



## RESEARCH ARTICLE

# Deep learning-based detection and stage grading for optimising diagnosis of diabetic retinopathy

Yuelin Wang<sup>1,2</sup>  | Miao Yu<sup>3</sup>  | Bojie Hu<sup>4</sup> | Xuemin Jin<sup>5</sup> | Yibin Li<sup>6,7</sup> | Xiao Zhang<sup>1,2</sup> | Yongpeng Zhang<sup>7</sup> | Di Gong<sup>8</sup> | Chan Wu<sup>1,2</sup> | Bilei Zhang<sup>1,2</sup> | Jingyuan Yang<sup>1,2</sup> | Bing Li<sup>1,2</sup> | Mingzhen Yuan<sup>1,2</sup> | Bin Mo<sup>7</sup> | Qijie Wei<sup>9</sup> | Jianchun Zhao<sup>9</sup> | Dayong Ding<sup>9</sup> | Jingyun Yang<sup>10</sup> | Xirong Li<sup>11</sup> | Weihong Yu<sup>1,2</sup> | Youxin Chen<sup>1,2</sup>

<sup>1</sup>Department of Ophthalmology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China

<sup>2</sup>Key Lab of Ocular Fundus Disease, Chinese Academy of Medical Sciences, Beijing, China

<sup>3</sup>Department of Endocrinology, Key Laboratory of Endocrinology, National Health Commission, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

<sup>4</sup>Department of Ophthalmology, Tianjin Medical University Eye Hospital, Tianjin, China

<sup>5</sup>Department of Ophthalmology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China

<sup>6</sup>Department of Ophthalmology, Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing, China

<sup>7</sup>Beijing Key Laboratory of Ophthalmology and Visual Science, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China

<sup>8</sup>Department of Ophthalmology, China-Japan Friendship Hospital, Beijing, China

<sup>9</sup>Vistel AI Lab, Visionary Intelligence Ltd., Beijing, China

<sup>10</sup>Department of Neurological Sciences, Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, USA

<sup>11</sup>Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China

## Correspondence

Youxin Chen and Weihong Yu, Key Lab of Ocular Fundus Disease, Chinese Academy of Medical Sciences; Department of Ophthalmology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, No. 1 Shuaifuyan Road, Dongcheng District, Beijing 100730, China.  
Email: [chenyx@pumch.cn](mailto:chenyx@pumch.cn) and [yuweihong.pumch@vip.126.com](mailto:yuweihong.pumch@vip.126.com)

Youxin Chen and Weihong Yu should be considered joint senior authors.

## Funding information

the non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences, Grant/Award Number: 2018PT32029; National Key Research and Development Project, Grant/Award Number: SQ2018YFC200148; Beijing Natural Science Foundation, Grant/Award Number: 4202033; the Beijing Natural Science Foundation Haidian Original Innovation Joint Fund, Grant/Award Number: 19L2062; Pharmaceutical Collaborative Innovation Research Project of Beijing Science and Technology Commission,

## Abstract

**Aims:** To establish an automated method for identifying referable diabetic retinopathy (DR), defined as moderate nonproliferative DR and above, using deep learning-based lesion detection and stage grading.

**Materials and Methods:** A set of 12,252 eligible fundus images of diabetic patients were manually annotated by 45 licenced ophthalmologists and were randomly split into training, validation, and internal test sets (ratio of 7:1:2). Another set of 565 eligible consecutive clinical fundus images was established as an external test set. For automated referable DR identification, four deep learning models were programmed based on whether two factors were included: DR-related lesions and DR stages. Sensitivity, specificity and the area under the receiver operating characteristic curve (AUC) were reported for referable DR identification, while precision and recall were reported for lesion detection.

**Results:** Adding lesion information to the five-stage grading model improved the AUC (0.943 vs. 0.938), sensitivity (90.6% vs. 90.5%) and specificity (80.7% vs. 78.5%) of the model for identifying referable DR in the internal test set. Adding stage information to the lesion-based model increased the AUC (0.943 vs. 0.936) and

Grant/Award Number: Z191100007719002;  
CAMS Innovation Fund for Medical Sciences  
(CIFMS): Grant/Award Number: CAMS-I2M;  
2018-I2M-AI-001

sensitivity (90.6% vs. 76.7%) of the model for identifying referable DR in the internal test set. Similar trends were also seen in the external test set. DR lesion types with high precision results were preretinal haemorrhage, hard exudate, vitreous haemorrhage, neovascularisation, cotton wool spots and fibrous proliferation.

**Conclusions:** The herein described automated model employed DR lesions and stage information to identify referable DR and displayed better diagnostic value than models built without this information.

#### KEYWORDS

deep learning, diabetic retinopathy, lesion detection, screening, stage grading

## 1 | INTRODUCTION

Diabetic retinopathy (DR), a major complication of diabetes, has become the leading cause of vision loss worldwide.<sup>1-3</sup> The situation is similarly severe in China, which has more than 100 million diabetic patients with a DR prevalence rate of approximately 30%.<sup>4</sup> Early screening for DR based on colour fundus images is well recognised as an effective measure to prevent blindness.<sup>5,6</sup> It is recommended that patients with no retinopathy or mild nonproliferative DR (NPDR) undergo annual screening, while those with moderate to severe NPDR and proliferative DR (PDR) need to be referred to retina specialists for further evaluation, closer follow-up or treatment.<sup>7</sup> A delayed diagnosis of DR could lead to an elevation of the rate of referable DR, with the risk of proliferative retinopathy four times higher in those for whom screening is delayed 3 years or more.<sup>8,9</sup> Given the high demand for large-scale DR screening, the lack of ophthalmologists has become a bottleneck for timely DR screening, especially in developing countries.<sup>10-12</sup> Therefore, a high-priority task is to identify, among patients with diabetes, those with retinopathy that require a referral to an ophthalmologist.<sup>13</sup>

In the last few years, the automated identification of DR using convolutional neural network (CNN)-based deep learning techniques has gained increasing attention.<sup>14-17</sup> A CNN model imitates to some extent the hierarchical information processing mechanism of the human brain. Given an input image, the CNN model starts from processing raw pixels, learns to perform information abstraction layer by layer and eventually produces high-level semantics such as objects and scenes presented in the image. Existing CNN-based approaches to DR identification follow the above procedure and make a binary prediction<sup>18,19</sup>; a colour fundus image is classified as either referable or nonreferable DR.<sup>7</sup>

In a typical clinical scenario, the diagnosis of referable DR by ophthalmologists is based on the presence of specific lesions which also translate to a specific DR stage. For the diagnosis of ophthalmologists, each fundus photograph is primarily focused on DR-related lesions.<sup>20</sup> Based on these lesions, images are classified into DR stages according to the International Clinical DR Classification System. Those with moderate or severe NPDR are referred to retina specialists for further evaluation. But for the diagnosis of machines,

the state-of-the-art methods, such as Gulshan et al.,<sup>18</sup> Gargeya et al.<sup>21</sup> and Ting et al.<sup>15</sup> are able to identify referable DR identification without taking lesion information into account.<sup>22</sup> As a consequence, their predictions lack clinical interpretation, albeit their high accuracy. As opposed to the previous models, in this study we developed a CNN-based model that explicitly detects lesions from a given fundus image and subsequently analyses the detected lesions for five-stage DR grading. Lesion detection not only improves referable DR identification, but also provides a means for decision interpretation.

## 2 | MATERIALS AND METHODS

### 2.1 | Data source

This study adhered to the tenets of the Declaration of Helsinki. The colour fundus images used in this study were retrospectively obtained and deidentified. Ethics review and Institutional Review Board exemption were obtained. Our study dataset consisted of 22,948 fundus images of patients with diabetes mellitus. Among these images, 10,678 highly suspected DR images, which may contain DR lesions, were selected from the publicly available Kaggle DR dataset,<sup>23</sup> while the rest of the images were randomly selected from Peking Union Medical College Hospital Database and the Tianjin Medical University Eye Hospital Diabetic Retinopathy Screening Program Database, which contain highly suspect DR images. In addition, we also collected 750 consecutive images from diabetes mellitus patients who underwent fundus photography in the Peking Union Medical College Hospital Examination Room between January 1, 2016, and January 1, 2018, which differed from the above database. Each eye was captured as an image of a macular-centred, 35–45° field of view by various camera models, including Canon CR1/DGi/CR2, Topcon NW and Optovue iCam cameras. The presence of laser scars in a fundus image suggested a history of laser photocoagulation therapy. As such, the International Clinical DR Classification System was no longer applied. Therefore, manual inspection was performed to exclude images with laser scars from all datasets.

## 2.2 | Data annotation

Our annotation team consisted of 45 licenced ophthalmologists in China (including five authors on the paper). All of them were trained by a course designed by a panel of retinal specialists and achieved a consistency of 85% in a DR grading test of consisting of 50 standard DR images. Each image from the datasets was graded and annotated independently by at least three ophthalmologists. In addition, consistency checks on the grading and lesion labels were performed by a trained quality control retinal specialist. Questionable images were submitted to a review panel of at least five ophthalmologists for further discussion; majority voting was used to reach a final decision. In previous studies, referable diabetic macular oedema was defined as any hard exudate (HE) within one-disc diameter (1.5 mm) of the centre of the macula,<sup>18</sup> in this study, we did not consider the presence of diabetic macular oedema since a single two-dimensional fundus image alone is inadequate to determine the existence of macular oedema.<sup>24</sup>

For the ease of manual grading and annotation, we developed a web-based annotation system and provided adaptively enhanced (high contrast) versions of the original images for reference (Figure 1A,B). Each image was graded as one of the five DR levels, that is, no DR, mild NPDR, moderate NPDR, severe NPDR and PDR, as specified in the 2003 International Clinical DR Classification System.<sup>25</sup> Image qualities were also classified as 'excellent' if all of the DR lesions were gradable (see Figure S1A); 'good' if there was a problem with 1–2 image quality factors (focus, illumination, artifacts or not including the entire optic nerve head and macula) (see Figure S1B); 'adequate' if there were problems with 3–4 of the image quality factors but all the DR lesions were gradable (see Figure S1C); 'insufficient for full interpretation' if one or more DR lesions could not be graded (see Figure S1D); and as 'insufficient for any interpretation' (see Figure S1E).

Referable DR-related lesions with varying colours and arbitrarily shaped lesion boundaries, including preretinal haemorrhage, HE, vitreous haemorrhage (VH), neovascularisation (NV), cotton wool spots (CWS), fibrous proliferation, intraretinal haemorrhage, intraretinal microvascular abnormality (IRMA), venous beading and microaneurysm (MA) were annotated by ophthalmologists (Figure 1C).

## 2.3 | Lesion-based DR classification models

Our lesion detector was based on a CNN, which was named the Lesion-Net. The proposed Lesion-Net has two branches (Figure 2A). The upper branch adopted the convolutional layers of the well-recognised Inception-v3 model,<sup>18,26</sup> which was originally designed for referable/nonreferable DR classification.<sup>15,18,27</sup> In the lower branch, we used a fully convolutional network (FCN-32s) to give pixelwise lesion segmentation predictions. The threshold is set to 0.5 for all lesions as common practice. Learning the threshold from the

validation set may help in improving the performance but it is not what this paper focuses on. Notice that for normal images, the predicted lesion scores would not be all zeros, more details are shown in Supporting Information Material.

A convolutional block was used to convert the lesion segmentation predictions into weights, which were then added via elementwise multiplication to the feature maps obtained in the upper branch. With this architecture, the model takes advantage of the predicted lesions. For DR grading, the shape of output scores would be  $5 \times 1$ , which is the probability of DR0–DR4 (five-stage model, where DR0 corresponds to 'no DR', DR1 corresponds to 'mild NPDR', DR2 corresponds to 'moderate NPDR', DR3 corresponds to 'severe NPDR', and DR4 corresponds to 'PDR'). For referable/nonreferable DR classification, the shape would be  $2 \times 1$ , which was the probability of referable DR and nonreferable DR (a two-class model).

To make small lesions more visible, we used high-resolution images as input, the size of which was  $3 \times 896 \times 896$  pixels (with RGB channels). Although the feature size would be down sampled to  $2048 \times 28 \times 28$  during convolutions in which representations of small lesions may be invisible, but the lesion features still exist. When resizing the lesion features, the pixels of the feature map are no longer RGB pixels, but high dimensional vectors. These vectors reserve the lesion features and can restore to the segmentation images with the same resolution as the original images by up sampling.

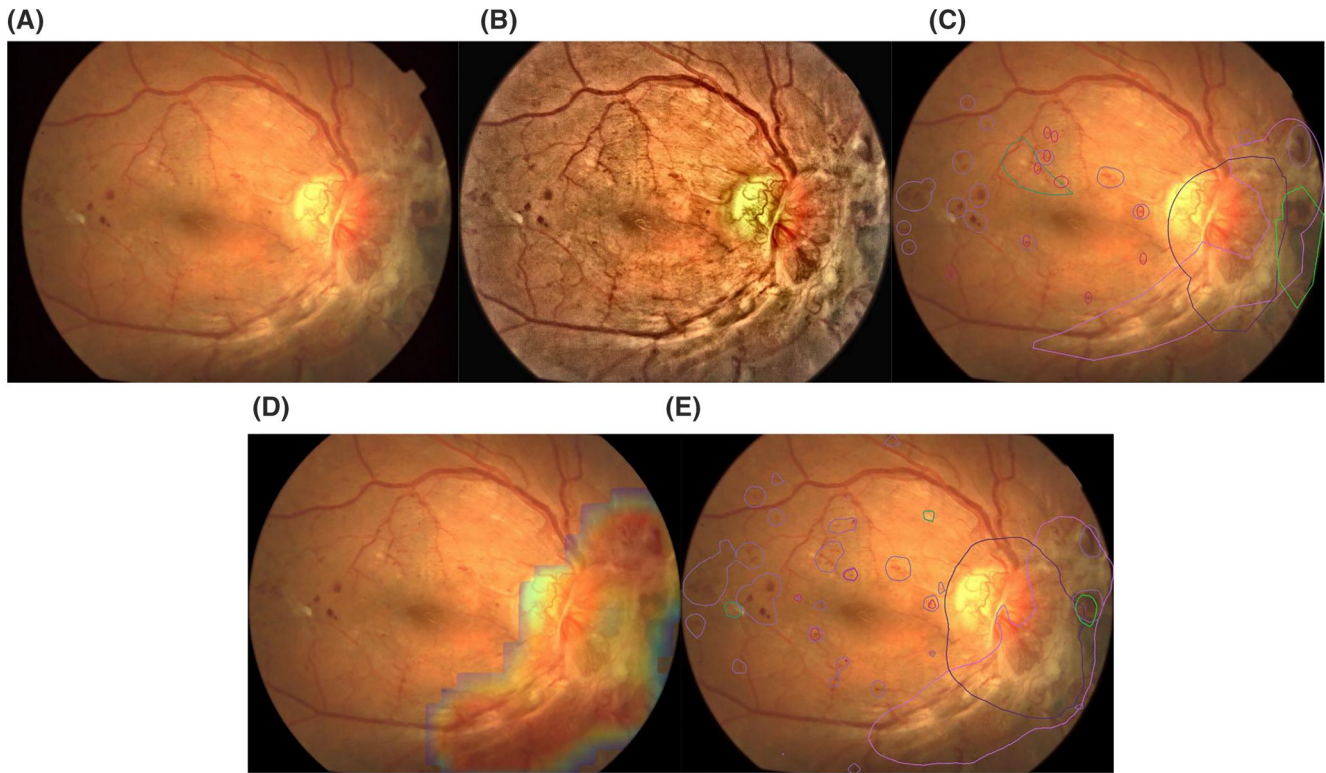
For network training, we adopted stochastic gradient descent with a weight decay of 0.0001 and a momentum of 0.95 for optimisation. The batch size was set to 4. Validation occurred every 1000 batches. If the validation performance did not improve after four consecutive validations, the learning rate was divided by 10. If the validation performance did not improve after 10 consecutive validations, training was stopped.

## 2.4 | Controlled models

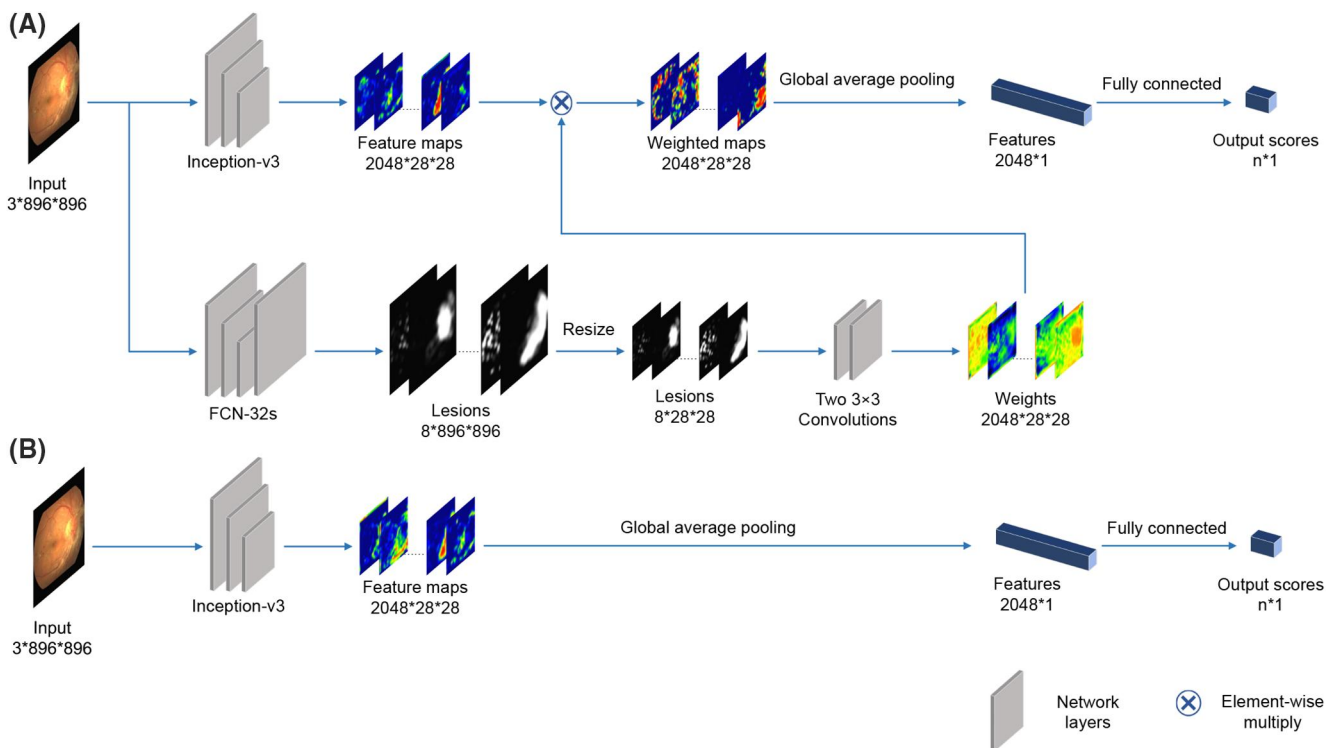
We also established controlled models that did not consider any lesion information (Figure 2B). We trained a five-stage model for DR stages (DR0–DR4) and a two-class model for referable/nonreferable DR.

## 2.5 | Statistical analysis

For referable/nonreferable DR classification, we reported the area under the receiver operating characteristic (ROC) curve and area under the curve (AUC), sensitivity, specificity, referral kappa and F1-score to evaluate the model performance in the test set.<sup>27–29</sup> The sensitivity was calculated as the number of correctly predicted positive examples divided by the total number positive examples. The specificity was calculated as the number of correctly predicted negative examples divided by the total number negative examples.



**FIGURE 1** Examples of colour fundus images: (A) original image; (B) contrast-enhanced image; (C) image with lesion annotation by ophthalmologists (schematic diagram); (D) pattern recognition heat map of the Inception-v3 model; (E) image with lesion recognition by the Lesion-Net model



**FIGURE 2** Two network architectures for referable/nonreferable DR classification. (A) Inception-v3 enhanced by the proposed Lesion-Net, with output lesion annotations. (B) Inception-v3. DR, diabetic retinopathy

**TABLE 1** Statistics for the manual DR grading and lesion annotation analysis of the fundus images

	Training and validation set		Internal test set		External test set		Total	
No DR (%)	1018 (10.39)		268 (10.94)		61 (10.80)		1347	
Mild NPDR (%)	1384 (14.12)		364 (14.86)		36 (6.47)		1784	
Moderate NPDR (%)	5497 (56.08)		1344 (54.86)		323 (57.16)		7164	
Severe NPDR (%)	896 (9.14)		234 (9.55)		97 (17.17)		1227	
PDR (%)	1007 (10.27)		240 (9.80)		48 (8.50)		1295	
Total	9802		2450		565		12,817	
Lesions <sup>a</sup>	Connected component	Images	Connected component	Images	Connected component	Images	Connected component	Images
Microaneurysm	84,348	7869	20,809	1823	6353	493	111,510	10,185
Intraretinal haemorrhage	110,100	8281	26,169	1697	5475	448	141,744	10,426
Preretinal haemorrhage	1009	501	280	80	52	24	1341	605
Vitreous haemorrhage	360	290	103	69	10	7	473	366
Hard exudate	28,612	5471	7116	1358	2235	405	37,963	7234
Cotton wool spot	8066	3,025	2,196	782	844	186	11,106	3993
Neovascularization	1829	875	351	169	66	37	2246	1081
Fibrous proliferation	622	488	160	123	23	18	805	629

Abbreviations: DR, diabetic retinopathy; NPDR, nonproliferative diabetic retinopathy; PDR, proliferative diabetic retinopathy.

<sup>a</sup>Connected component, union subgraphs composed of pixels with similar values and adjacent locations annotated by ophthalmologists.

The referral kappa was used to evaluate the referral consistency between two annotators. The F1-score was calculated as the harmonic mean of the sensitivity and specificity. We regarded DR0 and DR1 as nonreferable DR, and DR2, DR3 and DR4 as referable DR. Therefore, we also reported the sensitivity, specificity and referral kappa for the DR stages converted to referable/nonreferable DR prediction from the five-stage models.

To evaluate the effectiveness of our lesion detection models, we calculated the precision and recall based on the connected components of the predictions and ground truth.<sup>30,31</sup> For each predicted connected component of the lesion, if over 50% of the pixels were covered by the ground truth, the predicted lesion was considered correct. Precision was calculated as the number of correctly predicted connected components divided by the total number of predicted connected components. Similarly, for each connected component of the lesion in the ground truth, if over 50% of the pixels were also covered by the prediction, the ground-truth lesion was considered to be successfully detected. Recall was calculated as the number of successfully detected connected components divided by the total number of connected components in the ground truth. Only the lesions with acceptable precision (>0.5) and recall (>0.1) values were included in the Lesion-Net model. After a preliminary experiment, we found that it was difficult for the model to detect venous beading and IRMA in the single-

colour fundus images. The F1-score was 0 for venous beading and 0.06 for IRMA. Therefore, eight types of lesions (preretinal haemorrhage, HE, VH, NV, CWS, fibrous proliferation, intraretinal haemorrhage and MA) were included in the lesion detection models. For each input image with size  $3 \times 896 \times 896$ , FCN-32s outputs eight  $896 \times 896$  lesion predictions, represent eight kind of lesions correspondingly. A single pixel is allowed to be predicted as multiple lesions for lesions may have overlaps in practice.

### 3 | RESULTS

#### 3.1 | Data analytics of DR grading and lesion annotation

After excluding low-quality (e.g., Figure S1D,E), inconsistent annotation images (such as Figure S1F, which can be annotated as DR2 or DR3), and fundus photographs with laser spots (e.g., Figure S1G), 12,252 images were eligible for inclusion in the internal dataset (see Figure S2). After annotation, 1018 images (10.39%) in the training and validation sets were graded as no DR, 1384 (14.12%) as mild NPDR, 5497 images (56.08%) were graded as moderate NPDR, 896 images (9.14%) were graded as severe



TABLE 2 Performances of the different referable/nonreferable DR classification models

Type of model	AUC		Sensitivity		Specificity		Kappa		F1-score	
	Internal	External	Internal	External	Internal	External	Internal	External	Internal	External
Lesion-Net + five stages	0.943	0.982	90.60%	95.70%	80.70%	93.80%	0.696	0.815	0.854	0.948
Lesion-Net + referable	0.936	0.977	76.70%	95.10%	93.70%	90.70%	0.588	0.812	0.844	0.929
Fundus image + five stages	0.938	0.982	90.50%	93.80%	78.50%	92.80%	0.677	0.807	0.841	0.933
Fundus image + referable	0.928	0.964	81.70%	88.30%	88.80%	93.80%	0.621	0.684	0.851	0.910

Abbreviations: AUC, area under the curve; DR, diabetic retinopathy.

NPDR and 1007 images (10.27%) were graded as PDR (see Table 1). Therefore, 7400 images (75.49%) were graded as referable DR and 2402 images (24.51%) were graded nonreferable DR. Among these images, eight unique types of lesions and 234,946 annotations were identified. Table 1 shows the distribution of these lesions. The most common type of lesion was intraretinal haemorrhage, followed by MA and HE, with 110,100, 84,348 and 28,612 annotations, respectively.

For the internal test set, the distribution of DR gradings was roughly the same as in the training and validation sets (see Table 1). In the external test set, of which 565 images were eligible for inclusion after annotation, 61 images (10.80%) were graded as no DR, 36 images (6.47%) were graded as mild NPDR, 323 images (57.16%) were graded as moderate NPDR, 97 images (17.17%) were graded as severe NPDR and 48 images (8.50%) were graded as PDR. Therefore, 468 images (82.83%) were graded as referable DR and 97 images (17.17%) were graded as nonreferable DR.

### 3.2 | Performances of DR classification models

The performances of the Inception-v3 and our proposed Lesion-Net are shown in Table 2. The ROC curves of these models are shown in Figure 3.

Concerning whether adding the lesion information to the models improved the diagnostic value, for the five-stage models, Lesion-Net could increase the AUC (0.943 vs. 0.938), sensitivity (90.6% vs. 90.5%) and specificity (80.7% vs. 78.5%), referral kappa (0.696 vs. 0.677) and F1-score (0.854 vs. 0.841) in the internal test set. In the external test set, we found the same trends in the diagnostic value of the two models, excluding the AUC (0.982 vs. 0.982). For the binary models for referable/nonreferable DR, the Lesion-Net increased the AUC in both test sets (0.936 vs. 0.928 for the internal test set and 0.977 vs. 0.964 for the external test set) compared with the models simply based on the fundus images.

To determine whether the stage information affected the diagnostic value of the models, in the Lesion-Net models, we converted the DR stages to referable/nonreferable DR, and increases in the AUC (0.943 vs. 0.936, 0.982 vs. 0.977), sensitivity (90.6% vs. 76.7%, 95.7% vs. 95.1%), referral kappa (0.696 vs. 0.677, 0.815 vs. 0.812) and F1-score (0.854 vs. 0.844, 0.948 vs. 0.929) were achieved in both the internal and external test sets of the five-stage models, compared

with those in models for referable/nonreferable DR, respectively. Similarly, the AUC, sensitivity and referral kappa were also increased in the five-stage models in the Inception-v3 models for the fundus images.

### 3.3 | Precision and recall values for lesions in the lesion detection model

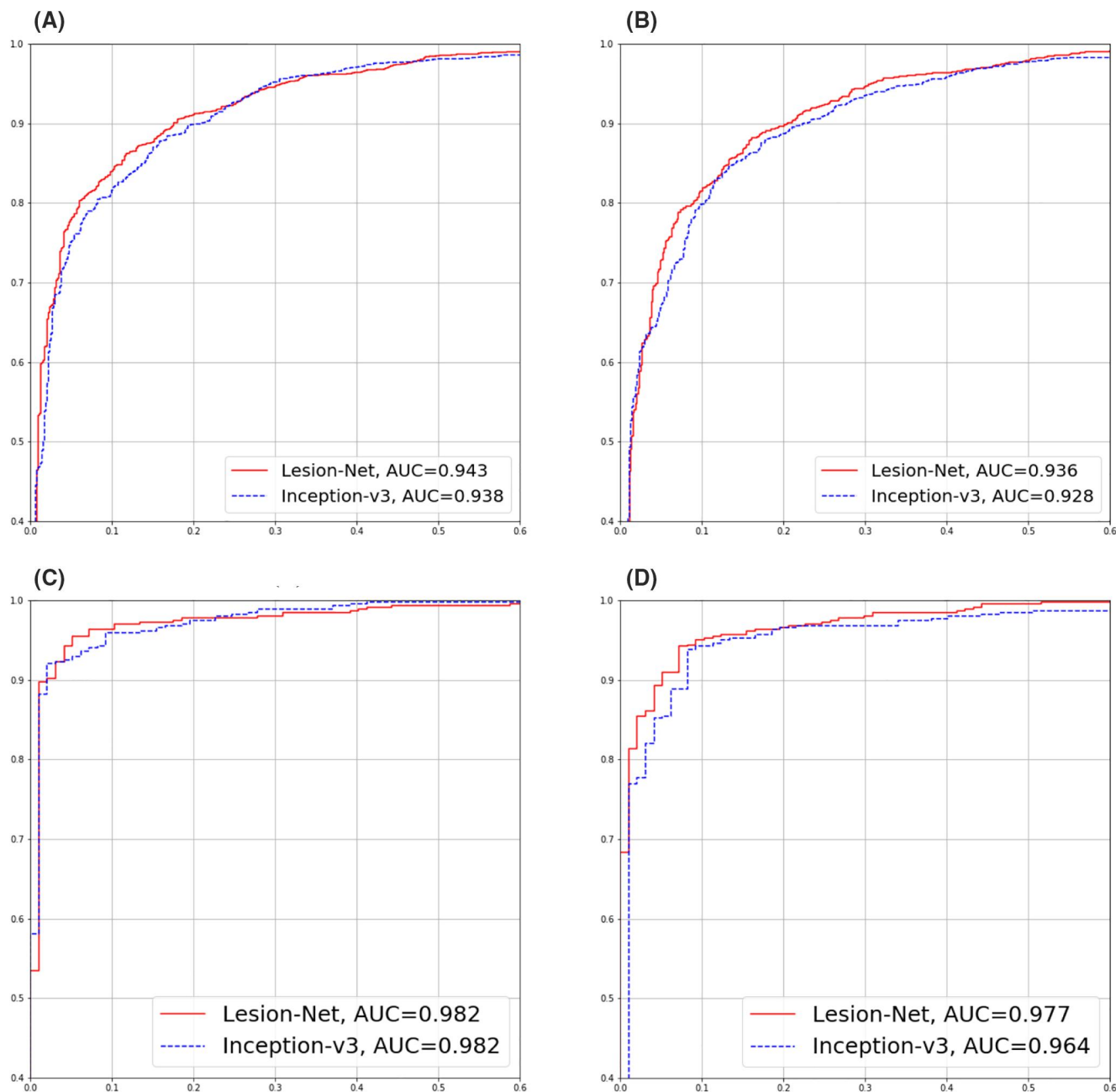
The precision and recall results of the lesions in the lesion detection models are summarised in Table 3. The precision results of lesions for DR grading in the lesion detection models were as follows: pre-retinal haemorrhage, 0.909; HE, 0.874; VH, 0.846; NV, 0.837; CWS, 0.801; fibrous proliferation, 0.780; intraretinal haemorrhage, 0.577; and MA, 0.498. The recall results of the lesions for DR grading in the lesion detection models were as follows: preretinal haemorrhage, 0.607; HE, 0.495; VH, 0.283; NV, 0.363; CWS, 0.573; fibrous proliferation, 0.687; intraretinal haemorrhage, 0.798; and MA, 0.164. These lesions were distinctive for DR grading.

## 4 | DISCUSSION

Our study demonstrates that deep learning networks could be used to train lesion-based models for referable DR with acceptable AUC, sensitivity and specificity values. Adding lesion information in the state-of-the-art CNN model, Inception-v3, improved the consistency of diagnosis in terms of the sensitivity and specificity of identifying referable DR, suggesting that lesion information improved automated referable DR identification. Converting the DR stages predicted by the five-stage models into referable/nonreferable DR achieved better AUCs than the two-class models, suggesting that learning from detailed stage information improved automated referable DR identification.

Deep learning relies heavily on the input and output information to construct complex neural networks in the hidden layers.<sup>31,32</sup> Therefore, we speculate that with detailed information given—for example, lesion information or DR stage information—the computer could ‘learn’ more and thus better perform referable DR identification.

Although some major models for referable DR identification, such as the methods proposed by Gulshan et al.,<sup>18</sup> Ting et al.<sup>15</sup> and Li



**FIGURE 3** The ROC curves for the different models. (A) The curves of five-stage models converted to referable/nonreferable DR in the internal test set, (B) the curves of the models for identifying referable/nonreferable DR in the internal test set. (C) The curves of five-stage models converted to referable/nonreferable DR in the external test set, (D) the curves of the models for identifying referable/nonreferable DR in the external test set. DR, diabetic retinopathy; ROC, receiver operating characteristic

et al.,<sup>27</sup> have achieved acceptable sensitivity (87%–100%), specificity (73%–99%) and AUC (0.89–0.99) results, these models were trained based on only colour fundus images. The limitation of these models was their lack of interpretability, as they could not present the diagnostic basis of referable DR. In addition, these models could not provide the precise results of lesion annotations or even grades of DR, providing clinicians limited reference for identifying referable DR.

The principle of referable DR was based on DR stages, and the DR stages depended on DR-related lesions. For lesion-based

automated or semiautomated DR evaluation, previous studies have made tremendous progress in detecting a few lesions (mostly MA, intraretinal haemorrhage and HE) using low-level image processing techniques. For example, Fleming et al.<sup>33</sup> and Pires et al.<sup>34</sup> performed research on utilising lesion detection results for improving DR screening performance. However, since this lesion detection method was based on low-level image processing, ad hoc detectors need to be carefully designed per lesion, with carefully engineered image features. For instance, Niemeijer et al.<sup>35</sup> used a 69-dimensional feature vector for red lesion detection, while another 83-dimensional

feature vector was used for bright lesion detection. These traditional computer-based techniques rely heavily on handcrafted features and ad hoc rules, and hence are not easily generalisable for detection of other types of lesions.

Recently, some authors employed CNN visualisation techniques and used feature importance heatmaps (red, green and blue) to show abnormal areas in fundus images.<sup>21</sup> According to the 2003 DR classification guidelines,<sup>25</sup> for moderate or severe NPDR, the model needs to specifically identify moderate NPDR lesions (intraretinal haemorrhage, HE, CWS and etc.) and PDR lesions (preretinal haemorrhage, VH, NV and fibrous proliferation). In clinical practice, it is important not only to make an accurate diagnosis, but also to look at the principle behind a decision. Our Lesion-Net model, which accommodated eight common types of DR lesions simultaneously, provides a fine diagnostic interpretation for referable DR. Our lesion detection model can remind physicians to double check for DR-related lesions and help them make a more accurate diagnosis.

However, the precision and recall results for some lesions in our models were not high enough. For IRMA and venous beading lesions, for which we found a low recognition ability in our preliminary experiment, there were also controversial lesions in our ophthalmologists' annotations. IRMA is often subtle in appearance and obscured by adjacent diabetic pathology and could be challenging to diagnose with clinical examination.<sup>36</sup> The same problem occurred in the study by Li et al.,<sup>27</sup> as they found that undetected IRMA accounted for 77.3% of all false-negative cases in their deep learning DR models. On the other hand, venous beading, which presents as vein dilation in DR, has a low prevalence in DR, as described by Chen et al.<sup>37</sup> (severe NPDR, 2.1%  $\geq 2$  quadrants; PDR, 27.1%  $\geq 2$  quadrants). Due to the distortion of veins and blurred images, the precision and recall for this type of lesion was limited. Tiny intraretinal haemorrhage lesions can be indiscernible in blurred images and can be fused with other lesions, which may increase the identification difficulties by machine learning methods. For the limited recall results for some DR lesions, we found that our model occasionally misidentified drusen as HE, and subretinal haemorrhage as VH in feature maps, which were the same colour. For MA, the sensitivity and specificity of our model is also limited, which may be due to blurred images or the fusion of multiple lesions being considered as a small haemorrhage, as described by Abramoff et al.<sup>7</sup> However, in the screening project, MA is decisive for only no DR and mild NPDR, both of which belong to nonreferable DR and therefore may have minimal influence on identifying referable DR.

Our study has some limitations. First, due to limited number of samples in the external test set, which were collected from consecutive clinical DR patients, the proportion of referable DR was high in the external test set. This may be because referable DR was more common in the tertiary hospital than in the screening project, which could affect the diagnostic value of our models. Second, the precision and recall results for some types of lesions in our study, which we mentioned above, were limited. More training data for these lesions should be collected to improve the performance of our model. Third, the

relatively lower occurrence of IRMA, venous beading, preretinal haemorrhage and NV (NV of the disc/NV elsewhere) in our training data may provide limited information for lesion detection, as they are known to be conclusive evidence for severe NPDR and PDR. Forth, images of our study were collected from highly suspect DR databases, which may influence the results of DR referral in real-world circumstances.

In conclusion, an automated and interpretable referable DR screening model was established. Our model can better perform referable DR identification and provide lesion and stage information, wherein doctors can 'double check' their findings, as these are important in clinical practice. Our work is important as it advances deep learning-based DR identification by providing clinical interpretability, as well as opening up the possibility of making precise screening and monitoring for referable DR more accessible for patients with diabetes.

## ACKNOWLEDGMENTS

The authors would like to extend a special thank to Ruohan Han, Huan Chen, Chenxi Zhang, Xinyu Zhao, Shuran Wang, Song Xia and Wenda Sui from Peking Union Medical College Hospital, Beijing, China; Yuan-yuan Xiao, Di Sun, Jingjing Huang, Haixia Bai, Dongxia Yang, Jing Zhao, Yu Mao, Mengyu Zhang, Shuang Wang, Yongpeng Zhang, Rong Huang, Dandan Ma, Qi Zhang and Bin Mo from Beijing Tongren Hospital, Beijing, China; Meng Cheng, You Wang, Hui Liu, Shaohui Gao, Zhaoxia Zhao, Xiao Chen, Hongling Chen, Shuyin Li, Dong Qin and Yanting Wang from Henan Eye Hospital, Henan, China; and Yue Yin, Tong Zhao, Yang Gao from China-Japan Friendship Hospital, Beijing, China for helping with image annotation. This work was supported by the following sources: The Chinese Academy of Medical Sciences (CAMS) Initiative for Innovative Medicine (Grant Number CAMS-I2M; 2018-I2M-AI-001); Pharmaceutical Collaborative Innovation Research Project of Beijing Science and Technology Commission (Grant Number Z191100007719002); National Key Research and Development Project (Grant Number SQ2018YFC200148); the nonprofit Central Research Institute Fund of CAMS (Grant Number 2018PT32029); the Beijing Natural Science Foundation Haidian Original Innovation Joint Fund (Grant Number 19L2062); and Beijing Natural Science Foundation (Grant Number 4202033).

## CONFLICT OF INTERESTS

Qijie Wei, Vistel(E); Jianchun Zhao, Vistel(E); Dayong Ding, Vistel(E); other authors declare no conflict of interests.

## AUTHOR CONTRIBUTIONS

Weihong Yu and Yuelin Wang designed the study and drafted the manuscript; Xuemin Jin, Bojie Hu, Yibin Li, Xiao Zhang, Yongpeng Zhang, and Bin Mo designed the study and acquired the data; Di Gong, Chan Wu, Bilei Zhang, Jingyuan Yang, Bing Li, and Mingzhen Yuan acquired the data and conducted the data analysis; Qijie Wei, Jianchun Zhao, Dayong Ding, and Xirong Li designed the study, analysed the data and revised the manuscript; Mingzhen Yuan and



Jingyuan Yang, critically revised the manuscript for important intellectual content; Youxin Chen designed the study, conducted data collection, revised the manuscript and approved the final version. All authors read and approved the final version of the manuscript.

## DATA AVAILABILITY

The data used to support the findings of this study are available from the corresponding author upon request.

## ETHICS STATEMENT

This study was approved by the Institutional Review Board of the Chinese Academy of Medical Sciences, Beijing, China. This study adhered to the tenets of the Declaration of Helsinki. The colour fundus images used in this study were retrospectively obtained and deidentified. An ethical review form is obtained.

## ORCID

Yuelin Wang  <https://orcid.org/0000-0001-9843-6280>

Miao Yu  <https://orcid.org/0000-0001-7232-8150>

## REFERENCES

- Sabanayagam C, Banu R, Chee ML, et al. Incidence and progression of diabetic retinopathy: a systematic review. *Lancet Diabetes Endocrinol*. 2019;7:140-149.
- Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: the technical and clinical considerations. *Prog Retin Eye Res*. 2019;72:100759.
- Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. 2012;35:556-564.
- Yang W, Lu J, Weng J, et al. Prevalence of diabetes among men and women in China. *N Engl J Med*. 2010;362:1090-1101.
- Li HK, Horton M, Bursell SE, et al. Telehealth practice recommendations for diabetic retinopathy, second edition. *Telemed J E Health*. 2011;17:814-837.
- Leasher JL, Bourne RR, Flaxman SR, et al. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 1990 to 2010. *Diabetes Care*. 2016;39(9):1643-1649.
- Abramoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol* 2013;131:351-357.
- Scanlon PH, Aldington SJ, Stratton IM. Delay in diabetic retinopathy screening increases the rate of detection of referable diabetic retinopathy. *Diabet Med*. 2014;31:439-442.
- Scanlon PH, Stratton IM, Leese GP, et al. Screening attendance, age group and diabetic retinopathy level at first screen. *Diabet Med*. 2016;33:904-911.
- Stein JD, Kapoor KG, Tootoo JL, et al. Access to ophthalmologists in states where optometrists have expanded scope of practice. *JAMA Ophthalmol* 2018;136:39-45.
- Kapoor R, Walters SP, Al-Aswad LA. The current state of artificial intelligence in ophthalmology. *Surv Ophthalmol*. 2019;64:233-240.
- Sasso FC, Pafundi PC, Gelso A, et al. Telemedicine for screening diabetic retinopathy: the NO BLIND Italian multicenter study. *Diabetes Metab Res Rev*. 2019;35(3):e3113.
- Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5:e1221-e1234.
- Abramoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57:5200-5206.
- Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *J Am Med Assoc*. 2017;318:2211-2223.
- Raumviboonsuk P, Krause J, Chotcommwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2:25.
- Son J, Shin JY, Kim HD, Jung KH, Park KH, Park SJ. Development and validation of deep learning models for screening multiple findings in retinal fundus images. *Ophthalmology*. 2019;127:85-94.
- Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J Am Med Assoc*. 2016;316:2402-2410.
- Raman R, Srinivasan S, Virmani S, Sivaprasad S, Rao C, Rajalakshmi R. Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. *Eye (Lond)*. 2019;33:97-109.
- Bhaskaranand M, Ramachandra C, Bhat S, et al. Automated diabetic retinopathy screening and monitoring using retinal fundus image analysis. *J Diabetes Sci Technol*. 2016;10:254-261.
- Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124:962-969.
- Wang YL, Yang JY, Yang JY, Zhao XY, Chen YX, Yu WH. Progress of artificial intelligence in diabetic retinopathy screening. *Diabetes Metab Res Rev*. 2020:e3414.
- Diabcometic Retinopathy Detection*. Kaggle. <https://www.kaggle.com/c/diabetic-retinopathy-detection>.
- Rudnisky CJ, Tennant MT, de Leon AR, Hinz BJ, Greve MD. Benefits of stereopsis when identifying clinically significant macular edema via teleophthalmology. *Can J Ophthalmol*. 2006;41:727-732.
- Wilkinson CP, Ferris FL, 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110:1677-1682.
- Li F, Liu Z, Chen H, Jiang M, Zhang X, Wu Z. Automatic detection of diabetic retinopathy in retinal fundus photographs based on deep learning algorithm. *Transl Vis Sci Technol*. 2019;8:4.
- Li Z, Keel S, Liu C, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes Care*. 2018;41:2509-2516.
- Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264-1272.
- Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527.
- Dai L, Fang R, Li H, et al. Clinical report guided retinal microaneurysm detection with multi-sieving deep learning. *IEEE Trans Med Imag*. 2018;37:1149-1161.
- Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167-175.
- Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunovic H. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1-29.
- Fleming AD, Goatman KA, Philip S, et al. The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy. *Br J Ophthalmol*. 2010;94:706-711.
- Pires R, Carvalho T, Spurling G, et al. Automated multi-lesion detection for referable diabetic retinopathy in indigenous health care. *PLoS One*. 2015;10:e0127664.

35. Niemeijer M, Abramoff MD, van Ginneken B. Information fusion for diabetic retinopathy CAD in digital color fundus photographs. *IEEE Trans Med Imag*. 2009;28:775-785.
36. Arya M, Sorour O, Chaudhri J, et al. Distinguishing intraretinal microvascular abnormalities from retinal neovascularization using optical coherence tomography angiography. *Retina*. 2020;40(9):1686-1695.
37. Chen L, Zhang X, Wen F. Venous beading in two or more quadrants might not be a sensitive grading criterion for severe nonproliferative diabetic retinopathy. *Graefes Arch Clin Exp Ophthalmol*. 2018;256:1059-1065.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Wang Y, Yu M, Hu B, et al. Deep learning-based detection and stage grading for optimising diagnosis of diabetic retinopathy. *Diabetes Metab Res Rev*. 2021;37:e3445. <https://doi.org/10.1002/dmrr.3445>