Graefe's Archive for Clinical and Experimental Ophthalmology (2022) 260:849–856
https://doi.org/10.1007/s00417-021-05402-x

RETINAL DISORDERS

Automated detection of severe diabetic retinopathy using deep learning method

Xiao Zhang[1,2] · Fan li[3] · Donghong Li[4] · Qijie Wei[4] · Xiaoxu Han[1,2] · Bilei Zhang[1,2] · Huan Chen[1,2] · Yongpeng Zhang[5] · Bin Mo[5] · Bojie Hu[6] · Dayong Ding[4] · Xirong Li[3] · Weihong Yu[1,2] · Youxin Chen[1,2]

Received: 11 May 2021 / Revised: 3 August 2021 / Accepted: 26 August 2021 / Published online: 30 September 2021

Abstract

Purpose The purpose of this study is to develop and validate the intelligent diagnosis of severe DR with lesion recognition based on color fundus photography.

Methods The Kaggle public dataset for DR grading is used in the project, including 53,576 fundus photos in the test set, 28,101 in the training set, and 7,025 in the validation set. We randomly select 4,192 images for lesion annotation. Inception V3 structure is adopted as the classification algorithm. Both 299 × 299 pixel images and 896 × 896 pixel images are used as the input size. ROC curve, AUC, sensitivity, specificity, and their harmonic mean are used to evaluate the performance of the models.

Results The harmonic mean and AUC of the model of 896 × 896 input are higher than those of the 299 × 299 input model. The sensitivity, specificity, harmonic mean, and AUC of the method with 896 × 896 resolution images as input for severe DR are 0.925, 0.907, 0.916, and 0.968, respectively. The prediction error mainly occurs in moderate NPDR, and cases with more hard exudates and cotton wool spots are easily predicted as severe cases. Cases with preretinal hemorrhage and vitreous hemorrhage are easily identified as severe cases, and IRMA is the most difficult lesion to recognize.

Conclusions We have studied the intelligent diagnosis of severe DR based on color fundus photography. This artificial intelligence–based technology offers a possibility to increase the accessibility and efficiency of severe DR screening.

Keywords Severe diabetic retinopathy · Color fundus photography · Deep learning · Intraretinal microvascular abnormality

Xiao Zhang and Fan Li contributed equally to this article.

✉ Weihong Yu
yuweihong.pumch@vip.126.com

✉ Youxin Chen
chenyouxinpumch@163.com

[1] Department of Ophthalmology, Union Medical College Hospital, Chinese Academy of Medical Sciences, PekingBeijing, China

[2] Key Laboratory of Ocular Fundus Diseases, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[3] Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China

[4] Vistel AI Lab, Visionary Intelligence Ltd, Beijing, China

[5] Ophthalmology and Visual Science Key Lab, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, BeijingBeijing, China

[6] Department of Retina, Tianjin Medical University Eye Hospital, Tianjin, China

Key messages

- DR diagnosis and screening comprise the most widely used field of artificial intelligence in ophthalmology and is mainly used to screen DR patients who need to be referred based on color fundus photography, that is, to distinguish mild NPDR and moderate NPDR.
- In this study, on the basis of our experiences in referral DR screening, we have focused on the intelligent diagnosis of severe DR with lesion recognition based on color fundus photography,including DR grading analysis, lesion analysis, and heat map analysis.
- This artificial intelligence–based technology offers a possibility to increase the accessibility and efficiency of severe DR screening.

Introduction

Diabetic retinopathy (DR) is one of the most important microvascular complications of diabetes mellitus. Visual threatening diabetic retinopathy (VTDR), including severe nonproliferative diabetic retinopathy (NPDR), proliferative diabetic retinopathy (PDR), and diabetic macular edema (DME), is the main cause of visual impairment and needs to be actively treated in clinic.

According to public reports and the results of DR screening in recent years, the incidence of diabetes in China has reached as high as 11.6% [1, 2]. Patients with DR may have no obvious symptoms in the early stage. When they go to see the doctor after their vision has deteriorated, the best opportunity for treatment has been lost. Therefore, there is a need to carry out regular fundus screening in diabetic patients; find DR as early as possible, especially VTDR; and provide timely treatment to delay the progress of DR and improve or stabilize the vision of patients to reduce the heavy economic burden to patients' families and society in general. However, at present, there are no more than 6,000 specialized doctors in retina diseases in China [3], which is far from the number of doctors needed for large-scale screening. Some primary general ophthalmologists have caused delays in diagnosis and treatment of DR due to the lack of experience in the diagnosis of ocular fundus diseases. In addition, some patients lack awareness of diabetic retinal complications, which may eventually lead to irreversible vision loss.

Expanding the scope of screening and assisting in improving the diagnosis and treatment of primary-level ophthalmologists through medical artificial intelligence is an effective way to resolve this contradiction. At present, the diagnosis and screening of DR is the most widely used field of artificial intelligence in ophthalmology and is mainly used for screening diabetic patients who need to be referred to hospital based on color fundus photography. However, much less work has concentrated on the recognition of severe DR, which also needs to be focused on in screening, and there are still many limitations on it. In this study, on the basis of our experiences in referral DR screening, we will focus on the intelligent diagnosis of severe DR with lesion recognition based on color fundus photography.

Dataset

The Kaggle public dataset for DR grading (https://www.kaggle.com/c/diabetic-retinopathy-detection) is used in the project and includes a large number of high-resolution color fundus photos of diabetic patients. We will refer to the dataset as "Kaggle DR." The test set of Kaggle DR, which contains 53,576 fundus photos, is used as the test set in this experiment; the training set of Kaggle DR, which contains 35,126 fundus photos, is randomly divided into two parts in this study, including 28,101 photos as the training set and 7,025 photos as the validation set (Table 1). The original

Table 1 Dataset partition of our experiments

	Normal	Mild NPDR	Moderate NPDR	Severe NPDR	PDR	Sum
Training set	20,626	1,961	4,256	670	588	28,101
Validation set	5,184	482	1,036	203	120	7,025
Test set	39,533	3,762	7,861	1,214	1,206	53,576
Sum	65,343	6,205	13,153	2,087	1,914	88,702

Table 2 International DR grading standard

Grading	Fundoscopy after mydriasis
No DR	Normal
Mild NPDR	Only MA
Moderate NPDR	Not only MA, but the lesion is lighter than severe NPDR
Severe NPDR	Have any of the following and no signs of proliferative DR: - There were more than 20 intraretinal hemorrhage points in any quadrant of four quadrants; - More than two quadrants had clear venous beading; - Clear microvascular abnormalities in retina in more than one quadrant
PDR	One or more of the following changes occurred: Neovascularization, vitreous hemorrhage or preretinal hemorrhage

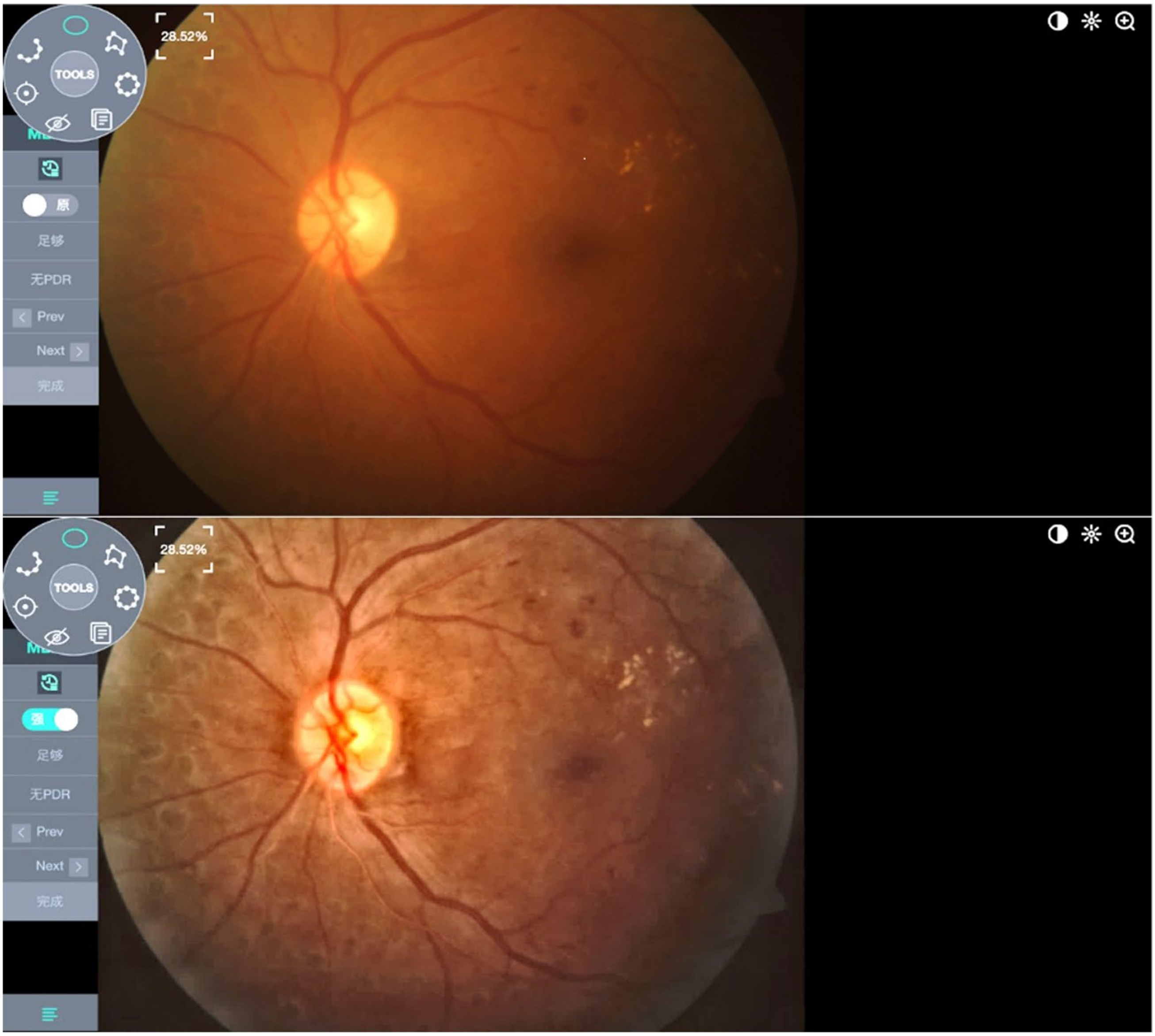

Fig. 1 Interface of our web-based annotation system with the original fundus image (top row) and an enhanced version of the original fundus image (bottom row)

training set of Kaggle DR is divided into two disjoint subsets, one for training and the other for validation, by random sampling without replacement as follows. Images are first randomly shuffled by the Fisher-Yates shuffle algorithm [4]. Then, the first 80% of the shuffled images (28,101 in total) are used for model training, while the remaining 20% (7,025 images) form the internal validation set. All images were classified by ophthalmologists according to the international grade of diabetic retinopathy (Table 2), including no DR, mild NPDR, moderate NPDR, and severe NPDR and PDR. Severe NPDR and PDR are called severe DR in this experiment and need timely diagnosis and treatment.

Kaggle DR provides DR grading labels without lesion annotations; therefore, we randomly select some images from the moderate NPDR, and severe NPDR and PDR cases of part of the test set by the Fisher-Yates shuffle algorithm for lesion annotation. The characteristic fundus lesions of DR are included in the annotation, including microaneurysm, intraretinal hemorrhage, hard exudate, cotton wool spot, intraretinal microvascular abnormality (IRMA), venous beading, neovascularization elsewhere (NVE), neovascularization of the optic disk (NVD), preretinal hemorrhage, vitreous hemorrhage, retinal proliferative membrane, and retinal detachment. To simplify the manual annotation, we develop a web-based annotation system and provide an adaptive enhanced version of the original image (Fig. 1). Each image is individually graded and annotated by at least three ophthalmologists. If the grades of the three ophthalmologists are inconsistent, an extra retinal specialist will arbitrate. The union of the three annotations is taken as the final lesion annotation. In addition, a trained quality control inspector will check the consistency of pathological content, which means that the lesions that conflict with the labeled DR grade are removed. As a result, 196 images are excluded. We ultimately obtain 4,192 images with lesion annotations.

Method

Inception V3 [5] structure is adopted as the classification algorithm. The grading of DR is based on the recognition of different lesions. Some lesions (i.e., microaneurysm) are tiny, and some lesions (i.e., IRMA) are difficult to recognize, which leads to the difficulty in DR grading. Therefore, considering that image resolution has a great impact on grading, images are downsized to 299×299 pixels and 896×896 pixels, respectively, for comparison. Among them, 299×299 pixels is the image size used in another study on the detection of VTDR [6]. Due to the imbalance of the proportions of severe DR (severe NPDR and PDR) and nonsevere cases (No DR, mild NPDR and moderate NPDR) in the training set (1,258/26,843), the strategy of weighted random sampling is used to make a balance between positive and negative cases in the training set. For model training, data augmentation in the form of random horizontal flip, rotation and random adjustment in brightness and contrast is performed in training images. The cross-entropy loss is used as the loss function. The initial learning rate is 0.001, and “reduce learning rate on plateau” is adopted to adjust the learning rate during training. SGD with a momentum of 0.9 is employed as the optimizer. Class activation mapping (CAM) technology is used to give us an idea about where our model pays attention for severe and nonsevere DR categories. In our network, CAM is obtained by a weighted sum of the feature maps of the last convolutional layer. By upsampling the class activation map to the size of the input image, we can identify the discriminative image regions most relevant to a specific category.

The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are used to evaluate the performance of the models. In addition, three criteria are also employed for evaluation, including sensitivity, specificity, and their harmonic mean, which are defined as:

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{TN + FP}$$

$$\text{harmonic mean} = 2 \times \text{sensitivity} \times \text{specificity}/(\text{sensitivity} + \text{specificity})$$

TP and TN represent the number of true positives and true negatives, respectively, while FP and FN represent the number of false positives and false negatives, respectively.

Our experiment is implemented with Python 2.7 and PyTorch 0.4.1 on a Linux server with 3 NVIDIA GeForce RTX 2080ti GPUs, configured with a 10-core CPU, 125 GPB memory, a 1 TB SSD hard disk, and an 11 TB mechanical hard disk. The study was approved by the Ethics Committee of the Peking Union Medical College Hospital. All investigations were carried out according to the Declaration of Helsinki.

Table 3 Results of different image resolutions using Inception V3

	Sensitivity	Specificity	Harmonic mean	AUC
InceptionV3_299	0.886	0.909	0.897	0.959
InceptionV3_896	0.925	0.907	0.916	0.968

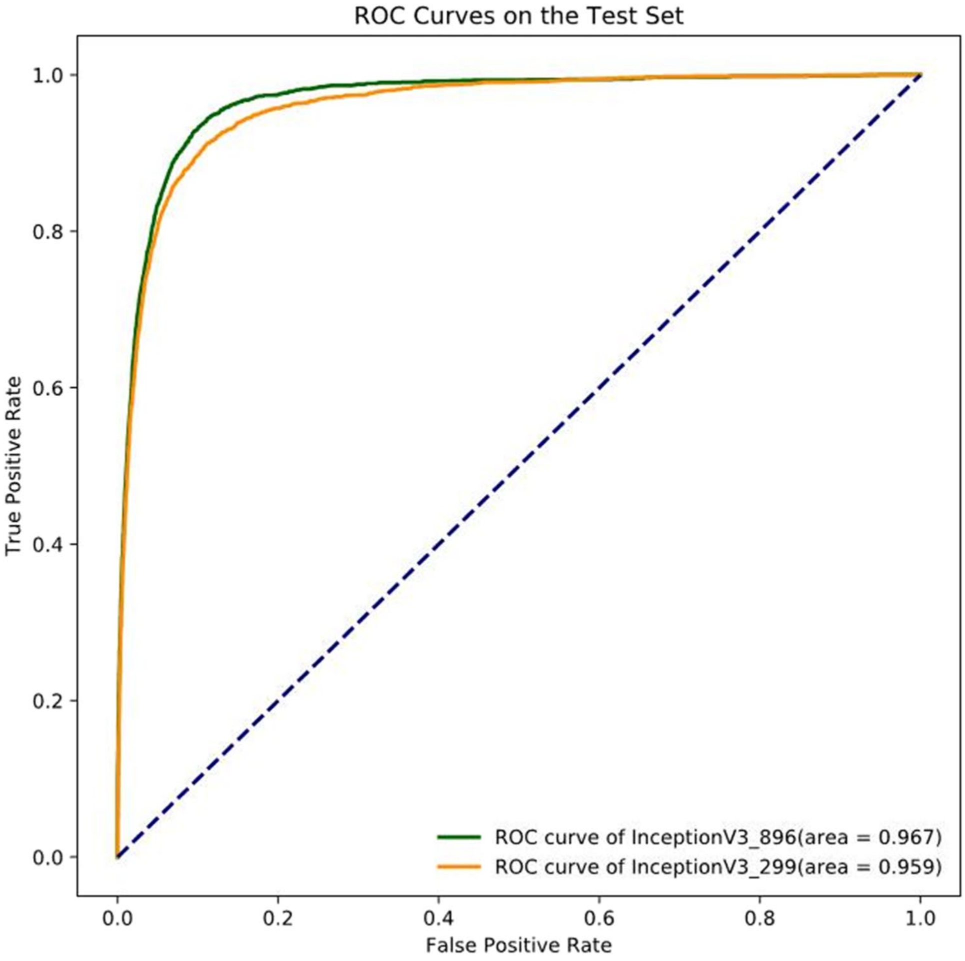

Fig. 2 ROC curves of models using Inception V3 architecture with 299 × 299 pixel images and 896 × 896 pixel images as input individually

Results

The experimental results are shown in Table 3, and the ROC curves are shown in Fig. 2. It can be seen from the results that the harmonic mean and AUC of model of 896 × 896 input are higher than those of 299 × 299 input. Further analysis for the results of the method with 896 × 896 resolution images as input is made in the following subsections.

DR grading analysis

Table 4 shows the number of severe DR and non-severe DR cases predicted by the model in each grade of DR. Moderate NPDR belongs to the category of nonsevere DR, but a certain proportion of cases (3,421/7,861 = 43.5%) is predicted to be severe. The ratio of mispredicted images in moderate NPDR to the number of all mispredicted images is 0.691 (3,421/(1,103 + 248 + 3,421 + 114 + 67)), which to some extent means the prediction error mainly occurs in moderate NPDR.

Lesion analysis

In the 4,192 images with lesion annotations, the DR grade distributions of predicted severe DR and predicted nonsevere DR are shown in Table 5. According to Table 5, 68 cases of severe NPDR and 29 cases of PDR are predicted as nonsevere DR. Among the 97 cases of severe DR that are not reported,

Table 4 The actual DR grades of the predicted severe DR set and the predicted nonsevere DR set

	Normal	Mild NPDR	Moderate NPDR	Severe NPDR	PDR	Sum
Predicted severe	1,103	248	3,421	1,100	1,139	7,011
Predicted nonsevere	38,430	3,514	4,440	114	67	46,565
Sum	39,533	3,762	7,861	1,214	1,206	53,576

Table 5 DR grade distributions of the predicted severe DR set and the predicted nonsevere DR set employed for lesion analysis

	Moderate NPDR	Severe NPDR	PDR	Sum
Predicted severe	1,096	795	718	2,609
Predicted nonsevere	1,486	68	29	1,583
Sum	2,582	863	747	4,192

11 cases have IRMA (11/97, 11.3%), 5 cases have retinal neovascularization (5/97, 5.2%), and 7 cases have proliferative membrane (7/97, 7.2%).

The predicted severe DR and nonsevere DR and their corresponding labeled lesions are listed in Table 6. Regarding the lesion analysis of severe DR, we have the following conclusions from the experiment: (1) In moderate NPDR, cases with more hard exudates and cotton wool spots are easily predicted as severe cases; (2) cases with preretinal hemorrhage and vitreous hemorrhage are easily identified as severe cases; and (3) IRMA is the most difficult lesion to recognize in DR cases.

Heat map analysis

We apply CAM to visualize which areas of the input image play an important role in models' decision-making. The visualization examples using CAM are shown as follows (Fig. 3).

Discussion

DR diagnosis and screening comprise the most widely used field of artificial intelligence in ophthalmology and are mainly used to screen DR patients who need to be referred based on color fundus photography, that is, to distinguish mild NPDR and moderate NPDR; for more severe cases, it is recommended to see a doctor in the hospital. There have been many studies in this field. In 2015, in the DR detection competition organized by Kaggle [7], more than 35,000 color fundus images were used to train algorithms for predicting disease severity. The results of four teams out of 661 were higher than that of humans. In 2016, Gulshan et al. developed and validated a deep learning algorithm for the detection of diabetic retinopathy in retinal fundus photographs and comprehensively analyzed DR detection [8]. Its performance was equivalent to that of a group of seven certified ophthalmologists, with sensitivity and specificity of more than 90%. Similar studies on intelligent diagnosis of DR cases that need to be referred have also been found in multiple literatures [7–11], with sensitivity of more than 90% and specificity of more than 80% [12]. In 2018, the U.S. Food and Drug Administration (FDA) approved the artificial intelligence product IDX-DR for DR detection, which became the world's first artificial intelligence medical equipment for disease diagnosis.

At present, there are few studies on artificial intelligence–assisted VTDR diagnosis. To our knowledge, studies on the intelligent diagnosis of VTDR that can be retrieved in the literature mainly include the following works: (1) in 2016, Abràmoff et al. [7] used IDx-DR × 2.1 system to recognize VTDR, and the sensitivity and specificity were 100.0% and 90.8%, respectively; (2) in 2017, Ting et al. [10] used a deep learning system to identify DR, and the sensitivity and specificity of VTDR recognition were 100% and 91.1%, respectively; (3) in 2018, Li et al. [6] detected VTDR based on convolutional neural network. The sensitivity and specificity reached 97.0% and 91.4% on the internal validation dataset and 92.5% and 98.5% on the external independent multiethnic dataset. In our study, we use the network

Table 6 The labeled lesions by retinal specialists in 4,192 images of the test set

	Moderate NPDR		Severe NPDR		PDR	
	Pred1	Pred0	Pred1	Pred0	Pred1#	Pred0#
MA	1,039	1,283	747	62	548	20
Intraretinal hemorrhage	1,064	1,211	790	63	669	17
Preretinal hemorrhage	0	0	13	0	72	0
Vitreous hemorrhage	0	0	3	0	31	0
Hard exudate	706	535	606	27	445	6
Cotton wool spot	382	284	434	11	180	4
Venous beading	0	0	91	2	72	1
IRMA	0	0	197	9	162	2
NVD	0	0	15	0	165	0
NVE	0	0	53	4	202	1
Proliferative membrane	0	0	4	0	160	3
Drusen	121	194	47	6	25	6

Pred1, predicted as severe DR; *Pred0*, predicted as nonsevere DR

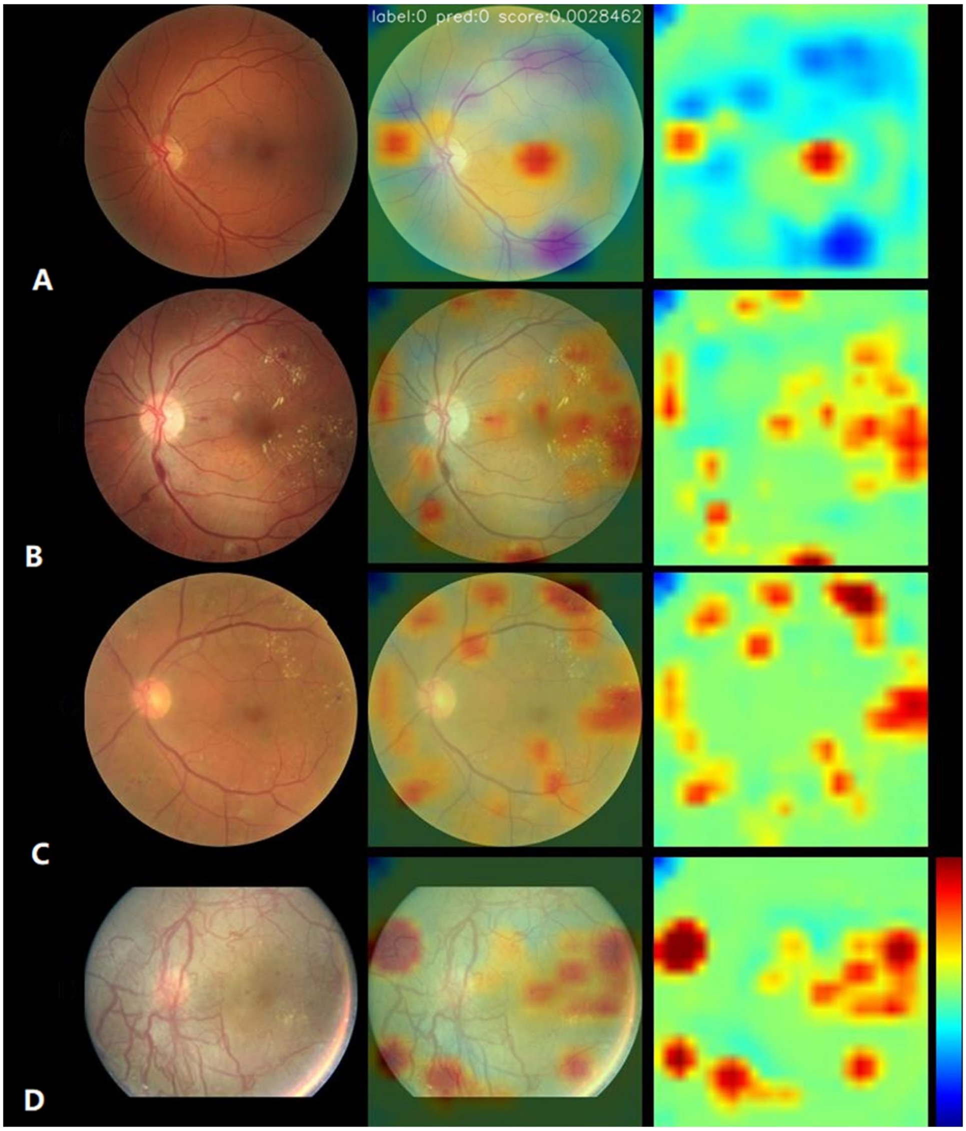

Fig. 3 CAM-based visualization examples. Left column: the original fundus images; middle column: the fusions of heat maps and original fundus images; right column: the heat maps. (A) A case of moderate NPDR is correctly predicted as nonsevere DR. (B) A case of severe NPDR is correctly predicted as severe DR. (C) A case of moderate NPDR is predicted as severe DR because hard exudates and cotton wool spots were activated. (D) A case of PDR is correctly predicted as severe DR, but only a few of the retinal neovascularizations are activated on the heat map, and some large new blood vessels are not activated

structure of Inception V3 to classify DR into severe and nonsevere cases, and the sensitivity and specificity reach 92.52% and 90.67%. Both 299 × 299 pixel images and 896 × 896 pixel images are used in this experiment as the input size, and the former were used in the research of Li et al. [6]. We find that the value of both harmonic mean of sensitivity and specificity and AUC are higher in 896 × 896 pixel images than those in 299 × 299 pixel images. Compared to the former studies, our study has the following advantages. First, the strategy of weighted random sampling is used to make a balance between positive and negative cases in the training set. Second, lesion analysis is included in our study, which is the exploration on the interpretability of models. Third, different pixel-resolution images are used in this experiment as input sizes to identify an optimal one.

Although the sensitivity and specificity of the published studies and our study are all above 90%, there are some limitations.

First, the positive and negative samples of the public datasets are seriously imbalanced. At present, public datasets are used for the training and testing of models in many studies. The number of DR patients is limited in these public datasets, which is far less than the number of normal people. If only DR samples are tested, the sensitivity and specificity will decrease to some extent or even a large extent.

Second, the positive predictive value of severe DR cases is relatively low and there is a lack of interpretability. In the study of Abràmoff [7], the positive predictive value of VTDR was only 56.4%. In the study of Li [6], the 77.3% false negative rate was because IRMA in fundus images was not recognized. In our study, 43.5% (3,421/7,861) of moderate NPDR patients are recognized as severe cases. Because of the "black box" of the machine learning model, it is difficult to explain why these misdiagnoses are caused. This issue needs to be addressed with the professional knowledge of ophthalmologists and the development of interpretability of the machine learning model.

Third, DME is an important part of VTDR, which is often difficult to identify manually from color fundus photos. At present, it mainly depends on OCT for diagnosis and follow-up, but due to the limitations of conditions, DR screening does not involve an OCT device in many cases. Therefore, if we can use artificial intelligence to diagnose DME accurately according to the color fundus photos and get rid of the dependence on OCT, we can easily solve this contradiction. The problem was not solved in the abovementioned studies concerning VTDR screening. In our study, only the recognition of severe NPDR and PDR is involved, and the recognition of DME is not included. In 2019, Arcadu et al. [13] retrospectively analyzed 17,997 color fundus photos and their corresponding OCT and made a preliminary exploration of predicting macular thickness from color fundus photos using deep learning. The sensitivity was 80.0 ~ 90.0%, and the specificity was 85.0 ~ 94.0%. However, there are no other relevant reports at present. The recognition of DME in color fundus photography is the next step we need to study.

In conclusion, we have studied the intelligent diagnosis of severe DR based on color fundus photography, including DR grading analysis, lesion analysis, and heat map analysis. Prediction errors mainly occur in moderate NPDR, and images of moderate NPDR are prone to be misreported as severe cases. Preretinal hemorrhage and vitreous hemorrhage are easily identified as lesions of severe cases, while IRMA is the most difficult lesion to recognize. Moderate NPDR with relatively more hard exudates and cotton wool spots is easily predicted as severe cases, while severe NPDR and PDR with relatively less intraretinal hemorrhage are easily predicted as nonsevere cases.

Funding 1. The Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2018PT32029)

2. CAMS Initiative for Innovative Medicine(CAMS-I2M, 2018-I2M-AI-001)

3. Pharmaceutical collaborative innovation research project of Beijing Science and Technology Commission (Z191100007719002)

4. National Key Research and Development Project(SQ2018YFC200148)

5. Beijing Natural Science Foundation Haidian original innovation joint fund (19L2062)

6. National Natural Science Foundation of China (NSFC No. 61672523)

7. Beijing Natural Science Foundation (BJNSF No. 4202033)

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Procedures performed in this study were in accordance with the ethical standards of the Peking Union Medical College Hospital and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

References

1. Yang W, Lu J, Weng J et al (2010) Prevalence of diabetes among men and women in China. N Engl J Med 362(12):1090–1101
2. Chan JC, Zhang Y, Ning G (2014) Diabetes in China: a societal solution for a personal challenge. Lancet Diabetes Endocrinol 2(1):969–979
3. Bellemo V, Lim G, Rim TH et al (2019) Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. Curr Diab Rep 19(9):72
4. Knuth DE (1997) The Art of Computer Programming, Volume 2: Seminumerical Algorithms. Addison Wesley Longman, 145–146
5. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.
6. Li Z, Keel S, Liu C et al (2018) An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. Diabetes Care 41(12):2509–2516
7. Abràmoff MD, Lou Y, Erginay A et al (2016) Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci 57(13):5200–5206
8. Gulshan V, Peng L, Coram M et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22):2402–2410
9. Gargeya R, Leng T (2017) Automated identification of diabetic retinopathy using deep learning. Ophthalmology 124(7):962–969
10. Ting DSW, Cheung CY, Lim G et al (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 318(22):2211–2223
11. Abràmoff MD, Lavin PT, Birch M et al (2018) Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 1:39
12. Ting DSW, Pasquale LR, Peng L et al (2019) Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 103(2):167–175
13. Arcadu F, Benmansour F, Maunz A et al (2019) Deep learning predicts OCT measures of diabetic macular thickening from color fundus photographs. Invest Ophthalmol Vis Sci 60(4):852–857