problems. They applied vision transformer to remote sensing image classification and tested multiple remote sensing image data sets. Experimental results showed that the vision transformer classification accuracy rate of remote sensing images exceeds the results based on the CNN.[19]

In this paper, we integrated vision transformer and OCT images to improve the diagnosis of retinal diseases. Aiming at the two common retinal diseases of AMD and DME, we collected OCT images of AMD and DME diseases, as well as images of normal ocular fundus. We then trained the vision transformer to classify these three types of OCT images in order to realize the diagnosis of retinal diseases. In Sec. 2, we introduced the dataset and vision transformer used in this paper. In Sec. 3, we analyzed the experimental results of this paper, discussed the impact of model pruning, and compared the results with that using the CNN. The full study was summarized in Sec. 4.

## 2. Materials and Methods

### 2.1. *Dataset*

The dataset used in this paper comes from the OCT fundus images of 15 normal people, 15 AMD patients, and 15 DME patients collected by Duke University.[20] The dataset includes 1407 OCT images of normal retinal, 723 OCT images of AMD, and 1101 OCT images of DME. Examples of these three types of OCT images are shown in Fig. 1.

As shown in Fig. 1(a), normal retina OCT images provide the inner retinal layers including the nerve fiber layer (NFL), ganglion cell layer (GCL), inner
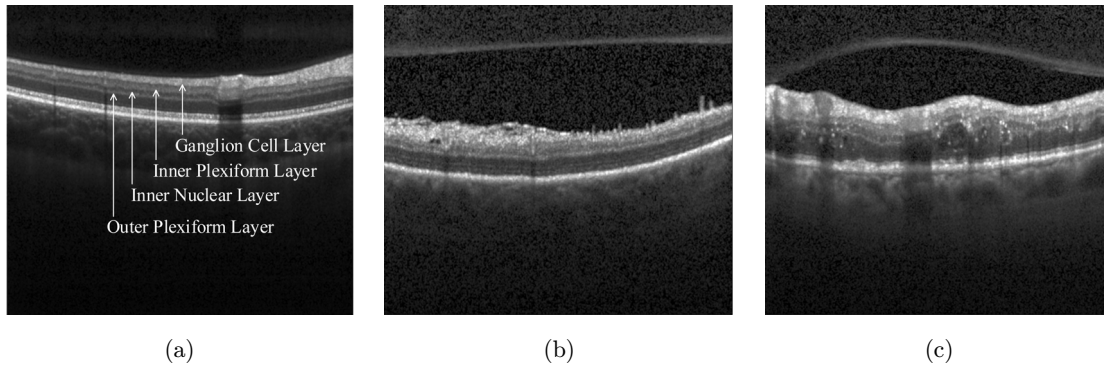
Fig. 1. Examples of these three types of OCT images. (a) Normal, (b) AMD, and (c) DME.

plexiform layer (IPL), inner nuclear layer (INL), and the outer layers including the outer plexiform layer (OPL), outer nuclear layer (ONL), myoid and ellipsoid zone (MEZ), the outer segment of photoreceptors (OS) and the retinal pigment epithelium (RPE). OCT images of AMD fundus shown in Fig. 1(b) present a large number of drusen/choroidal neovascularization (CNV) formations, which results in submacular hemorrhage and leakage. As we can see from Fig. 1(c), OCT images of DME present with retinal edema, hemorrhage, cystic macular edema, and subretinal fluid. OCT fundus images of AMD shown in Fig. 1(b) present a large number of larger drusen/CNV formations, which results in submacular blood and protein leakage. We can see from Fig. 1(c) that the OCT fundus images of DME are present with retinal swelling, hemorrhage, cystic macular edema, and subretinal fluid.

### 2.2. Data preprocessing

It is necessary to preprocess the images before inputting the images to model for recognition. Image resizing and normalization are the main data processing steps. In this study, the images' size was adjusted to $224 \times 224$, and then the outputs of them were normalized.[21–23]

### 2.3. Vision transformer

Vision transformer is completely implemented based on the transformer structure, which is widely used in the field of Natural Language Processing (NLP). The transformer structure consists of a set of encoder components and a group of decoder components. Vision transformer is an image classification model and does not require decoder

components, so there is only an encoder component in the vision transformer.

The encoder component is composed of a stack of six identical encoders. Each encoder consists of a multi-head attention layer and a feed-forward layer. Both layers contain residual connection and Layer-Norm. MLP is a multilayer perceptron. The structure of an encoder component is shown in Fig. 2.

The multi-head attention in the encoder component structure is a self-attention structure, which allows the model to focus on different aspects of information. Scaled dot-product attention is the attention score calculation method of multi-head attention. The structure of the multi-head attention and the structure of scaled dot-product attention are shown in Fig. 3.
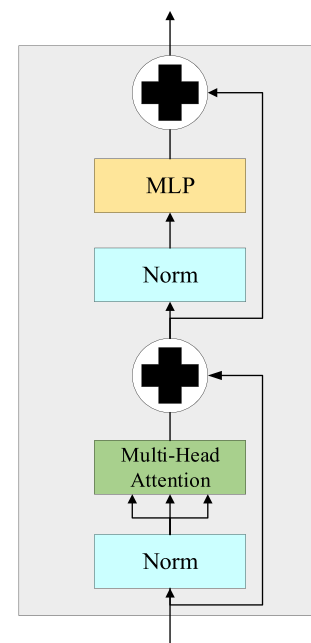


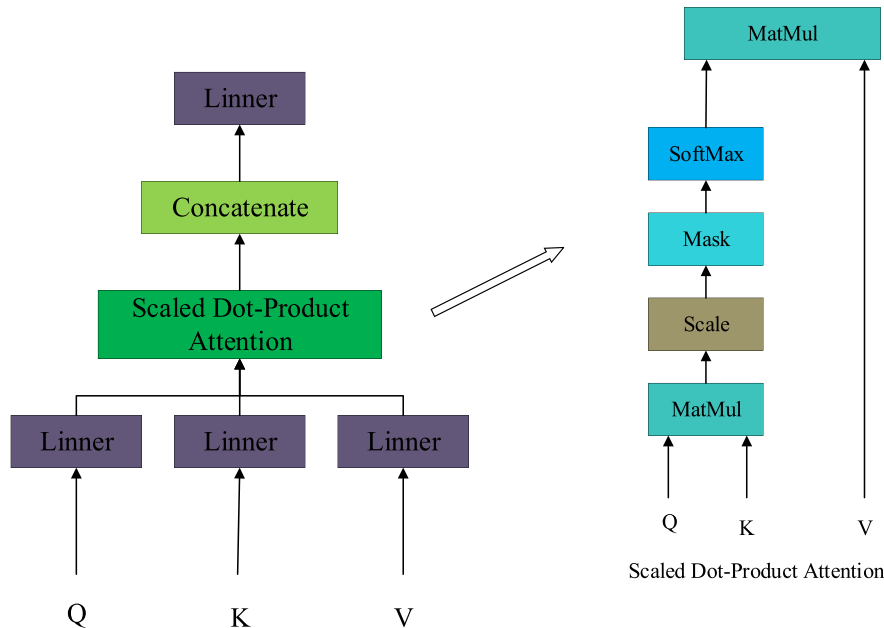Fig. 2. The structure of an encoder component.

Fig. 3.    The structure of the multi-head attention and the structure of scaled dot-product attention.

Multi-head attention is shown in the following equations:

$$Q_i = \mathrm{QW}_i^Q, \quad K_i = \mathrm{KW}_i^K,$$
$$V_i = \mathrm{VW}_i^V, \quad i = 1,\ldots,8, \tag{1}$$

$$\mathrm{head}_i = \mathrm{Attention}(Q_i, K_i, V_i), \quad i = 1,\ldots,8, \tag{2}$$

$$\mathrm{MultiHead}(Q, K, V)$$
$$= \mathrm{Concact}(\mathrm{head}_1,\ldots,\mathrm{head}_8)W^o, \tag{3}$$

where $Q$ represents a query vector, $K$ represents a key vector, $V$ represents a value vector, and $W$ represents a weight matrix.

Linear embedding layer is a crucial structure in vision transformer. The images are divided into multiple patches after being input into vision transformer, and each patch is implemented for patch embedding processing. Embedding is a spatial mapping method commonly used in NLP, which maps high-dimensional vectors to low-dimensional spaces. Patch embedding is the embedding operation of a patch, and each patch is flattened into a one-dimensional tensor. After the patch embedding operation, positional embedding and class embedding are added to the transformer encoder. After being output by the transformer encoder, it will go through an MLP head structure, which is composed of a fully connected layer and a Gaussian Error Linear Unit (GELU) activation function. The structure of the MLP is shown in Fig. 4.
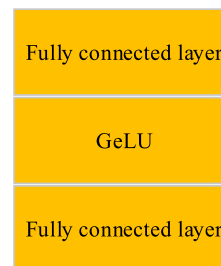


Fig. 4.    The structure of the MLP.

The equation of the GELU activation function is

$$\mathrm{GELU}(x)$$
$$= 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right). \tag{4}$$

After the output of the MLP head structure and the output of a layer of full connection layer, the model carries out the classification task. In this study, three types of data were classified, so the full connection layer output of vision transformer was changed to three categories. The structure of vision transformer is shown in Fig. 5.

### 2.4.  *Symmetric cross-entropy loss function*

The loss function used in this paper is the asymmetric cross-entropy loss function. The symmetrical
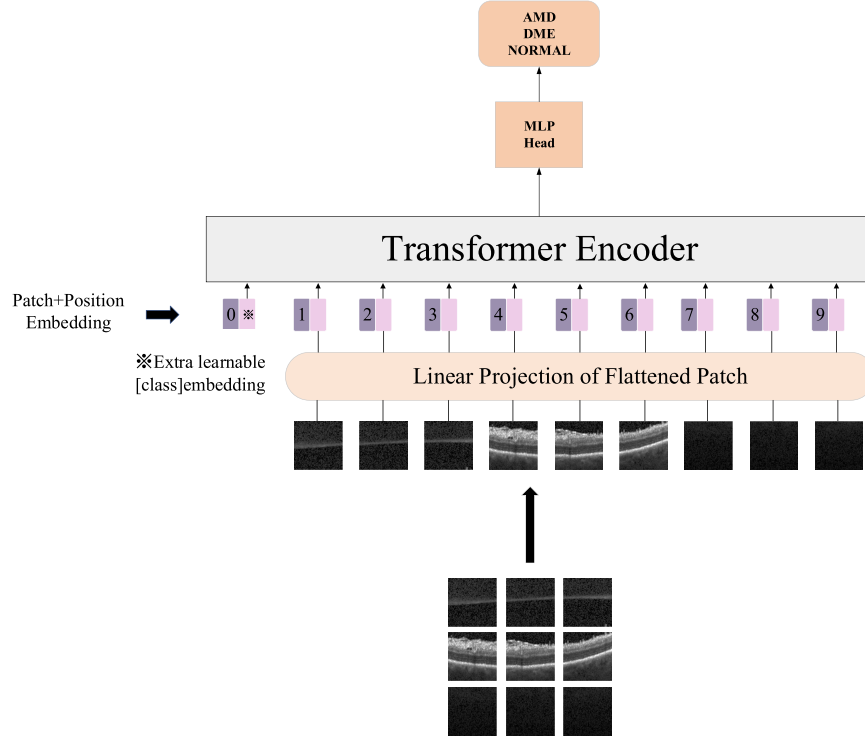
Fig. 5. The structure of vision transformer.

cross-entropy loss function can reduce the impact of noise in the data set on training and can prevent overfitting.[24] The definition equation of the symmetric cross-entropy loss function is

$$l_{\text{sce}} = l_{\text{ce}} + l_{\text{rce}}, \tag{5}$$

where $l_{\text{ce}}$ is the cross-entropy loss function and $l_{\text{rce}}$ is the reverse cross-entropy function. The definitions of $l_{\text{ce}}$ and $l_{\text{rce}}$ are, respectively,

$$l_{\text{ce}} = \frac{1}{N} \sum_i -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)], \tag{6}$$

$$l_{\text{rce}} = -\sum_{k=1}^{K} p(k \,|\, x) \log q(k \,|\, x). \tag{7}$$

## 2.5. *Experimental environment*

The experimental environment of this paper is as follows: The hardware environment is Intel Core i7-9700f processor, NVIDIA RTX2060s 8 GB graphics card and 16 GB memory; the software environment is Win10 system, Python 3.7, and the deep learning framework used is PyTorch.[25]

## 2.6. *Evaluation standard*

In this paper, accuracy was adopted as the standard to evaluate the performance of the model. The definition of accuracy is shown in Eq. (8). In addition, we also analyzed the recognition ability of each type of OCT image by using the confusion matrix, which is the comparison matrix between the predicted results and the real results,

$$\text{ACC} = \frac{(\text{TP} + \text{TN})}{\text{ALL}}, \tag{8}$$

where ACC represents the ratio of all true predictions of positive, TP represents true positive, TN represents true negative, and ALL represents all predicted samples.

## 2.7. *Experimental procedure and hyperparameter settings*

The experimental process and parameter settings of this paper were as follows:

(1) The data set was divided into training set, validation set, and test set in an 8:1:1 ratio. The training set was used for training the model, the validation set was used for verifying

whether the model had been fitted, and the test set was used for testing the classification ability of the model.

(2) The data of the model were preprocessed.

(3) We used the Ranger optimizer as the optimizer; the learning rate was set as 0.00003, which is the commonly used learning rate for the training classification model; we trained the model 100 times, which is sufficient for the model with small data set.

(4) The parameters of the model were saved with the highest accuracy of the model on the validation set.

(5) The test was performed on the test set and the accuracy of the test set was obtained.

## 3. Results and Discussion

### 3.1. *Experimental results*

The experiment was performed according to the experimental process described in Sec. 2.7, and the error changes of the training set and the accuracy changes of the validation set were recorded, which were used as the basis for analyzing the fitting ability of the model. The error changes of the training set and the accuracy changes of the validation set were shown in Fig. 6.

It can be seen that the error on the training set dropped very quickly and tends to be stable after the 40th epoch. At the same time, the accuracy rate on the validation set rises very fast and tends to be stable after the 40th epoch. As shown by the error change on the training set and the accuracy change on the validation set, it can be determined that the vision transformer fits normally without overfitting.

The accuracy of the model on the test set is 99.69% and the confusion matrix of the test set is shown in Fig. 7.

Meanwhile, we can conclude that the vision transformer's classification of OCT fundus images for AMD and DME is completely accurate, which indicates that there will be no missed diagnosis in practical applications. For the OCT images of normal fundus, there is one case of classification error. In the actual diagnosis, a patient will have multiple OCT images for diagnosis, so the misclassification of one normal OCT image will not affect the final diagnosis of retinopathy.

In order to verify the effectiveness of the method proposed in this paper, the model presented in this paper was compared with the traditional CNN. We chose VGG16,[26] Resnet50,[27] Densenet121,[28] and EfficentNet[29] as the comparison models. The experimental procedures and hyperparameter settings of these models were consistent with the vision transformer.

Model pruning is a method to reduce the memory demand and calculation demand of the model and improve the calculation speed of the model. By encouraging the sparsity of channels in the vision transformer, the important channels are selected and those channels with zero or small coefficients are discarded to achieve an efficient classification model.[30] In this paper, the model pruning is performed on the vision transformer to improve the recognition speed of a single OCT image under the condition that the recognition ability of the model does not decrease significantly.

The accuracy comparison of the test set between vision transformer, vision transformer after pruning, and four CNN models and the recognition time comparison of a single image are shown in Table 1.
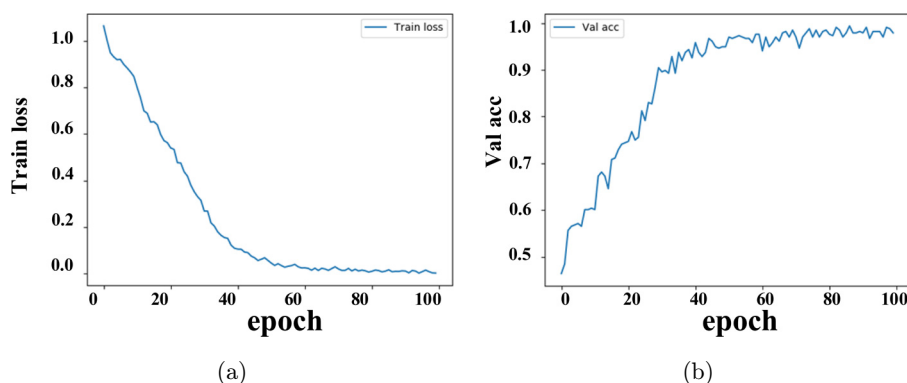


|  (a)  |  (b)  |

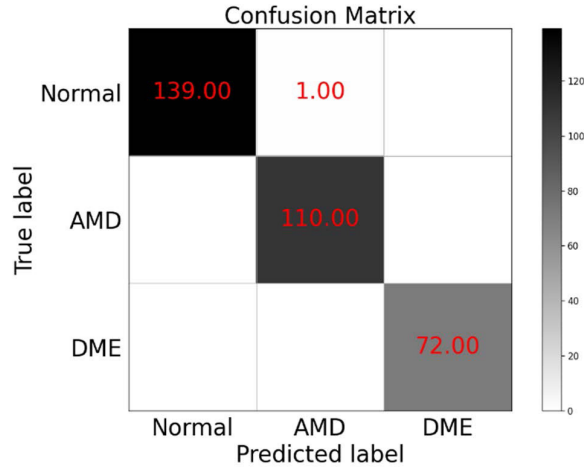Fig. 6.   (a) The error changes of the training set; (b) the accuracy changes of the validation set.

Fig. 7. The confusion matrix of the test set.

Table 1. Model comparison.

| Model | Accuracy | Time |
|---|---|---|
| VGG16 | 98.51% | 0.014 s |
| Resnet50 | 97.32% | 0.089 s |
| Densenet121 | 97.02% | 0.091 s |
| Efficientnet | 34.16% | 0.032 s |
| Vision transformer | **99.69%** | 0.017 s |
| Vision transformer after pruning | **99.69%** | **0.010 s** |

As we can see from Table 1, among the CNN image classification models, VGG16 has the highest classification accuracy of 98.51%, but it is still inferior to the classification accuracy of vision transformer. With attention mechanism, vision transformer can focus on image regions that are semantically relevant to the classification target, so it can obtain higher accuracy. In terms of the recognition speed of a single image, both VGG16 and vision transformer are faster than other CNN models in the recognition speed of a single image. The recognition speed of a single image of vision transformer after pruning is the fastest; it only costs 0.010 s. This speed is faster than all other models, and its recognition accuracy does not decrease, remaining at 99.69%. Considering the recognition accuracy and recognition speed, vision transformer after pruning is superior to CNN models in recognizing OCT images of fundus.

In order to further verify the validity of the model proposed in this paper, we compare the results of the model in this paper with the research results of related literature, and the comparison is shown in Table 2.

Table 2. Comparison of classification results with previous authors.

| Model | Methods | Diseases | Accuracy |
|---|---|---|---|
| Literature[20] | SVM | AMD DME | 95.56% |
| Literature[31] | Dictionary-learning | AMD DME | 98.38% |
| Literature[32] | CNN | AMD DME | 91.33% |
| Literature[33] | CNN | AMD DME | 96.66% |
| Literature[34] | CNN | AMD DME | 94.20% |
| Literature[35] | CNN | DME | 96% |
| This paper | Vision transformer | AMD DME | **99.69%** |

Table 2 shows that Refs. 20 and 31 used machine learning algorithms and the accuracy of the model using dictionary-learning reached 98.38%, which exceeds the results of other studies using CNN. In the literature using CNN, we can see that the recognition accuracy of the model using multiscale and multipath CNN architecture is higher than that of general CNN models. The classification accuracy of the model proposed in this paper is the highest, surpassing the other researches, which proves that the recognition ability of vision transformer for OCT fundus images is stronger than CNN models and traditional machine learning algorithms.

## 4. Conclusions

AMD and DME are two retinal diseases that seriously harm the eyesight of the elderly. Timely diagnosis and treatment are very important for patients. In this paper, we presented a CAD method using vision transformer to classify OCT fundus images, which effectively diagnosed AMD and DME retinal diseases. The data set used in this paper came from the OCT fundus images collected by Duke University. The loss function and optimizer of the model in this paper selected the symmetric cross-entropy loss function and the Ranger optimizer. The classification accuracy of our model for AMD, DME, and normal OCT fundus images reached 99.69%. Then, we pruned the model and compared the effect of vision transformer after pruning with the CNN image classification models. We proved that vision transformer had the best recognition ability, and vision transformer after pruning had the fastest recognition speed without any decrease in recognition accuracy. The model proposed in this paper can better realize the CAD of retinal diseases.

## References

1. J. Z. Nowak, "Age-related macular degeneration (AMD): Pathogenesis and therapy," *Pharmacol. Rep.* **58**(3), 353 (2006).
2. K. L. Pennington, M. M. DeAngelis, "Epidemiology of age-related macular degeneration (AMD): Associations with cardiovascular disease phenotypes and lipid factors," *Eye Vis.* **3**(1), 1–20 (2016).
3. K. Kaarniranta, A. Salminen, A. Haapasalo *et al.*, "Age-related macular degeneration (AMD): Alzheimer's disease in the eye?" *J. Alzheimer's Dis.* **24**(4), 615–631 (2011).
4. N. Bhagat, R. A. Grigorian, A. Tutela *et al.*, "Diabetic macular edema: Pathogenesis and treatment," *Surv. Ophthalmol.* **54**(1), 1–32 (2009).
5. J. Ding, T. Y. Wong, "Current epidemiology of diabetic retinopathy and diabetic macular edema," *Curr. Diabetes Rep.* **12**(4), 346–354 (2012).
6. Y. Lin *et al.*, "A measurement of epidermal thickness of fingertip skin from OCT images using convolutional neural network," *J. Innov. Opt. Health Sci.* **14**(1), 2140005 (2021).
7. H. Li *et al.*, "ID-OCTA: OCT angiography based on inverse SNR and decorrelation features," *J. Innov. Opt. Health Sci.* **14**(1), 2130001 (2021).
8. Y. Yu *et al.*, "Monitoring of edema progression in permanent and transient MCAO model using SS-OCT," *J. Innov. Opt. Health Sci.* **14**(1), 2140006 (2021).
9. N. Qin *et al.*, "Research on optical properties of cardiovascular tissues based on OCT data," *J. Innov. Opt. Health Sci.* **14**(1), 2140007 (2021).
10. F. Du, L. Gao, L. Li *et al.*, "In vivo evaluation of laser-induced choroidal neovascularization in rats simultaneously using optical coherence tomography and photoacoustic microscopy," *J. Innov. Opt. Health Sci.* **14**(3), 2140012 (2021).
11. M. R. Ibrahim, K. M. Fathalla, S. M. Youssef, "HyCAD-OCT: A hybrid computer-aided diagnosis of retinopathy by optical coherence tomography integrating machine learning and feature maps localization," *Appl. Sci.* **10**(14), 4716 (2020).
12. D. S. Kermany, M. Goldbaum, W. Cai *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell* **172**(5), 1122–1131 (2018).
13. F. Y. Shih, H. Patel, "Deep learning classification on optical coherence tomography retina images," *Int. J. Pattern Recognit. Artif. Intell.* **34**(8), 2052002 (2020).
14. Q. Zhang, Z. Liu, J. Li *et al.*, "Identifying diabetic macular edema and other retinal diseases by optical coherence tomography image and multiscale deep learning," *Diabetes Metab. Syndr. Obes. Targets Ther.* **13**, 4787 (2020).
15. C. L. Saratxaga, J. Bote, J. F. Ortega-Morán *et al.*, "Characterization of optical coherence tomography images for colon lesion differentiation under deep learning," *Appl. Sci.* **11**(7), 3119 (2021).
16. Y. Luo, X. Wang, X. Yu *et al.*, "Imaging sebaceous gland using optical coherence tomography with deep learning assisted automatic identification," *J. Biophoton.* **14**(6), e202100015 (2021).
17. I. Potapenko, M. Kristensen, B. Thiesson *et al.*, "Detection of oedema on optical coherence tomography images using deep learning model trained on noisy clinical data," *Acta Ophthalmol.* (2021).
18. A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929 (2020).
19. Y. Bazi, L. Bashmal, M. M. A. Rahhal *et al.*, "Vision transformers for remote sensing image classification," *Remote Sens.* **13**(3), 516 (2021).
20. P. P. Srinivasan, L. A. Kim, P. S. Mettu *et al.*, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomed. Opt. Exp.* **5**(10), 3568 (2014).
21. Z. Jiang *et al.*, "Recognition of rice leaf diseases and wheat leaf diseases based on multi-task deep transfer learning," *Comput. Electron. Agric.* **186**, 106184 (2021).
22. J. Zhen-Cun, W. E. N. Xiao-Jing, D. Zheng-Xin *et al.*, "Research on fire detection of improved VGG16 image recognition based on deep learning," *Fire Sci. Technol.* **40**(3), 375 (2021).
23. J. Wen-Ping, J. Zhen-Cun, "Research on early fire detection of Yolo V5 based on multiple transfer learning," *Fire Sci. Technol.* **40**(1), 109 (2021).
24. Y. Wang, X. Ma, Z. Chen *et al.*, Symmetric cross entropy for robust learning with noisy labels, *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp. 322–330, Seoul, Korea (2019).

25. A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," arXiv:1912.01703 (2019).

26. K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition" arXiv:1409.1556 (2014).

27. K. He, X. Zhang, S. Ren *et al.*, Deep residual learning for image recognition, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

28. G. Huang, Z. Liu, L. Van Der Maaten *et al.*, Densely connected convolutional networks, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, 2017.

29. M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, *Int. Conf. Machine Learning*, pp. 6105–6114 (PMLR, 2019).

30. M. Zhu, K. Han, Y. Tang *et al.*, Visual transformer pruning, arXiv:2104.08500 (2021).

31. E. Mousavi, R. Kafieh, H. Rabbani, "Classification of dry age-related macular degeneration and diabetic macular oedema from optical coherence tomography images using dictionary learning," *IET Image Process.* **14**(8), 1571–1579 (2020).

32. S. P. K. Karri, D. Chakraborty, J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomed. Opt. Exp.* **8**(2), 579–592 (2017).

33. A. Thomas, P. M. Harikrishnan, R. Ramachandran *et al.*, "A novel multiscale and multipath convolutional neural network based age-related macular degeneration detection using OCT images," *Comput. Meth. Prog. Biomed.* **209**, 106294 (2021).

34. Y. Luo, Q. Xu, R. Jin *et al.*, "Automatic detection of retinopathy with optical coherence tomography images via a semi-supervised deep learning method," *Biomed. Opt. Exp.* **12**(5), 2684–2702 (2021).

35. Z. Wang, W. Zhang, Y. Sun *et al.*, "Detection of diabetic macular edema in optical coherence tomography image using an improved level set algorithm," *BioMed Res. Int.* **2020**, 6974215 (2020).