





Development and evaluation of a deep learning model for the detection of multiple fundus diseases based on colour fundus photography

Bing Li ,^{1,2} Huan Chen,^{1,2} Bilei Zhang ,^{1,2} Mingzhen Yuan,³ Xuemin Jin,⁴ Bo Lei,⁵ Jie Xu,³ Wei Gu,⁶ David Chuen Soong Wong ,⁷ Xixi He,⁸ Hao Wang,⁸ Dayong Ding,⁸ Xirong Li,⁹ Youxin Chen ,^{1,2} Weihong Yu^{1,2}

► Additional material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bjophthalmol-2020-316290>).

For numbered affiliations see end of article.

Correspondence to

Dr Weihong Yu;
536273640@qq.com Professor
Youxin Chen, Ophthalmology,
Peking Union Medical College
Hospital, Beijing, China;
chenyx@pumch.cn

WY and YC contributed equally.

Received 23 March 2020
Revised 24 January 2021
Accepted 16 February 2021

ABSTRACT

Aim To explore and evaluate an appropriate deep learning system (DLS) for the detection of 12 major fundus diseases using colour fundus photography.

Methods Diagnostic performance of a DLS was tested on the detection of normal fundus and 12 major fundus diseases including referable diabetic retinopathy, pathologic myopic retinal degeneration, retinal vein occlusion, retinitis pigmentosa, retinal detachment, wet and dry age-related macular degeneration, epiretinal membrane, macula hole, possible glaucomatous optic neuropathy, papilledema and optic nerve atrophy. The DLS was developed with 56 738 images and tested with 8176 images from one internal test set and two external test sets. The comparison with human doctors was also conducted.

Results The area under the receiver operating characteristic curves of the DLS on the internal test set and the two external test sets were 0.950 (95% CI 0.942 to 0.957) to 0.996 (95% CI 0.994 to 0.998), 0.931 (95% CI 0.923 to 0.939) to 1.000 (95% CI 0.999 to 1.000) and 0.934 (95% CI 0.929 to 0.938) to 1.000 (95% CI 0.999 to 1.000), with sensitivities of 80.4% (95% CI 79.1% to 81.6%) to 97.3% (95% CI 96.7% to 97.8%), 64.6% (95% CI 63.0% to 66.1%) to 100% (95% CI 100% to 100%) and 68.0% (95% CI 67.1% to 68.9%) to 100% (95% CI 100% to 100%), respectively, and specificities of 89.7% (95% CI 88.8% to 90.7%) to 98.1% (95% CI 97.7% to 98.6%), 78.7% (95% CI 77.4% to 80.0%) to 99.6% (95% CI 99.4% to 99.8%) and 88.1% (95% CI 87.4% to 88.7%) to 98.7% (95% CI 98.5% to 99.0%), respectively. When compared with human doctors, the DLS obtained a higher diagnostic sensitivity but lower specificity.

Conclusion The proposed DLS is effective in diagnosing normal fundus and 12 major fundus diseases, and thus has much potential for fundus diseases screening in the real world.

INTRODUCTION

Colour fundus photography (CFP) plays an important role in detecting prevalent vision-threatening fundus diseases such as diabetic retinopathy (DR), retinal vein occlusion (RVO), age-related macular degeneration (AMD) and glaucoma. According to recent epidemiological studies, approximately 79.6 million people worldwide will have glaucoma by 2020,¹ while the

number of people with AMD is expected to reach around 200 million.² The prevalence of diabetes around the world will reach 592 million people by 2035,³ with one-third affected by DR.^{4,5} However, medical services are extremely limited worldwide. For example, in mainland China, the ophthalmic human resource at the country level was only 0.14 per thousand people according to a survey in 2014.⁶ This serious situation imposed a substantial burden on the large-scale screening of multiple fundus diseases for early detection.

Deep learning system (DLS)-based diagnosing and grading in ophthalmology has progressed rapidly in many conditions, including cataracts,^{7,8} DR,⁹⁻¹¹ glaucoma,¹² retinopathy of prematurity (ROP),^{13,14} AMD^{15,16} and macular telangiectasia type 2.^{17,18} However, current studies mostly focus on one or only a few (less than five) diseases.^{19,20} To the best of our knowledge, there are still lack of efficient DL models for multiple disease (especially more than 10) recognition using CFPs. We attribute this absence to two factors: the difficulties of establishing a large-scale multidisease data set for training and validation and the technical challenges of developing a DLS suited not only for separating abnormal and normal CFPs but also for distinguishing one disease from many others.

Recently, Son *et al*²¹ proposed a DLS for the detection of 12 major fundus abnormalities using 12 binary classification models, which could help greatly on the detection of retinal lesions. However, for disease recognition, it still needs professional interpretation, which may bring obstacles for screening and AI-assisted diagnosis if there is no trained ophthalmologists available. Also, the application of a panel of binary classification models will take much more time and computer resources than a single multiclassification model. This paper aims to develop an automated screening DLS for multiple major fundus diseases, which could be of great significance for clinical practice in future.

METHODS

The current study complied with the Declaration of Helsinki and was approved by the Ethics committee of Peking Union Medical College Hospital (Number S-K631). The review board waived the need to obtain informed patient consent



© Author(s) (or their employer(s)) 2021. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Li B, Chen H, Zhang B, *et al*. *Br J Ophthalmol* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bjophthalmol-2020-316290

because of the retrospective study design and the use of fully anonymised CFPs.

Image acquisition and data sets

The selection of diseases was decided according to their prevalence and morbidity, also taking into account their clinical potential for screening using CFPs. Hence, in addition to normal fundus images, we selected 12 major fundus diseases including nine retina diseases: referable DR, pathologic myopic (PM) retinal degeneration, RVO, retinitis pigmentosa (RP), retinal detachment (RD), wet and dry AMD, epiretinal membrane (ERM) and macula hole (MH) and three optic nerve disorders: possible glaucomatous optic neuropathy (GON), papilledema and optic nerve atrophy. The imaging diagnosis was made on standard diagnostic criteria (online supplemental eTable1). Although dry and wet AMD can be considered as the same disease of different stages,²² we still classified them into two categories considering their potential difference on treatments and prognoses.

Since there were no publicly available data sets for the detection of multiple fundus diseases, we acquired and annotated a data set for the development and internal test of the DLS. To test the generalisability of the model, we also collected CFPs from an independent tertiary medical centre forming the external test set A and three primary hospitals forming the external test set B.

Development set

A total of 56 738 CFPs taken between January 2014 and December 2018 were collected from three participating centres (Henan Provincial Peoples' Hospital, Zhengzhou, Henan, Beijing Tongren Hospital, Beijing and Beijing Aier Intech Eye Hospital, Beijing). These images formed the development data set for the models' training and validation.

Test sets

Another 8176 CFPs were collected for the DLS testing. Among them, 3579 were from the same source of the development set and ensured that the sample size of each disease reached over 100, forming the internal test set. Another 1245 CFPs from 757 patients were collected from another independent tertiary medical centre (Peking Union Medical College Hospital) from 1 January 2019 to 30 June 2019, as the external test set A. The

last 3352 CFPs from 2558 patients were collected from three primary hospitals from 4 July 2017 to 14 September 2020, as the external test set B.

For each patient enrolled, only one image of each eye could be included. The detailed inclusion and exclusion criteria are provided in the online material (online supplemental file).

After preprocessing and desensitisation, the development data set was separated into a training set and a validation set with the ratio of 4:1, according to the patients' number, which means that the bilateral CFPs of the same patient were assigned together to either the training set or validation set. This process was organised randomly. The three test sets were maintained independently to test the performance and generalisation of the DLS.

Online annotation was carried out to label the images as normal fundus or the 12 selected diseases. A total of 17 senior board-certified ophthalmologists (with 5–12 years of experience) were randomly assigned for image annotation. Thirteen of them were assigned to label the development data set and internal test set. The other four doctors were assigned to label the external test sets. Images in the test sets were labelled three times by different ophthalmologists to obtain high reliability. Consistent labels by all three doctors were retained. If the label was only agreed by two doctors, then the final decision would be made by a fourth, more senior ophthalmologists (with over 10 years of experience). Images with no consistent labels or those annotated with poor quality, such as loss of focus, misalignment, excessive brightness or dimness, were excluded.

Development of evaluation of the DLS

The DLS was designed using the convolutional neural network (CNN) of SeResNext50²³ network as a multilabel model selected from four-candidate CNNs with two parallel branches at the fully connected layer, one for the distinguish of normal and abnormalities and the other for the recognition of diseases it predicted to have, which could be more than one kind of diseases, simultaneously. The details are available in online materials (online supplemental eFigure1).

The performance of the DLS was evaluated on the three test sets. We used the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity and specificity for assessments. The metrics were calculated for each label instead of each image, since one image could be annotated with more than one label.

Table 1 The sample size of normal fundus and 12 fundus diseases in the five datasets

Label	Development set		Test sets		
	Training set N=46 501	Validation set N=10 237	Intern test set N=3 579	External test set A N=1 245	External test set B N=3 352
Normal fundus	19146 (41.2)	4315 (9.3)	1053 (29.4)	441 (12.3)	1804 (50.4)
Retinal vein occlusion	3528 (7.6)	967 (2.1)	531 (14.8)	54 (1.5)	123 (3.4)
Referable diabetic retinopathy	2701 (5.8)	642 (1.4)	285 (8.0)	292 (8.2)	388 (10.8)
Pathological myopic retinal degeneration	8243 (17.7)	989 (2.1)	192 (5.4)	84 (2.3)	113 (3.2)
Retinitis pigmentosa	587 (1.3)	137 (0.3)	130 (3.6)	62 (1.7)	38 (1.1)
Retinal detachment	315 (0.7)	88 (0.2)	110 (3.1)	5 (0.1)	14 (0.4)
Epiretinal membrane	2403 (5.2)	544 (1.2)	268 (7.5)	36 (1.0)	165 (4.6)
Dry age-related macular degeneration	2669 (5.7)	808 (1.7)	267 (7.5)	86 (2.4)	404 (11.3)
Wet age-related macular degeneration	1564 (3.4)	433 (0.9)	146 (4.1)	67 (1.9)	75 (2.1)
Macular hole	266 (0.6)	59 (0.1)	137 (3.8)	1 (0.0)	14 (0.4)
Possible glaucomatous optic neuropathy	3648 (7.8)	544 (1.2)	270 (7.5)	79 (2.2)	227 (6.3)
Papilledema	2882 (6.2)	682 (1.5)	228 (6.4)	78 (2.2)	82 (2.3)
Optic nerve atrophy	1459 (3.1)	462 (1.0)	202 (5.6)	23 (0.6)	150 (4.2)

The results are presented with: number (%).

Table 2 The model's performance on the three test sets

	Intern test set			External test set A			External test set B		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
Normal fundus	0.945 (0.938, 0.953)	0.967 (0.961, 0.973)	0.989 (0.985, 0.992)	0.951 (0.945, 0.958)	0.787 (0.774, 0.800)	0.956 (0.950, 0.963)	0.862 (0.855, 0.868)	0.895 (0.889, 0.901)	0.955 (0.951, 0.959)
Referable diabetic retinopathy	0.804 (0.791, 0.816)	0.897 (0.888, 0.907)	0.950 (0.942, 0.957)	0.990 (0.986, 0.993)	0.810 (0.797, 0.823)	0.965 (0.960, 0.971)	0.923 (0.918, 0.928)	0.881 (0.874, 0.887)	0.986 (0.984, 0.988)
Retinal vein occlusion	0.964 (0.958, 0.970)	0.969 (0.963, 0.974)	0.994 (0.992, 0.997)	0.963 (0.957, 0.969)	0.960 (0.953, 0.966)	0.992 (0.990, 0.995)	1.000 (1.000, 1.000)	0.986 (0.983, 0.988)	0.999 (0.998, 0.999)
Pathological myopic retinal degeneration	0.958 (0.952, 0.965)	0.971 (0.965, 0.976)	0.988 (0.984, 0.991)	0.952 (0.945, 0.959)	0.990 (0.986, 0.993)	0.992 (0.989, 0.995)	0.991 (0.989, 0.993)	0.938 (0.934, 0.943)	0.989 (0.988, 0.991)
Retinitis pigmentosa	0.962 (0.955, 0.968)	0.978 (0.973, 0.983)	0.996 (0.994, 0.998)	1.000 (1.000, 1.000)	0.988 (0.985, 0.992)	1.000 (0.999, 1.000)	0.895 (0.889, 0.901)	0.977 (0.974, 0.980)	0.996 (0.995, 0.998)
Retinal detachment	0.973 (0.967, 0.978)	0.981 (0.977, 0.986)	0.996 (0.993, 0.998)	0.800 (0.787, 0.813)	0.981 (0.977, 0.986)	0.992 (0.990, 0.995)	0.786 (0.778, 0.794)	0.987 (0.985, 0.990)	0.992 (0.990, 0.993)
Epilethelial membrane	0.918 (0.909, 0.927)	0.923 (0.915, 0.932)	0.968 (0.963, 0.974)	0.694 (0.679, 0.709)	0.940 (0.930, 0.947)	0.938 (0.931, 0.946)	0.745 (0.737, 0.754)	0.889 (0.883, 0.895)	0.934 (0.929, 0.938)
Dry age-related macular degeneration	0.858 (0.846, 0.869)	0.939 (0.931, 0.947)	0.976 (0.971, 0.981)	0.895 (0.885, 0.905)	0.975 (0.970, 0.978)	0.973 (0.967, 0.978)	0.718 (0.709, 0.727)	0.941 (0.937, 0.946)	0.968 (0.964, 0.971)
Wet age-related macular degeneration	0.842 (0.831, 0.854)	0.953 (0.946, 0.960)	0.964 (0.958, 0.970)	0.925 (0.917, 0.934)	0.894 (0.884, 0.904)	0.974 (0.969, 0.979)	0.920 (0.915, 0.925)	0.971 (0.968, 0.975)	0.988 (0.986, 0.990)
Macular hole	0.876 (0.865, 0.887)	0.963 (0.957, 0.970)	0.978 (0.973, 0.983)	1.000 (1.000, 1.000)	0.978 (0.974, 0.983)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	0.966 (0.962, 0.969)	1.000 (0.999, 1.000)
Possible GON	0.804 (0.791, 0.817)	0.934 (0.925, 0.942)	0.953 (0.946, 0.960)	0.646 (0.630, 0.661)	0.938 (0.930, 0.946)	0.931 (0.923, 0.939)	0.797 (0.790, 0.805)	0.930 (0.925, 0.935)	0.946 (0.942, 0.950)
Papilledema	0.904 (0.894, 0.913)	0.950 (0.943, 0.957)	0.980 (0.975, 0.985)	0.756 (0.742, 0.770)	0.990 (0.986, 0.993)	0.991 (0.989, 0.994)	0.756 (0.748, 0.764)	0.975 (0.972, 0.978)	0.990 (0.988, 0.992)
Optic nerve atrophy	0.950 (0.943, 0.958)	0.946 (0.938, 0.953)	0.989 (0.985, 0.992)	0.826 (0.814, 0.838)	0.996 (0.994, 0.998)	0.996 (0.994, 0.998)	0.680 (0.671, 0.689)	0.952 (0.947, 0.956)	0.955 (0.951, 0.959)

The results are presented with: percentage (95% CI).

AUC, area under the receiver operating characteristic curve; GON, glaucomatous optic neuropathy.

Information learnt in our automated method was visualised for further clinical review using Class Activation Map (CAM)²⁴ which is a CNN's visualisation technique that can identify the importance of the image regions by projecting back the weights of the classification layer on the convolutional feature maps obtained from the last convolution layer.

Comparison of the DLS with human doctors

To assess whether the DLS has reached a comparable diagnostic performance with human doctors, four ophthalmic residents were tested using the external test set B. Each of them was assigned randomly with one quarter samples of the whole set and annotated online and then compared the performance with DLS, which annotated the same images.

All statistical analyses, including ROC curves, were carried out using the programming language Python (V2.7; Python Software Foundation; Wilmington, Delaware, USA). The results of the indicators are presented as values with 95% CIs.

RESULTS

A total of 64 914 CFPs were enrolled in this study with the field of 35–55 degrees of the posterior pole covering the whole area of macula and the optic disc. The DLS was trained and validated using 46 501 and 10 237 images, respectively, and evaluated on the three test sets with 3 579 images (2 635 patients with a mean age (\pm SD) of 55.4 \pm 18.3 ranging from 2 to 96), 1 245 images (757 patients with a mean age (\pm SD) of 48.7 \pm 18.0 ranging from 4 to 89) and 3 352 images (2 558 patients with a mean age (\pm SD) of 52.6 \pm 20.6 ranging from 3 to 97), respectively. The numbers of images in each category of the internal test set were all over 100, which ensured the reliability of the test results. The two external test sets represented a real clinical scenario and the disease distribution of both tertiary medical centre and primary hospitals in China over a certain period of time (table 1). CFPs with more than one label in the training, validation internal test set, external test sets A and B were 3 202 (6.9%), 488 (4.8%), 334 (9.3%), 70 (5.6%) and 217 (6.5%), respectively.

The model performance on the test sets

We developed a late-fusion multilabel model as well as 12 binary classification models for comparison, and the former achieved a higher mean average precision on validation set with statistical significance ($p=0.020$) (online supplemental eTables 2 and 3). The ROC curves were also listed online (online supplemental eFigure 2 and 3). We, therefore, selected the late-fusion multilabel model for testing. The threshold of the model on validation set was listed in online materiel (online supplemental eTable 4). The AUCs in the internal test set and the two external test sets were 0.950 (95% CI 0.942 to 0.957) to 0.996 (95% CI 0.994 to 0.998), 0.931 (95% CI 0.923 to 0.939) to 1.000 (95% CI 0.999 to 1.000) and 0.934 (95% CI 0.929 to 0.938) to 1.000 (95% CI 0.999 to 1.000), with corresponding sensitivities of 80.4% (95% CI 79.1% to 81.6%) to 97.3% (95% CI 96.7% to 97.8%), 64.6% (95% CI 63.0% to 66.1%) to 100% (95% CI 100% to 100%) and 68.0% (95% CI 67.1% to 68.9%) to 100% (95% CI 100% to 100%), and corresponding specificities of 89.7% (95% CI 88.8% to 90.7%) to 98.1% (95% CI 97.7% to 98.6%), 78.7% (95% CI 77.4% to 80.0%) to 99.6% (95% CI 99.4% to 99.8%) and 88.1% (95% CI 87.4% to 88.7%) to 98.7% (95% CI 98.5% to 99.0%), respectively. For the major blindness leading diseases, the AUCs of referable DR, possible GON, dry and wet form AMD in the external test sets were 0.965 (95% CI 0.960 to 0.971) to 0.986 (95% CI 0.984 to 0.988), 0.931 (95% CI 0.923

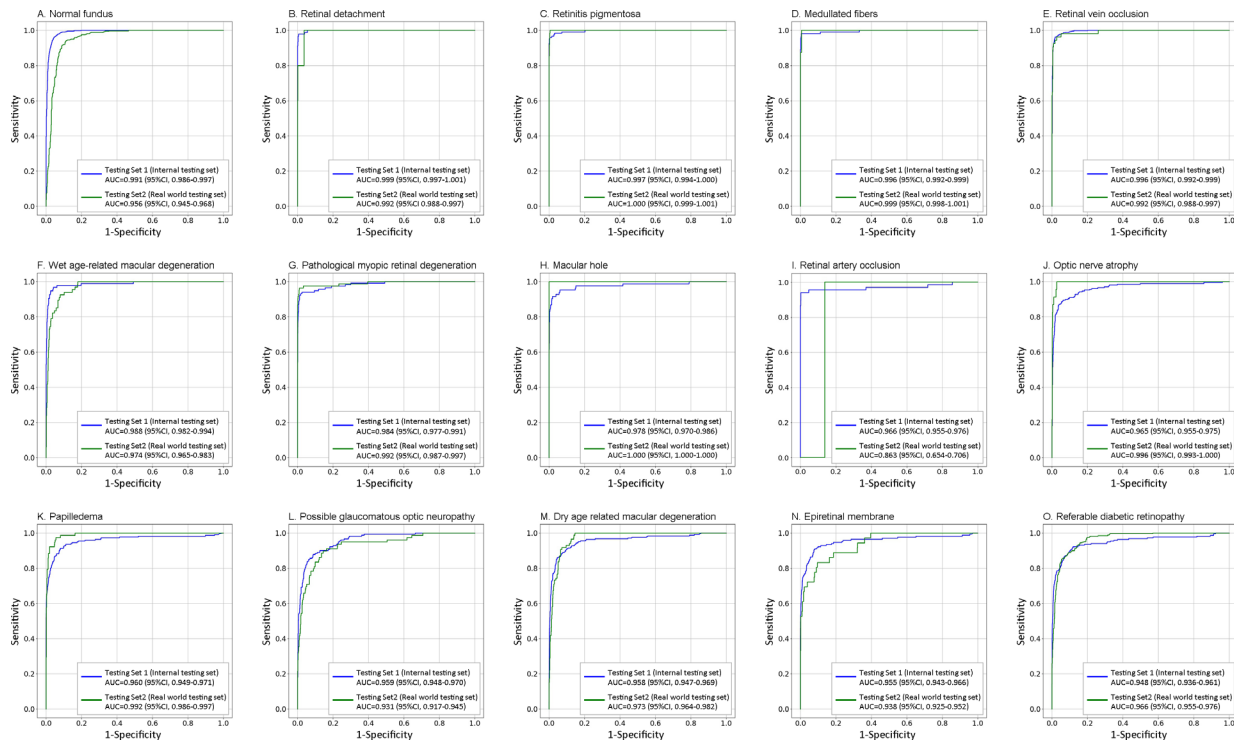


Figure 1 The receiver operating characteristic curves of the deep learning system tested in the internal test set.

to 0.939) to 0.946 (95% CI 0.942 to 0.950) and 0.968 (95% CI 0.964 to 0.971) to 0.988 (95% CI 0.986 to 0.990), respectively. **Table 2** shows the results of the AUC, sensitivity and specificity, of the DLS tested on the three test sets. The ROC curves of the DLS tested in the internal set were as **figure 1** shows. Other ROC results tested in the external sets are listed in the online material (online supplemental eFigure 4 and 5).

To further understand the model's performance, we used heat maps for visualisation and clinical review. **Figure 2** shows heat maps of the true-positive reports of normal fundus and 12 fundus diseases on the external test sets. Different colours mark subregions with different degrees of activation of the DLS, which increase progressively from blue to red as indicated by the colour bar. The heat maps indicate that the features extracted by the model generally present a high consistency with human doctors' diagnostic basis in real clinical work according to the specific lesions on CFPs. Some false-positive and false-negative cases indicated that the DLS seemed to miss some fine abnormalities like the change of the disc rim, optic disc pit in possible GON or small MH (**figure 3**).

We also noticed that the model achieved a relatively lower sensitivity on the detection of possible GON. To further interpret and prove the model's performance, we compared our DLS with some other specialised GON detecting models using public available data set. The test was performed on Retinal Fundus Glaucoma Challenge, REFUGE (<https://refuge.grand-challenge.org>) test set, which contains 400 fundus images with 360 normal fundus and 40 glaucoma. We achieved 0.955 AUC and 0.931 reference sensitivity, which rank six and four among all the 12 participating team, that is comparable to the state-of-the-art models (reference sensitivity: 0.725 to 0.976, AUC: 0.846 to 0.989).²⁵ The detailed comparison results were available in online material (online supplemental eTable 5 and eFigure 4).

The comparison between human doctors and the DLS model

The mean sensitivity and specificity of the four human doctors were 69.5%, 75.7%, 74.0% and 71.1%, and 98.1%, 97.8%, 97.8% and 97.6%, respectively. The corresponding DLS model's sensitivity and specificity were 90.2%, 86.8%, 84.0% and 82.4%, and 97.6%, 92.6%, 93.7% and 93.6%, respectively. Statistical analysis (Mann-Whitney U test) showed that the DLS achieves significant higher sensitivity comparing with two of the four doctors and lower specificity comparing with all four doctors. Detailed results are available in online materials (online supplemental eTable 6).

DISCUSSIONS

DL models for the detection of multiple fundus diseases

Previous studies have reported a large number of DLSs used for multiclassification, such as the detection of several diseases or severity of DR and AMD using CFPs or optical coherence topography.^{9 16 26} There have also been studies focused on the detection of multiple fundus lesions recently.²¹ The detection of certain fundus diseases using DLS exceeding 10 categories remains very rare. Choi *et al*²⁷ described automated differentiation between normal fundus and nine retinal diseases but achieved an accuracy of only 36.7% for all 10 classes. Comparing with their study, our work was carried out using a large data set with over 60 000 images acquired from real clinical patients. The DLS developed by Son *et al*²¹ proposed a deep learning method for detecting multiple lesion-level abnormalities in colour fundus images. The strength to their study is that the detected lesions provide a more intuitive interpretation than holistic predictions as made by the prior art. However, as there lacks a one-to-one correspondence between lesions and fundus diseases, a gap naturally exists when converting lesion-level findings to diseases, which is left untouched by Son *et al* in this work, we take a orthogonal direction, making a novel attempt to directly recognise 12 fundus

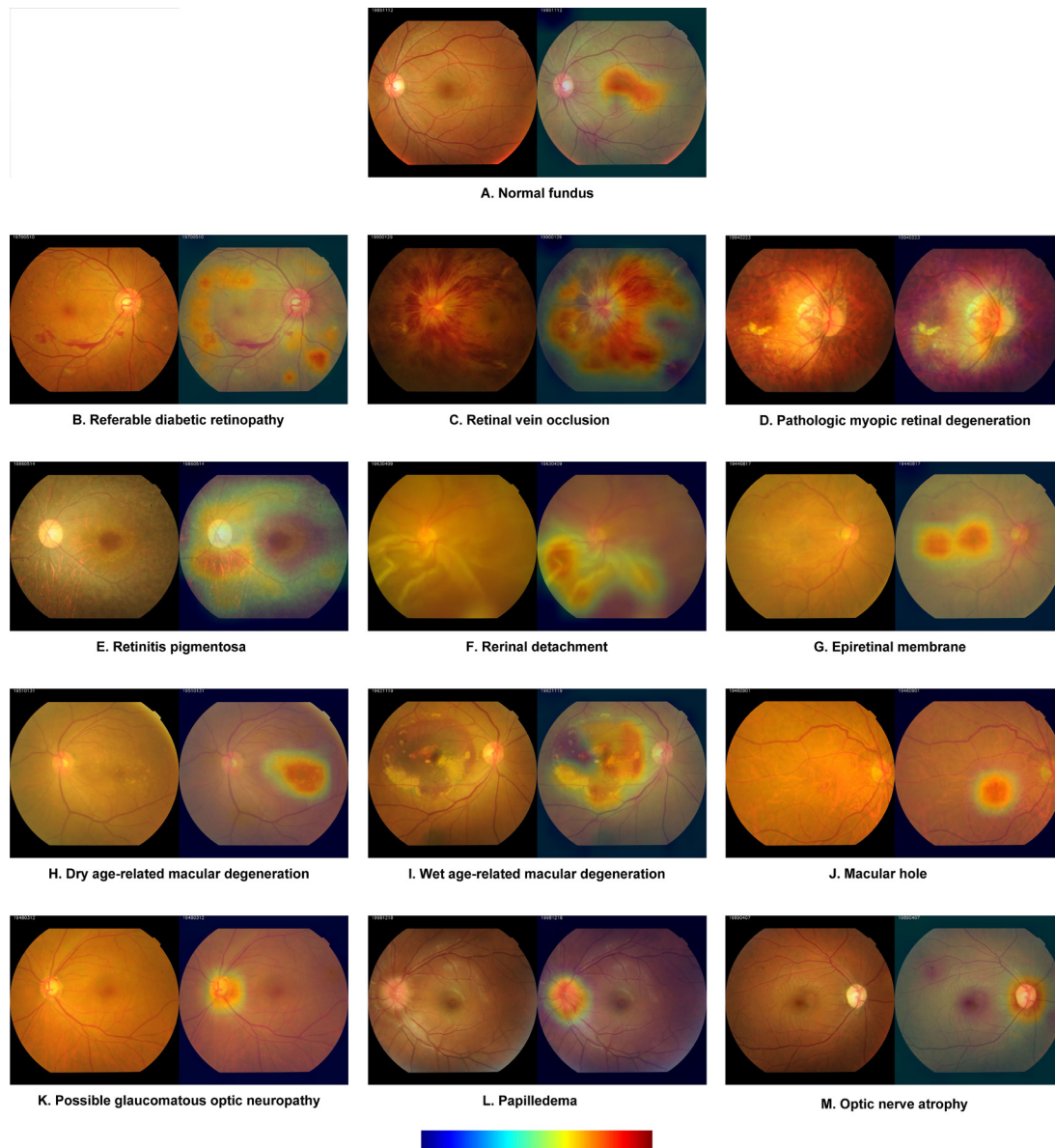


Figure 2 CFPs and visualisation heat maps of true-positive cases on the internal test set. The colour bar mark subregions with different active intensities of the model, which increase progressively from the blue end to the red end. These heat maps represent the ability of our method to objectively distinguish different diseases. CFP, colour fundus photography.

diseases from a given colour fundus image. Moreover, we adopt the CAM technique to visualise which part of the given image is responsible for the final prediction.

Furthermore, the diseases selected in this study mostly comprise leading causes of blindness that need early detection and intervention covering a broad spectrum including retinal vascular diseases (RVO, referable DR), retinal degeneration diseases (PM retinal degeneration, RP, RD), macular disease (ERM, AMD and MH) and optic nerve disorders (possible GON, papilledema and optic nerve atrophy). Most of them have rarely been reported in previous studies.

Development and selection of the models

The models developed for multidisease detection were diverse in previous studies. The scenario targeted most often by machine learning methods for applications in ophthalmology is image classification,²⁸ which is typically used in retinal analysis for automatic screening. Multiclass classification is used²⁸ to detect

the type of disease present or to accurately determine the stage of disease. This has been done for DR^{10 11} and ROP.^{29 30} In the case of multiclass classification, images belong to only one of the mutually exclusive categories. Choi *et al*²⁷ reported a multidisease recognition model that applied a method of classification to classify fundus images into different categories of retinal diseases for diagnosis. The authors attributed part of the dissatisfactory performance of the model to decreased expected accuracy as the number of categories multiplied, which has been demonstrated in previous studies.³¹ However, mutually exclusive multiclassification model may not be unsuitable for multiple disease recognition since some fundus diseases may coexist. For example, patients could have DR and ERM simultaneously,³² and the incidence rate of open-angle glaucoma in patients with RVO is significantly higher than that in the general population.³³ Our multilabel model was developed with the modified feature layer of SeResNext50 in order to simultaneously classify abnormal versus normal CFP images and to accurately detect the presence

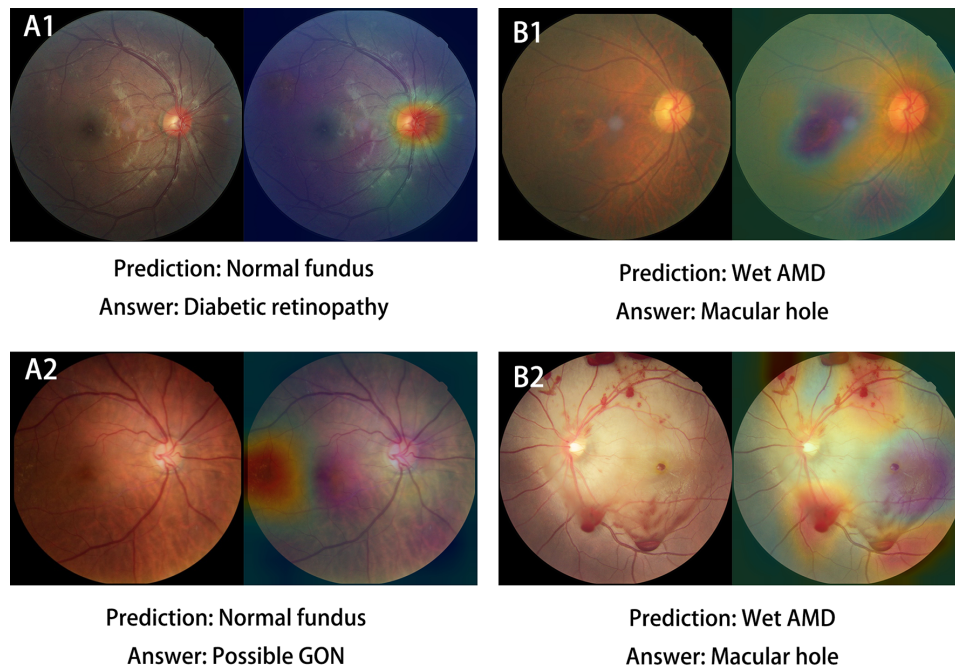


Figure 3 The fundus image and corresponding heat maps of some cases of false positive and false negative results predicted by the DLS in the validation set. A1 and A2 are false negative cases: the DLS miss diagnosed referable diabetic retinopathy (A1) and possible GON (A2) to normal fundus; B1 and B2 are false positive cases: the DLS miss diagnosed macular hole to wet age-related macular degeneration. AMD, age-related macular degeneration; DLS, deep learning system; GON, glaucomatous optic neuropathy.

of multiple diseases. We combined the two steps into a single model to simplify implementation in future clinical practice.

The data sets and the model's performance

Our model was trained and tested in real clinical data sets, and this was an important feature of the study, mimicking real screening scenarios as closely as possible at this early stage of development. To assure the accuracy, diversity and reliability of the data sets, we used CFPs from real-life data sets from three different clinical centres that were annotated by 17 experienced ophthalmologists. The amount of work involved in annotating the images was formidable, and this data set was much larger than in previous studies on multidisease classification with only 279 images.²⁷ To our knowledge, this is also the largest multidisease recognition data set thus far.

Considering the future application scenarios of the model is screening especially in lower level medical places, which may be accompanied with more complex conditions and interferences while screening, we provided two external test sets from tertiary medical centre and primary hospitals, respectively. The results showed that the disease distribution was different from that of tertiary hospital. For example, the proportion of dry AMD and possible GON was much higher. Even so, the results still supported, that the DLS could do well in both scenarios, which proved the possibility of large-scale screening in the future work.

Notice that for glaucoma detection, the sensitivity of our DLS varies, which is 0.913, 0.797 and 0.646 on REFUGE, the external test set B and the external test set A, respectively. We attribute this variation to the distinct sources of the three test sets. REFUGE, as a public benchmark data set, tends to include images of less ambiguity to ensure the reliability of its ground truth. Indeed, we observed that images from this data set are typical with respect to glaucoma. Recall that the external test sets B and A were collected from primary hospitals and tertiary hospitals, respectively. Given the common practice of a referral

medical system, where cases that are less typical and thus more difficult to diagnose are to be referred from a primary hospital to a tertiary hospital, it is fair to claim that images from A were the most challenging. The increasing difficulty in glaucoma diagnosis from REFUGE to the test set B and to the test set A explains the decreasing sensitivity of the DLS to detect this condition.

The interpretation of the heat maps

The 'black box' problem of DLS has greatly limited its application and acceptance in real clinical practice. In this study, we used heat maps for visualisation. As the heat maps indicated, the features extracted by the model for prediction are very similar to human doctors' considerations. Taking referable DR as an example (figure 2B), the model precisely extracted the appropriate retinal lesions (intraretinal and preretinal haemorrhages) and provided a correct prediction. The heatmaps are also helpful on understanding the false results. For example, the heatmap indicated that in false-negative case of possible GON (figure 3 A2), the model paid almost no attention on the optic disc and failed to give the correct answer. The DLS model presented a limited performance on the detection of specific diseases like possible GON. To further interpret the results, we tested the model in a public available REFUGE dataset and proved that our DLS model presented a comparable performance with some of the other specialised GON detecting models. We attribute this variation to the distinct source of the test sets. REFUGE as a public benchmark dataset tends to include images of less ambiguity to ensure the reliability of its ground truth. Indeed, we observed that images from this dataset are typical with respect to glaucoma. Recall that the external test set A and B were collected from primary hospitals and tertiary hospitals respectively. Given the common practice of a referral medical system, where cases that are less typical and thus more difficult to diagnose are to be referred from a primary hospital to a tertiary hospital, it is fair to claim that images from A were the most challenging.

Limitations and future works

Our work has some limitations. First, while we have spent much efforts to expand our external test sets, the testing sample sizes for MH and RD, which are 19 and 15 in total, remain relatively small, as compared with the other conditions. To improve the reliability of the detection performance of the two diseases, more test samples need to be collected for future exploration. Second, the external evaluation on a clinical data set collected from tertiary hospitals (external test set A) shows that our DLS detects glaucoma with a relatively lower sensitivity of 0.646. Given that glaucoma is a major blinding disease, much work remains to be done for real-world deployment. Third, some diseases included in this study initiate from the peripheral retinal area such as RP and RD, but most of the images we used for analysis were centred by the macula fovea with the maximal field of 55 degree. Therefore, the detection of these diseases may be limited. With the future common use of ultrawide fundus camera, DLS model for this kind of CFP is of high research value. Finally, future prospective trials are needed to assess the DLS in multiple independent real clinical scenarios.

CONCLUSION

The proposed DLS showed well performance on the three test sets for the detection of normal fundus as well as 12 major fundus diseases. The application of this model may alleviate the workloads of trained specialists and provide an efficient, low-cost approach for preliminary screening in places with scarce medical resources and ophthalmologists. Further acquisition of data to broaden the extent of screening for more fundus diseases will be the next step of our work.

Author affiliations

- ¹Department of Ophthalmology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China
- ²Key Laboratory of Ocular Fundus Diseases, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China
- ³Department of Ophthalmology, Beijing Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China
- ⁴Department of Ophthalmology, Zhengzhou University First Affiliated Hospital, Zhengzhou, Henan, China
- ⁵Clinical Research Center, Henan Eye Institute, Henan Eye Hospital, Clinical Research Center, Henan Provincial People's Hospital, Zhengzhou, Henan, China
- ⁶Department of Ophthalmology, Beijing Aier Intech Eye Hospital, Beijing, China
- ⁷University of Cambridge School of Clinical Medicine, Cambridge, UK
- ⁸Vistel AI Lab, Visionary Intelligence Ltd, Beijing, China
- ⁹Key Lab of DEKE, Renmin University of China, Beijing, China

Acknowledgements The authors thank Di Gong, Hong Du, Ning Chen, Dongmei Huo, Nan Chen, Hongling Chen, Donghui Li, Meiyuan Zhu, Yanting Wang, Xiao Chen, Hui Liu, Huan Chen and Tong Zhao for their valuable contribution to this research. They devoted considerable time and effort to this work during the process of online annotation that lasted for more than 8 months.

Contributors BL contributed to the statistical analysis, drafting and revising of the manuscript. HC, BZ and MY contributed to the standard operating procedure and quality control of the datasets. XJ, BL, JX and WG contributed to the acquisition of the color fundus photograph of the datasets. DCSW contributed to the revision of the manuscript. XH and HW contributed to the models' developing, statistical analysis and preparing of the figures for the work. XL and DD contributed to the development of the models and interpretation of data, and revision of the manuscript for this study. YC and WY contributed to the conception and design of the work, revision of the manuscript and will final approval of the version to be published.

Funding CAMS Initiative for Innovative Medicine (CAMS-12M)(2018-12M-AI-001). Pharmaceutical collaborative innovation research project of Beijing Science and Technology Commission (Z191100007719002). Beijing Natural Science Foundation Haidian original innovation joint fund (19L2062). Natural Science Foundation of Beijing Municipality 4202033. The priming scientific research foundation for the junior researcher in Beijing Tongren Hospital, Capital Medical University (2018-YJJ-ZZL-052).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

- Bing Li <http://orcid.org/0000-0001-8958-2972>
 Bilei Zhang <http://orcid.org/0000-0001-9368-4939>
 David Chuen Soong Wong <http://orcid.org/0000-0002-1712-9527>
 Youxin Chen <http://orcid.org/0000-0002-7231-5058>

REFERENCES

- 1 Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol* 2006;90:262–7.
- 2 Wong WL, Su X, Li X, *et al.* Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health* 2014;2:e106–16.
- 3 Guariguata L, Whiting DR, Hambleton I, *et al.* Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract* 2014;103:137–49.
- 4 Yau JWY, Rogers SL, Kawasaki R, *et al.* Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* 2012;35:556–64.
- 5 Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* 2010;87:4–14.
- 6 Feng JJ, An L, Wang ZF, *et al.* [Analysis on ophthalmic human resource allocation and service delivery at county level in Mainland China in 2014]. *Zhonghua Yan Ke Za Zhi* 2018;54:929–34.
- 7 Gao X, Lin S, Wong TY. Automatic feature learning to grade nuclear cataracts based on deep learning. *IEEE Trans Biomed Eng* 2015;62:2693–701.
- 8 Caixinha M, Amaro J, Santos M, *et al.* In-Vivo automatic nuclear cataract detection and classification in an animal model by ultrasounds. *IEEE Trans Biomed Eng* 2016;63:2326–35.
- 9 Ting DSW, Cheung CY-L, 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. *JAMA* 2017;318:2211–23.
- 10 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;41:2509–16.
- 11 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. *JAMA* 2016;316:2402–10.
- 12 Li Z, He Y, Keel S, *et al.* Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus Photographs. *Ophthalmology* 2018;125:1199–206.
- 13 Redd TK, Campbell JP, Brown JM, *et al.* Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *Br J Ophthalmol* 2018. doi:10.1136/bjophthalmol-2018-313156. [Epub ahead of print: 23 Nov 2018].
- 14 Brown JM, Campbell JP, Beers A, *et al.* Automated diagnosis of plus disease in retinopathy of prematurity using deep Convolutional neural networks. *JAMA Ophthalmol* 2018;136:803–10.
- 15 Li F, Chen H, Liu Z, *et al.* Fully automated detection of retinal disorders by image-based deep learning. *Graefes Arch Clin Exp Ophthalmol* 2019;257:495–505.
- 16 Grassmann F, Mengelkamp J, Brandl C, *et al.* A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology* 2018;125:1410–20.
- 17 Loo J, Fang L, Cunefare D, *et al.* Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2. *Biomed Opt Express* 2018;9:2681–98.
- 18 Kihara Y, Heeren TFC, Lee CS, *et al.* Estimating retinal sensitivity using optical coherence tomography with Deep-Learning algorithms in macular telangiectasia type 2. *JAMA Netw Open* 2019;2:e188029.
- 19 Lu W, Tong Y, Yu Y, *et al.* Deep Learning-Based automated classification of Multi-Categorical abnormalities from optical coherence tomography images. *Transl Vis Sci Technol* 2018;7:741.
- 20 Liu Y-Y, Ishikawa H, Chen M, *et al.* Computerized macular pathology diagnosis in spectral domain optical coherence tomography scans based on multiscale texture and shape features. *Invest Ophthalmol Vis Sci* 2011;52:8316–22.

- 21 Son J, Shin JY, Kim HD, *et al.* Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology* 2020;127:85–94.
- 22 Ferris FL. 3Rd, Wilkinson CP, Bird A, *et al.* Clinical classification of age-related macular degeneration. *Ophthalmology* 2013;120:844–51.
- 23 Hu J, Shen L, Albanie S, *et al.* Squeeze-and-Excitation networks. *IEEE Trans Pattern Anal Mach Intell* 2020;42:2011–23.
- 24 Zhou B, Khosla A, Lapedriza A. Learning deep features for discriminative localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016:2921–9.
- 25 Orlando JI, Fu H, Barbosa Breda J, *et al.* REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal* 2020;59:101570.
- 26 Keel S, Wu J, Lee PY, *et al.* Visualizing deep learning models for the detection of Referable diabetic retinopathy and glaucoma. *JAMA Ophthalmol* 2019;137:288–92.
- 27 Choi JY, Yoo TK, Seo JG, *et al.* Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database. *PLoS One* 2017;12:e0187336.
- 28 Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, *et al.* Artificial intelligence in retina. *Prog Retin Eye Res* 2018;67:1–29.
- 29 Zhang Y, Wang L, Wu Z, *et al.* Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images. *IEEE Access* 2019;7:10232–41.
- 30 Wang J, Ju R, Chen Y, *et al.* Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine* 2018;35:361–8.
- 31 Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
- 32 Jackson TL, Nicod E, Angelis A, *et al.* Vitreous attachment in age-related macular degeneration, diabetic macular edema, and retinal vein occlusion: a systematic review and metaanalysis. *Retina* 2013;33:1099–108.
- 33 Na KI, Jeoung JW, Kim YK, *et al.* Incidence of open-angle glaucoma in newly diagnosed retinal vein occlusion: a nationwide population-based study. *J Glaucoma* 2019;28:111–8.

Online material

1. The inclusion and exclusion criteria of the datasets
2. The development and selection of the deep learning system.
3. eTable1. The standard criteria of the 12 selected fundus diseases
4. eTable2. The four multilabel models' performance using different convolutional neural networks (CNNs) tested in the validation set
 - 4.1 eTable2-1. The area under the curve (AUC) results
 - 4.2 eTable2-2. The average precision (AP) results
 - 4.3 eTable2-3. The sensitivity results
 - 4.4 eTable2-4. The specificity results
5. eTable3. The comparison of the selected multilabel model, binary classification models and the late-fusion multilabel model tested in the validation set
 - 5.1 eTable3-1. The area under the curve (AUC) results
 - 5.2 eTable3-2. The average precision (AP) results
 - 5.3 eTable3-3. The sensitivity results
 - 5.4 eTable3-4. The specificity results
6. eTable 4. The threshold point and the corresponding results of the model tested in the validation set.
7. eTable5. The DLS performance tested on the REFUGE challenge dataset
8. eTable6 The comparison between human doctors and the DLS model in the validation set.
 - 8.1 eTable6-1 The diagnostic sensitivity and specificity of ophthalmic resident 1 and the DLS model in the subset of the external test set B.
 - 8.2 eTable6-2 The diagnostic sensitivity and specificity of ophthalmic resident 2 and the DLS model in the subset of the external test set B.
 - 8.3 eTable6-3 The diagnostic sensitivity and specificity of ophthalmic resident 3 and the DLS model in the subset of the external test set B
 - 8.4 eTable6-4 The diagnostic sensitivity and specificity of ophthalmic resident 4 and the DLS model in the subset of the external test set B.
9. eFigures and legends
 - 9.1 eFigure1
 - 9.2 eFigure2
 - 9.3 eFigure3
 - 9.4 eFigure4
 - 9.5 eFigure5
 - 9.6 eFigure6
8. References

1. The inclusion and exclusion criteria of the datasets

The inclusion criterion of the development dataset and internal test set included: ①the initial diagnosis were normal fundus and 12 selected fundus diseases with a standard diagnostic criteria (eTable1); ②for each patient, bilateral fundus images could be enrolled but only one for each eye; ③the CFPs should be posterior fundus photograph centralized by the macular fovea and contain the whole area of the optic disc. The exclusion criteria included: ①images identified with insufficient quality by the doctors; ②images annotated without consistent labels at the annotation stage.

For the CFPs in the external test set, the original diagnosis was not limited considering the real clinical condition. Other inclusion and exclusion criteria remained the same.

2. The development and selection of the deep learning system.

We first selected four state-of-the-art convolutional neural network (CNN) architectures as candidate multilabel classification models: Inception-V3, ResNet101, DenseNet121 and SeResNext50. They were all pretrained in ImageNet Datasets. To meet the joint requirement of abnormal versus normal classification and fine-grained recognition of multiple diseases, we modified the feature layer of the four CNNs to adapt to a certain situation. Two different branches were applied, with one branch for normal and abnormal classification and the other branch for diseases annotation (eFigure1). The binary classification models were trained with only one branch in the feature layer to predict disease or non-disease. The prediction of the late-fusion multilabel model is based on the average predictive value of 3 multilabel models. The detail of the cross entropy of the models, the postprocessing and conflict resolutions are available in the online materials.

The input size of the images was 512x512. Through several convolution layers and blocks, we obtained a feature vector. The cross-entropy loss was applied for the first branch since predicting abnormalities from normal images is a binary classification problem. The definition of cross entropy loss is as follows:

$$Loss_{cross-entropy} = -[y \cdot \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y})]$$

y is the target label of an image, and \hat{y} is the predicted score of the model. We utilized the multi-label loss function logits binary cross entropy for the second branch, and the definition of this loss is as follows:

$$Loss_{BCE} = - \left[\sum_{i=1}^N y_i \cdot \log \delta(\hat{y}_i) + \sum_{i=1}^N (1 - y_i) \cdot \log \delta(\hat{y}_i) \right]$$

N is the total number of disorders, y_i and \hat{y}_i are the target label and predict score of the model, respectively, and $\delta(\cdot)$ is the sigmoid function to transform the predict score from infinity to (0.1). Then, we involved weights of the 2 branches as the hyperparameter to combine two losses together:

$$Loss = w_1 * Loss_{cross-entropy} + w_2 * Loss_{BCE}$$

w_1 and w_2 are the weights of the first and second branches, respectively. This reconstruction makes it possible to combine a binary classification system with multi-label recognition into a single network.

The model was validated in the validation set for model selection and hyperparameter optimization, and was trained using the remaining images in the training set. We utilized stochastic gradient descent (SGD) optimizer to train the model with an initial learning rate of 0.001, a momentum of 0.9, and a weight decay of 10^{-4} . We used the initial learning rate to train 5 epoch. If the mean average precision in the validation doesn't increase in two consecutives then the learning rate was halved. The Batch size was set to 10 and the following hyperparameters: $w_1=0.5$ and $w_2=1$. We obtained the model with the best mean average precision score in the validation set as the final model.

The sensitivity and specificity were calculated by the decision threshold, which was selected when the model obtained the optimal harmonic average of the two indexes in the validation set. Since the two branches of the task may provide conflict results like predicting “normal fundus” and “Possible GON” together if they both reach the threshold in two branches. To solve this kind of conflict, we designed a conversion method to unify the predictive values of the 12 categories based on the probability. We divided [0.1] into 100 equal intervals and analyze the possibility distribution of positive images in each category. Then the predictive value of the testing image was introduced into the calculation to get a new predictive value. Considering the screening application scenarios, we set that the label of “normal fundus” would be provided as the predictive label only when the predictive value was the highest of all the predicted labels.

3. eTable1. The standard criteria of the 12 selected fundus diseases

Diseases	Standard diagnostic criteria
Referable diabetic retinopathy	①The patients should have clear medical history of diabetes. ②The fundus images presented a diabetic retinopathy severity level of moderate nonproliferative diabetic retinopathy or worse, diabetic macular edema, and/or ungradable image ¹
Retinal vein occlusion	The fundus appearance presented with flame-shaped hemorrhage, dilation of the involved veins, cotton-wool spots, with or without papilledema ²
Pathologic myopic retinal degeneration maculopathy	①The myopic diopter should be over -6.0DS. ②The fundus images presented myopic maculopathy lesions in category 2-4 according to the META-PM (meta analyses of pathologic myopia) study classification ⁴
Retinitis pigmentosa	Fundus appearance with peripheral bone spicule pigmentations and thin retinal arteries ⁵
Retinal detachment	The elevation of the sensory retina presented as translucent membranoid structure with vessels on. This cartogory included three major types of RD: rhegmatogenous, tractional and exudative.
Epiretinal membrane	A thin glistening membrane over the macula with or without retinal wrinkling ⁶
Dry age-related macular degeneration	Early, intermediate AMD ⁷ and geographic atrophy
Wet age-related macular degeneration	Neovascular AMD which belongs to the late stage of AMD ⁷
Macular hole	Stage 2-4 according to Gass's classification ⁸ of macular hole.
Possible glaucomatous optic neuropathy	With cup disc ratio greater than 0.7 with or without corresponding retinal fiber layer defect
Papilledema	Swollen and elevation of the optic disc, blurred disc range, with or without hemorrhage and retinal vein dilation ⁹
Optic nerve atrophy	Pallor of the optic nerve ⁹

4. eTable2. The four multilabel models' performance using different convolutional neural networks tested in the validation set

4.1 eTable2-1. The area under the curve (AUC) results

	Area under the curve (AUC) values (95% confidence interval)			
	SeResNext50	ResNet101	Inception-V3	DenseNet121
Normal fundus	0.987 (0.983, 0.991)	0.984 (0.980, 0.988)	0.982 (0.978, 0.986)	0.987 (0.983, 0.991)
Retinal vein occlusion	0.941 (0.933, 0.948)	0.939 (0.931, 0.947)	0.916 (0.907, 0.925)	0.946 (0.938, 0.953)
Referable diabetic retinopathy	0.990 (0.987, 0.993)	0.991 (0.988, 0.994)	0.989 (0.986, 0.993)	0.988 (0.985, 0.992)
Pathological myopic retinal degeneration	0.985 (0.981, 0.989)	0.985 (0.981, 0.989)	0.986 (0.982, 0.989)	0.989 (0.986, 0.993)
Retinitis pigmentosa	0.996 (0.994, 0.998)	0.998 (0.996, 0.999)	0.989 (0.986, 0.993)	0.997 (0.995, 0.999)
Retinal detachment	0.995 (0.993, 0.998)	0.991 (0.988, 0.994)	0.970 (0.965, 0.976)	0.995 (0.993, 0.997)
Epiretinal membrane	0.960 (0.954, 0.967)	0.971 (0.965, 0.976)	0.950 (0.943, 0.957)	0.969 (0.964, 0.975)
Dry age-related macular degeneration	0.966 (0.960, 0.972)	0.963 (0.957, 0.970)	0.937 (0.929, 0.945)	0.968 (0.963, 0.974)
Wet age-related macular degeneration	0.950 (0.943, 0.957)	0.963 (0.957, 0.970)	0.925 (0.916, 0.934)	0.956 (0.950, 0.963)
Macular hole	0.969 (0.963, 0.975)	0.958 (0.952, 0.965)	0.959 (0.953, 0.965)	0.967 (0.961, 0.973)
Possible glaucomatous optic neuropathy	0.941 (0.934, 0.949)	0.950 (0.942, 0.957)	0.943 (0.935, 0.950)	0.948 (0.941, 0.955)
Papilledema	0.971 (0.965, 0.976)	0.977 (0.972, 0.982)	0.981 (0.976, 0.985)	0.979 (0.974, 0.984)
Optic nerve atrophy	0.985 (0.981, 0.989)	0.985 (0.981, 0.989)	0.986 (0.983, 0.990)	0.988 (0.984, 0.991)

4.2 eTable2-2. The average precision (AP) results

	Average precision values (95% confidence interval)			
	SeResNext50	ResNet101	Inception-V3	DenseNet121
Normal fundus	0.961 (0.954, 0.967)	0.946 (0.939, 0.954)	0.938 (0.930, 0.946)	0.954 (0.947, 0.960)
Retinal vein occlusion	0.747 (0.732, 0.761)	0.758 (0.744, 0.772)	0.695 (0.680, 0.710)	0.796 (0.783, 0.809)
Referable diabetic retinopathy	0.962 (0.956, 0.968)	0.944 (0.937, 0.952)	0.935 (0.927, 0.943)	0.918 (0.909, 0.927)
Pathological myopic retinal degeneration	0.887 (0.877, 0.898)	0.893 (0.883, 0.903)	0.874 (0.864, 0.885)	0.907 (0.897, 0.916)
Retinitis pigmentosa	0.967 (0.961, 0.973)	0.967 (0.961, 0.972)	0.918 (0.909, 0.926)	0.969 (0.964, 0.975)
Retinal detachment	0.954 (0.947, 0.961)	0.921 (0.912, 0.930)	0.880 (0.870, 0.891)	0.944 (0.937, 0.952)
Epiretinal membrane	0.862 (0.850, 0.873)	0.873 (0.862, 0.884)	0.838 (0.826, 0.850)	0.880 (0.870, 0.891)
Dry age-related macular degeneration	0.780 (0.767, 0.794)	0.723 (0.709, 0.738)	0.645 (0.629, 0.660)	0.756 (0.742, 0.770)
Wet age-related macular degeneration	0.764 (0.750, 0.778)	0.798 (0.785, 0.811)	0.723 (0.709, 0.738)	0.777 (0.764, 0.791)
Macular hole	0.842 (0.830, 0.854)	0.799 (0.786, 0.812)	0.784 (0.771, 0.798)	0.816 (0.803, 0.828)
Possible glaucomatous optic neuropathy	0.710 (0.696, 0.725)	0.704 (0.689, 0.719)	0.662 (0.647, 0.677)	0.680 (0.665, 0.696)
Papilledema	0.869 (0.858, 0.880)	0.889 (0.879, 0.899)	0.859 (0.847, 0.870)	0.896 (0.886, 0.905)
Optic nerve atrophy	0.851 (0.839, 0.862)	0.854 (0.842, 0.865)	0.869 (0.858, 0.880)	0.868 (0.857, 0.879)

4.3 eTable2-3. The sensitivity results

	Sensitivity values (95% confidence interval)			
	SeResNext50	ResNet101	Inception-V3	DenseNet121
Normal fundus	0.932 (0.924, 0.940)	0.933 (0.925, 0.941)	0.924 (0.915, 0.932)	0.943 (0.935, 0.951)
Retinal vein occlusion	0.793 (0.780, 0.806)	0.761 (0.748, 0.775)	0.758 (0.744, 0.772)	0.811 (0.798, 0.823)
Referable diabetic retinopathy	0.960 (0.954, 0.967)	0.964 (0.958, 0.970)	0.942 (0.934, 0.949)	0.947 (0.940, 0.955)
Pathological myopic retinal degeneration	0.937 (0.930, 0.945)	0.953 (0.946, 0.960)	0.922 (0.913, 0.931)	0.953 (0.946, 0.960)
Retinitis pigmentosa	0.954 (0.947, 0.961)	0.969 (0.964, 0.975)	0.954 (0.947, 0.961)	0.962 (0.955, 0.968)
Retinal detachment	0.936 (0.928, 0.944)	0.918 (0.909, 0.927)	0.864 (0.852, 0.875)	0.927 (0.919, 0.936)
Epiretinal membrane	0.862 (0.851, 0.873)	0.884 (0.874, 0.895)	0.869 (0.858, 0.880)	0.922 (0.913, 0.930)
Dry age-related macular degeneration	0.861 (0.850, 0.873)	0.816 (0.804, 0.829)	0.764 (0.750, 0.778)	0.839 (0.827, 0.851)
Wet age-related macular degeneration	0.863 (0.852, 0.874)	0.849 (0.838, 0.861)	0.842 (0.831, 0.854)	0.822 (0.809, 0.834)
Macular hole	0.832 (0.820, 0.844)	0.745 (0.730, 0.759)	0.803 (0.790, 0.816)	0.788 (0.775, 0.802)
Possible glaucomatous optic neuropathy	0.719 (0.704, 0.733)	0.793 (0.779, 0.806)	0.756 (0.742, 0.770)	0.793 (0.779, 0.806)
Papilledema	0.886 (0.876, 0.896)	0.886 (0.876, 0.896)	0.890 (0.880, 0.901)	0.846 (0.835, 0.858)
Optic nerve atrophy	0.911 (0.902, 0.920)	0.926 (0.917, 0.934)	0.876 (0.866, 0.887)	0.931 (0.922, 0.939)

4.4 eTable2-4. The specificity results

	Specificity values (95% confidence interval)			
	SeResNext50	ResNet101	Inception-V3	DenseNet121
Normal fundus	0.977 (0.972, 0.982)	0.967 (0.961, 0.973)	0.965 (0.959, 0.971)	0.966 (0.960, 0.972)
Retinal vein occlusion	0.889 (0.879, 0.900)	0.892 (0.882, 0.902)	0.877 (0.867, 0.888)	0.906 (0.897, 0.916)
Referable diabetic retinopathy	0.951 (0.944, 0.958)	0.959 (0.953, 0.966)	0.963 (0.957, 0.969)	0.961 (0.954, 0.967)
Pathological myopic retinal degeneration	0.976 (0.971, 0.981)	0.958 (0.952, 0.965)	0.969 (0.963, 0.975)	0.963 (0.957, 0.969)
Retinitis pigmentosa	0.982 (0.978, 0.987)	0.981 (0.977, 0.986)	0.976 (0.971, 0.981)	0.985 (0.981, 0.989)
Retinal detachment	0.990 (0.986, 0.993)	0.987 (0.984, 0.991)	0.988 (0.984, 0.991)	0.978 (0.973, 0.983)
Epiretinal membrane	0.950 (0.943, 0.957)	0.940 (0.932, 0.948)	0.920 (0.911, 0.929)	0.935 (0.927, 0.943)
Dry age-related macular degeneration	0.921 (0.912, 0.929)	0.934 (0.925, 0.942)	0.921 (0.913, 0.930)	0.911 (0.902, 0.920)
Wet age-related macular degeneration	0.918 (0.909, 0.927)	0.942 (0.935, 0.950)	0.924 (0.915, 0.932)	0.951 (0.944, 0.958)
Macular hole	0.962 (0.956, 0.968)	0.986 (0.982, 0.990)	0.969 (0.963, 0.974)	0.968 (0.963, 0.974)
Possible glaucomatous optic neuropathy	0.940 (0.932, 0.947)	0.913 (0.904, 0.922)	0.919 (0.910, 0.928)	0.923 (0.915, 0.932)
Papilledema	0.938 (0.930, 0.946)	0.949 (0.942, 0.956)	0.927 (0.918, 0.935)	0.955 (0.948, 0.962)
Optic nerve atrophy	0.953 (0.946, 0.960)	0.959 (0.952, 0.965)	0.939 (0.931, 0.947)	0.948 (0.940, 0.955)

5. eTable3. The comparison of the selected multilabel model, binary classification models and the late-fusion multilabel model tested in the validation set

5.1 eTable3-1. The area under the curve (AUC) results

	Area under the curve (AUC) values (95% confidence interval)		
	Model A	Model B	Model C
Normal fundus	0.987 (0.983, 0.991)	0.988 (0.983, 0.994)	0.989 (0.985, 0.992)
Retinal vein occlusion	0.941 (0.933, 0.948)	0.957 (0.946, 0.969)	0.950 (0.942, 0.957)
Referable diabetic retinopathy	0.990 (0.987, 0.993)	0.996 (0.993, 1.000)	0.994 (0.992, 0.997)
Pathological myopic retinal degeneration	0.985 (0.981, 0.989)	0.989 (0.983, 0.995)	0.988 (0.984, 0.991)
Retinitis pigmentosa	0.996 (0.994, 0.998)	0.992 (0.987, 0.997)	0.996 (0.994, 0.998)
Retinal detachment	0.995 (0.993, 0.998)	0.998 (0.996, 1.000)	0.996 (0.993, 0.998)
Epiretinal membrane	0.960 (0.954, 0.967)	0.974 (0.965, 0.983)	0.968 (0.963, 0.974)
Dry age-related macular degeneration	0.966 (0.960, 0.972)	0.954 (0.942, 0.966)	0.976 (0.971, 0.981)
Wet age-related macular degeneration	0.950 (0.943, 0.957)	0.984 (0.978, 0.991)	0.964 (0.958, 0.970)
Macular hole	0.969 (0.963, 0.975)	0.963 (0.952, 0.973)	0.978 (0.973, 0.983)
Possible glaucomatous optic neuropathy	0.941 (0.934, 0.949)	0.959 (0.948, 0.970)	0.953 (0.946, 0.960)
Papilledema	0.971 (0.965, 0.976)	0.962 (0.951, 0.973)	0.980 (0.975, 0.985)
Optic nerve atrophy	0.985 (0.981, 0.989)	0.978 (0.970, 0.986)	0.989 (0.985, 0.992)
Model A=the multilabel model; Model B=the combination of binary classification models; Model C=the late fusion model			

5.2 eTable3-2. The average precision (AP) results

	Average precision (AP) value (95% confidence interval)		
	Model A	Model B	Model C
Normal fundus	0.961 (0.954, 0.967)	0.975 (0.966, 0.984)	0.969 (0.964, 0.975)
Retinal vein occlusion	0.747 (0.732, 0.761)	0.775 (0.752, 0.799)	0.809 (0.797, 0.822)
Referable diabetic retinopathy	0.962 (0.956, 0.968)	0.984 (0.977, 0.991)	0.966 (0.960, 0.972)
Pathological myopic retinal degeneration	0.887 (0.877, 0.898)	0.889 (0.871, 0.906)	0.895 (0.885, 0.905)
Retinitis pigmentosa	0.967 (0.961, 0.973)	0.970 (0.961, 0.980)	0.971 (0.965, 0.976)
Retinal detachment	0.954 (0.947, 0.961)	0.968 (0.958, 0.977)	0.965 (0.959, 0.971)
Epiretinal membrane	0.862 (0.850, 0.873)	0.822 (0.801, 0.844)	0.884 (0.874, 0.895)
Dry age-related macular degeneration	0.780 (0.767, 0.794)	0.753 (0.729, 0.777)	0.811 (0.798, 0.824)
Wet age-related macular degeneration	0.764 (0.750, 0.778)	0.815 (0.794, 0.837)	0.810 (0.797, 0.823)
Macular hole	0.842 (0.830, 0.854)	0.806 (0.784, 0.828)	0.871 (0.860, 0.882)
Possible glaucomatous optic neuropathy	0.710 (0.696, 0.725)	0.714 (0.689, 0.739)	0.742 (0.728, 0.756)
Papilledema	0.869 (0.858, 0.880)	0.851 (0.832, 0.871)	0.901 (0.892, 0.911)
Optic nerve atrophy	0.851 (0.839, 0.862)	0.819 (0.797, 0.840)	0.873 (0.862, 0.884)
Model A=the multilabel model; Model B=the combination of binary classification models; Model C=the late fusion model			

5.3 eTable3-3. The sensitivity results

	Sensitivity Value (95% confidence interval)		
	Model A	Model B	Model C
Normal fundus	0.932 (0.924, 0.940)	0.964 (0.954, 0.974)	0.945 (0.938, 0.953)
Retinal vein occlusion	0.793 (0.780, 0.806)	0.822 (0.801, 0.843)	0.804 (0.791, 0.816)
Referable diabetic retinopathy	0.960 (0.954, 0.967)	0.974 (0.965, 0.983)	0.964 (0.958, 0.970)
Pathological myopic retinal degeneration	0.937 (0.930, 0.945)	0.949 (0.936, 0.961)	0.958 (0.952, 0.965)
Retinitis pigmentosa	0.954 (0.947, 0.961)	0.976 (0.967, 0.984)	0.962 (0.955, 0.968)
Retinal detachment	0.936 (0.928, 0.944)	0.956 (0.945, 0.967)	0.973 (0.967, 0.978)
Epiretinal membrane	0.862 (0.851, 0.873)	0.869 (0.850, 0.888)	0.918 (0.909, 0.927)
Dry age-related macular degeneration	0.861 (0.850, 0.873)	0.825 (0.804, 0.846)	0.858 (0.846, 0.869)
Wet age-related macular degeneration	0.863 (0.852, 0.874)	0.917 (0.901, 0.932)	0.842 (0.831, 0.854)
Macular hole	0.832 (0.820, 0.844)	0.783 (0.760, 0.806)	0.876 (0.865, 0.887)
Possible glaucomatous optic neuropathy	0.719 (0.704, 0.733)	0.720 (0.695, 0.745)	0.804 (0.791, 0.817)
Papilledema	0.886 (0.876, 0.896)	0.818 (0.797, 0.840)	0.904 (0.894, 0.913)
Optic nerve atrophy	0.911 (0.902, 0.920)	0.876 (0.858, 0.894)	0.950 (0.943, 0.958)
Model A=the multilabel model; Model B=the combination of binary classification models; Model C=the late fusion model			

5.4 eTable3-4. The specificity results

	Specificity value (95% confidence interval)		
	Model A	Model B	Model C
Normal fundus	0.977 (0.972, 0.982)	0.941 (0.928, 0.954)	0.967 (0.961, 0.973)
Retinal vein occlusion	0.889 (0.879, 0.900)	0.856 (0.837, 0.876)	0.897 (0.888, 0.907)
Referable diabetic retinopathy	0.951 (0.944, 0.958)	0.967 (0.958, 0.977)	0.969 (0.963, 0.974)
Pathological myopic retinal degeneration	0.976 (0.971, 0.981)	0.964 (0.953, 0.974)	0.971 (0.965, 0.976)
Retinitis pigmentosa	0.982 (0.978, 0.987)	0.973 (0.963, 0.982)	0.978 (0.973, 0.983)
Retinal detachment	0.990 (0.986, 0.993)	0.991 (0.986, 0.996)	0.981 (0.977, 0.986)
Epiretinal membrane	0.950 (0.943, 0.957)	0.945 (0.932, 0.957)	0.923 (0.915, 0.932)
Dry age-related macular degeneration	0.921 (0.912, 0.929)	0.905 (0.888, 0.921)	0.939 (0.931, 0.947)
Wet age-related macular degeneration	0.918 (0.909, 0.927)	0.938 (0.925, 0.951)	0.953 (0.946, 0.960)
Macular hole	0.962 (0.956, 0.968)	0.982 (0.974, 0.989)	0.963 (0.957, 0.970)
Possible glaucomatous optic neuropathy	0.940 (0.932, 0.947)	0.927 (0.912, 0.941)	0.934 (0.925, 0.942)
Papilledema	0.938 (0.930, 0.946)	0.961 (0.950, 0.971)	0.950 (0.943, 0.957)
Optic nerve atrophy	0.977 (0.972, 0.982)	0.941 (0.928, 0.954)	0.967 (0.961, 0.973)
Model A=the multilabel model; Model B=the combination of binary classification models; Model C=the late fusion model			

6. eTable 4. The threshold point and the corresponding results of the model tested in the validation set.

	Validation set			
	Sensitivity	Specificity	AP	AUC
Normal fundus	0.891 (0.885, 0.897)	0.913 (0.907, 0.918)	0.936 (0.932, 0.941)	0.960 (0.956, 0.964)
Referable diabetic retinopathy	0.688 (0.680, 0.697)	0.923 (0.918, 0.928)	0.651 (0.642, 0.660)	0.904 (0.899, 0.910)
Retinal vein occlusion	0.965 (0.961, 0.968)	0.979 (0.976, 0.981)	0.963 (0.959, 0.967)	0.993 (0.992, 0.995)
Pathological myopic retinal degeneration	0.941 (0.937, 0.946)	0.973 (0.970, 0.976)	0.948 (0.944, 0.952)	0.987 (0.984, 0.989)
Retinitis pigmentosa	0.949 (0.945, 0.953)	0.986 (0.984, 0.988)	0.914 (0.909, 0.919)	0.992 (0.990, 0.993)
Retinal detachment	0.898 (0.892, 0.904)	0.989 (0.987, 0.991)	0.861 (0.854, 0.867)	0.985 (0.983, 0.988)
Epiretinal membrane	0.862 (0.855, 0.869)	0.943 (0.939, 0.948)	0.838 (0.830, 0.845)	0.975 (0.972, 0.978)
Dry age-related macular degeneration	0.752 (0.744, 0.761)	0.919 (0.914, 0.925)	0.612 (0.603, 0.621)	0.932 (0.927, 0.936)
Wet age-related macular degeneration	0.894 (0.888, 0.900)	0.948 (0.944, 0.952)	0.750 (0.742, 0.759)	0.975 (0.972, 0.978)
Macular hole	0.593 (0.584, 0.603)	0.979 (0.976, 0.982)	0.401 (0.392, 0.411)	0.920 (0.914, 0.925)
Possible glaucomatous optic neuropathy	0.664 (0.654, 0.673)	0.892 (0.886, 0.898)	0.413 (0.404, 0.423)	0.912 (0.907, 0.918)
Papilledema	0.818 (0.811, 0.826)	0.969 (0.965, 0.972)	0.862 (0.855, 0.869)	0.978 (0.975, 0.981)
Optic nerve atrophy	0.803 (0.795, 0.811)	0.939 (0.934, 0.944)	0.662 (0.653, 0.671)	0.972 (0.969, 0.975)

7. eTable5 Comparing our DLS with the state-of-art for glaucoma detection on the REFUGE challenge dataset. Performance scores of the individual teams are cited from the REFUGE challenge paper

Rank	Team	AUC	Reference sensitivity
1	VRT	0.9885	0.9752
2	SDSAIRC	0.9817	0.9760
3	CUHKMED	0.9644	0.9500
4	NKSG	0.9587	0.8917
5	Mammoth	0.9555	0.8918
6	Our DLS	0.9546	0.9305
7	Masker	0.9524	0.8500
8	SMILEDeepDR	0.9508	0.8750
9	BUCT	0.9348	0.8500
10	WinterFell	0.9327	0.9250
11	NightOwl	0.9101	0.9000
12	Cvblab	0.8806	0.7318
13	AIML	0.8458	0.7250
Ground truth	vCDR	0.9471	0.8750

8 eTable6 The comparison between human doctors and the DLS model in the validation set.

8.1 eTable6-1 The diagnostic sensitivity and specificity of ophthalmic resident 1 and the DLS model in the subset of the external test set B.

	No	Sensitivity		Specificity	
		Doctor	DLS model	Doctor	DLS model
Normal fundus	647	0.958	0.923	0.827	0.957
Retinal vein occlusion	39	0.769	0.949	0.994	0.958
Referable diabetic retinopathy	21	0.810	1.000	0.996	0.990
Pathological myopic retinal degeneration	16	0.500	1.000	0.996	0.982
Retinitis pigmentosa	15	0.467	0.867	0.999	0.994
Retinal detachment	2	1.000	1.000	1.000	0.992
Epiretinal membrane	21	0.762	0.857	0.982	0.971
Dry age-related macular degeneration	25	0.560	0.800	0.993	0.977
Wet age-related macular degeneration	15	0.867	1.000	0.999	0.994
Macular hole	4	0.750	1.000	1.000	0.981
Possible glaucomatous optic neuropathy	44	0.568	0.909	0.981	0.969
Papilledema	10	0.800	0.800	0.993	0.982
Optic nerve atrophy	13	0.231	0.615	0.995	0.948
Mean		0.695	0.902	0.981	0.976
Mann-Whitney U test		<i>P=0.005</i>		<i>P=0.003</i>	

8.2 eTable6-2 The diagnostic sensitivity and specificity of ophthalmic resident 2 and the DLS model in the subset of the external test set B.

	No	Sensitivity		Specificity	
		Doctor	DLS model	Doctor	DLS model
Normal fundus	340	0.876	0.844	0.877	0.850
Retinal vein occlusion	110	0.745	0.927	0.980	0.851
Referable diabetic retinopathy	28	0.857	1.000	0.994	0.976
Pathological myopic retinal degeneration	34	0.735	1.000	0.997	0.910
Retinitis pigmentosa	5	0.800	1.000	1.000	0.961
Retinal detachment	6	0.833	0.833	1.000	0.980
Epiretinal membrane	45	0.533	0.711	0.978	0.855
Dry age-related macular degeneration	145	0.641	0.697	0.973	0.932
Wet age-related macular degeneration	23	0.652	0.957	0.991	0.960
Macular hole	2	1.000	1.000	0.999	0.945
Possible glaucomatous optic neuropathy	78	0.744	0.846	0.958	0.914
Papilledema	27	0.704	0.778	0.979	0.967
Optic nerve atrophy	43	0.721	0.698	0.986	0.936
Mean		0.757	0.868	0.978	0.926
Mann-Whitney U test		P=0.057		P <0.001	

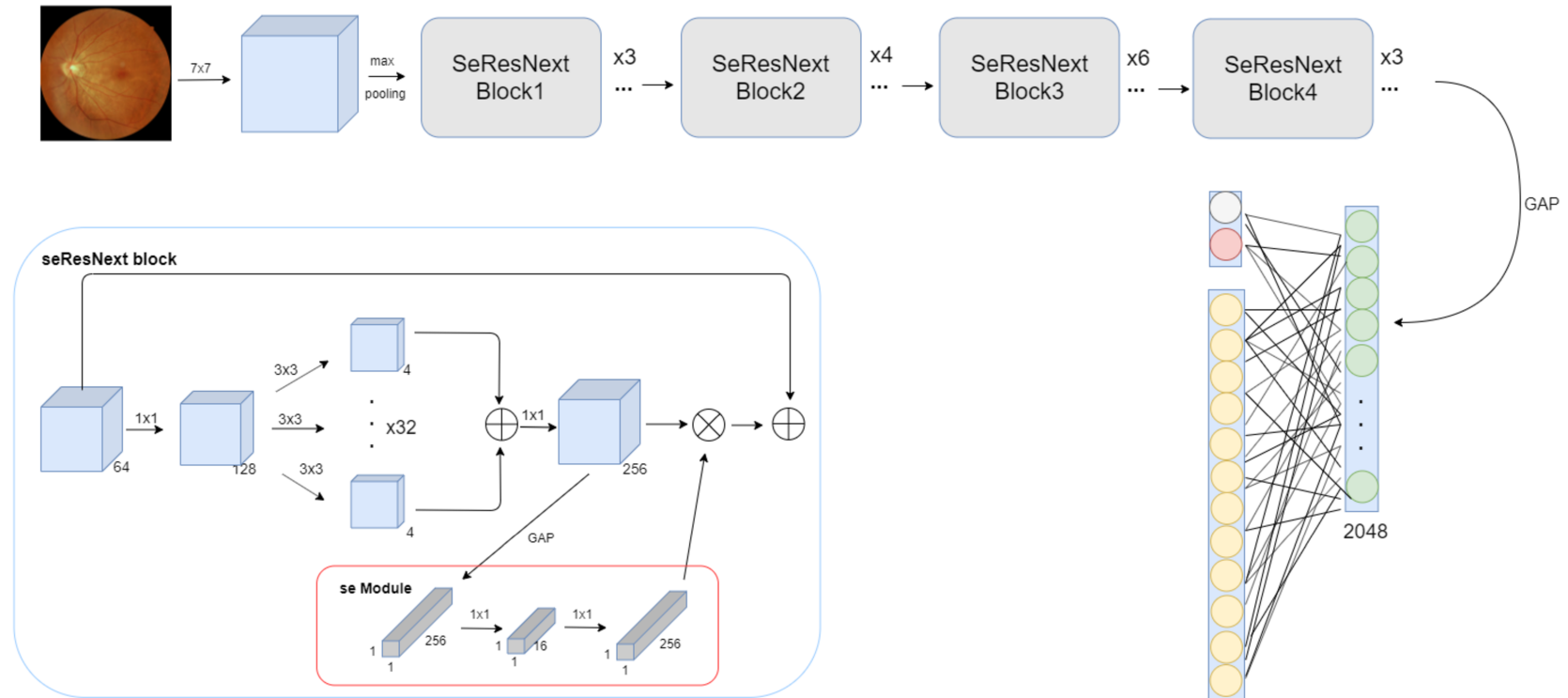
8.3 eTable6-3 The diagnostic sensitivity and specificity of ophthalmic resident 3 and the DLS model in the subset of the external test set B.

	No	Sensitivity		Specificity	
		Doctor	DLS model	Doctor	DLS model
Normal fundus	382	0.809	0.846	0.905	0.880
Retinal vein occlusion	113	0.681	0.885	0.990	0.847
Referable diabetic retinopathy	37	0.865	1.000	0.995	0.991
Pathological myopic retinal degeneration	36	0.861	1.000	0.984	0.927
Retinitis pigmentosa	2	0.500	1.000	0.995	0.974
Retinal detachment	3	1.000	0.667	1.000	0.992
Epiretinal membrane	55	0.636	0.764	0.967	0.861
Dry age-related macular degeneration	137	0.613	0.745	0.972	0.934
Wet age-related macular degeneration	21	0.667	0.810	0.993	0.970
Macular hole	3	1.000	1.000	0.995	0.962
Possible glaucomatous optic neuropathy	57	0.702	0.737	0.947	0.911
Papilledema	19	0.737	0.789	0.988	0.981
Optic nerve atrophy	56	0.554	0.679	0.989	0.956
Mean		0.740	0.840	0.978	0.937
Mann-Whitney U test		P=0.091		P=0.009	

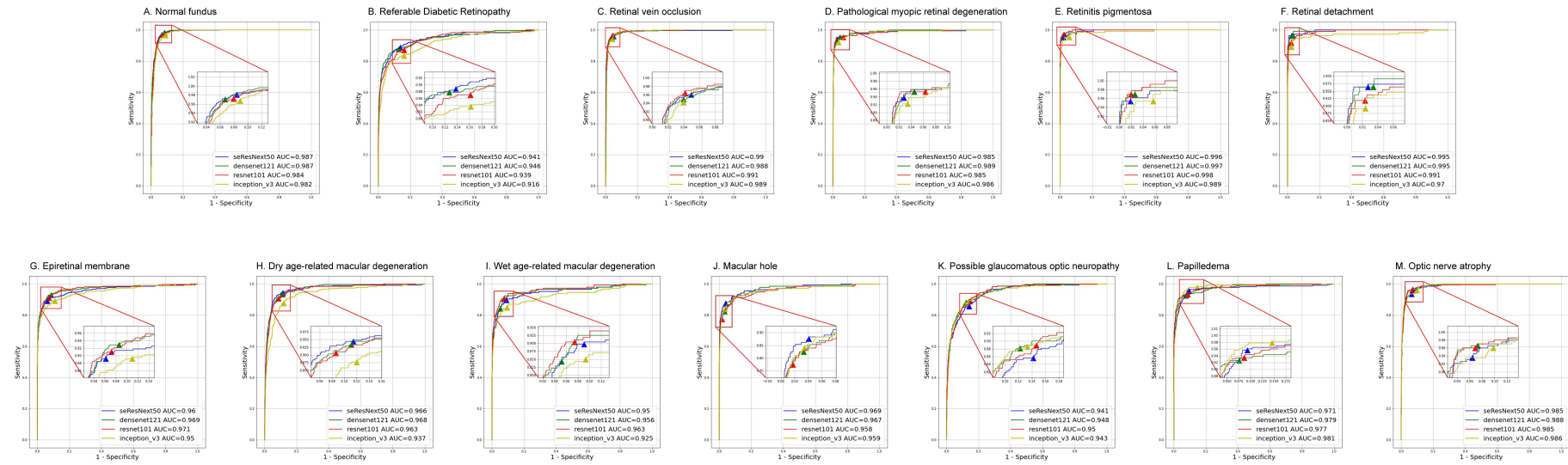
8.4 eTable6-4 The diagnostic sensitivity and specificity of ophthalmic resident 4 and the DLS model in the subset of the external test set B.

	No	Sensitivity		Specificity	
		Doctor	DLS model	Doctor	DLS model
Normal fundus	435	0.860	0.871	0.883	0.851
Retinal vein occlusion	126	0.754	0.944	0.969	0.858
Referable diabetic retinopathy	37	0.568	1.000	0.999	0.985
Pathological myopic retinal degeneration	27	0.778	0.963	0.991	0.933
Retinitis pigmentosa	16	0.500	0.875	0.998	0.979
Retinal detachment	3	0.667	0.667	0.999	0.986
Epiretinal membrane	44	0.659	0.705	0.971	0.866
Dry age-related macular degeneration	97	0.588	0.691	0.975	0.918
Wet age-related macular degeneration	16	0.875	0.937	0.987	0.961
Macular hole	5	1.000	1.000	1.000	0.974
Possible glaucomatous optic neuropathy	48	0.604	0.687	0.955	0.925
Papilledema	26	0.654	0.692	0.989	0.971
Optic nerve atrophy	38	0.737	0.684	0.973	0.967
Mean		0.711	0.824	0.976	0.936
Mann-Whitney U test		<i>P</i> < 0.001		<i>P</i> = 0.007	

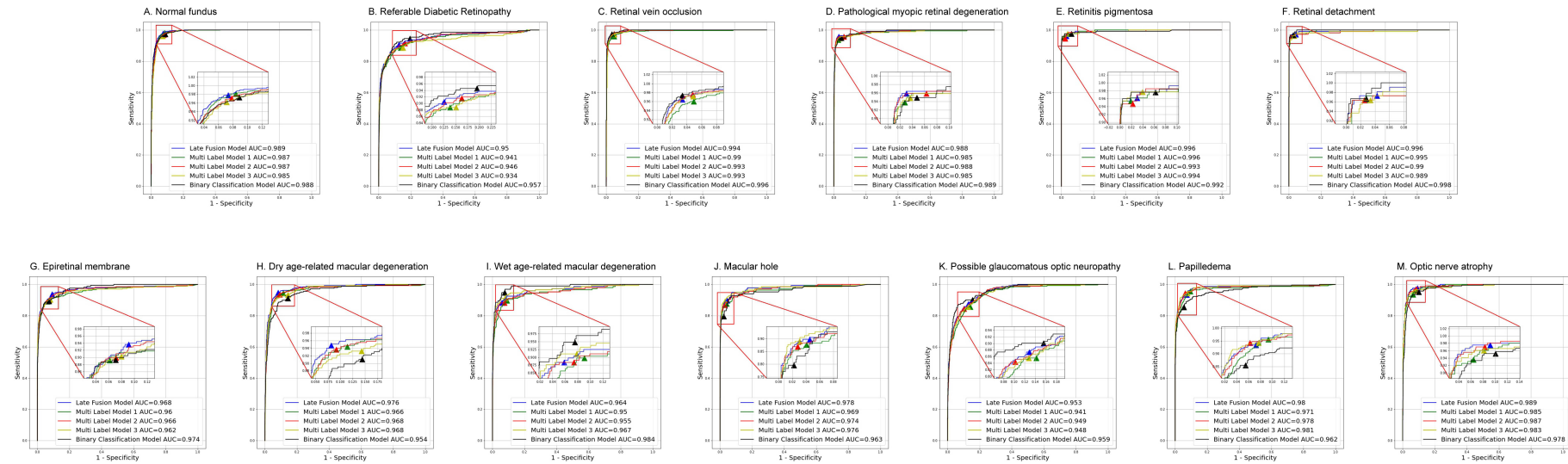
9. eFigures and legends



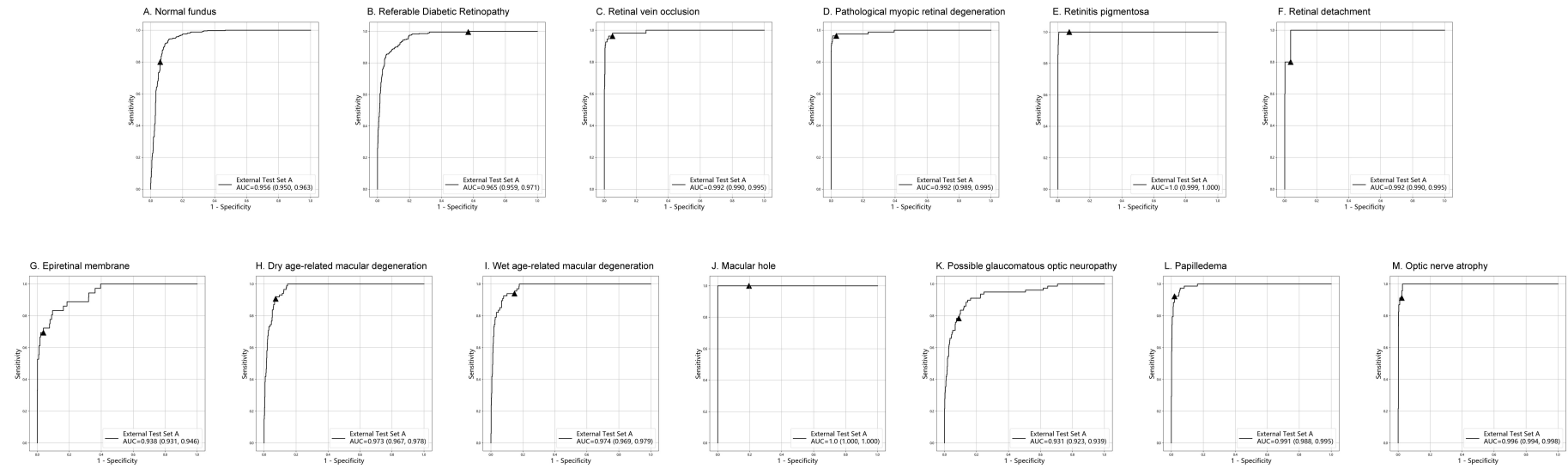
9.1 eFigure1 Convolutional neural network (CNN) architecture of the model. A parallel 2-branch network was applied for the task layer to distinguish abnormal and normal images in the first branch and to report diseases it predicted in the second branch.



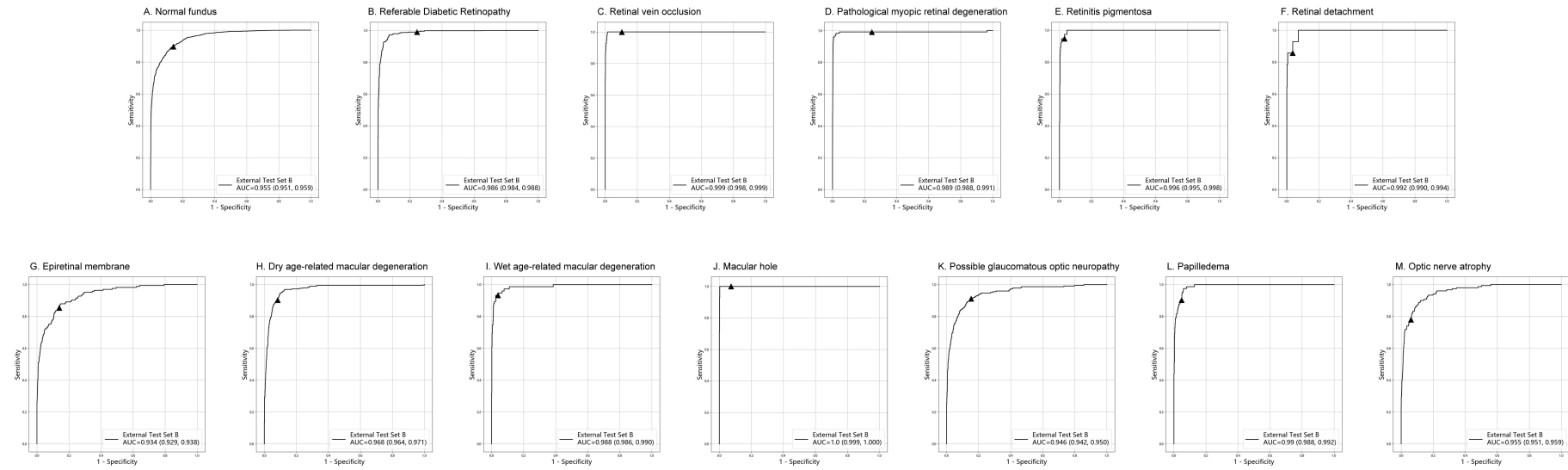
9.2 eFigure2. The receiver operating characteristic (ROC) curves of the four-candidate convolutional neural networks (CNNs) tested in the internal test set. Among all CNNs, SeResNext50 showed slightly better performance than others with the mean average precision reached 0.878.



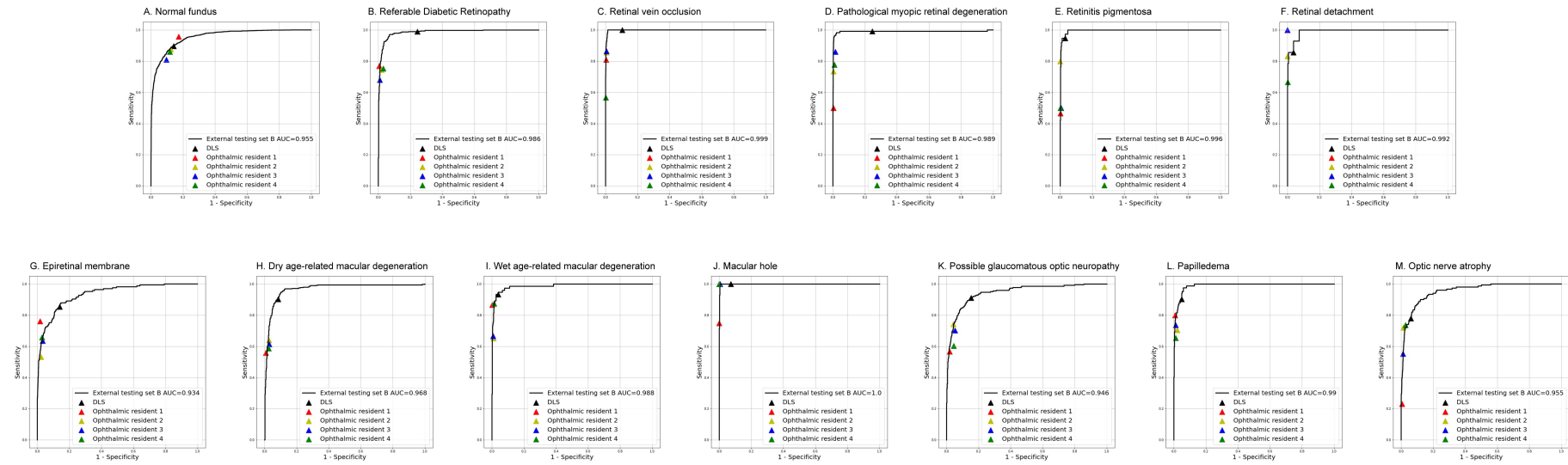
9.3 eFigure3 The receiver operating characteristic (ROC) curves of the single multilabel model, combination of the 13 binary classification models and the late-fusion multilabel model. The multilabel model showed comparable performance with the binary classification models. The late-fusion model presented more stable performance than the single multilabel model. It also showed the best performance with the mean average precision reached 0.889 in the validation set.



9.4 eFigure4. The ROC curved of the selected DLS tested on the internal test set A.



9.5 eFigure5. The ROC curved of the selected DLS tested on the internal test set B.



9.6 eFigure6 The receiver operating characteristic (ROC) curves and the results of the DLS and the diagnostic sensitivity and specificity of the four participate four ophthalmic residents.

10. References

1. 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. *JAMA* 2017,318(22):2211-2223.
2. Robinson MK and Halpern JI. Retinal vein occlusion. *Am Fam Physician* 1992,45(6):2661-2666.
3. Hayreh SS and Zimmerman MB. Fundus changes in central retinal artery occlusion. *Retina* 2007,27(3):276-289.
4. Ohno-Matsui K, Kawasaki R, Jonas JB, et al. International photographic classification and grading system for myopic maculopathy. *Am J Ophthalmol* 2015,159(5):877-883 e877.
5. Hartong DT, Berson EL and Dryja TP. Retinitis pigmentosa. *Lancet* 2006,368(9549):1795-1809.
6. Bu SC, Kuijer R, Li XR, Hooymans JM and Los LI. Idiopathic epiretinal membrane. *Retina* 2014,34(12):2317-2335.
7. Ferris FL, 3rd, Wilkinson CP, Bird A, et al. Clinical classification of age-related macular degeneration. *Ophthalmology* 2013,120(4):844-851.
8. Johnson RN and Gass JD. Idiopathic macular holes. Observations, stages of formation, and implications for surgical intervention. *Ophthalmology* 1988,95(7):917-924.
9. Biousse V and Newman NJ. Diagnosis and clinical features of common optic neuropathies. *Lancet Neurol* 2016,15(13):1355-1367.



Contents available at [ScienceDirect](https://www.sciencedirect.com)

Diabetes Research
and Clinical Practice

journal homepage: www.elsevier.com/locate/diabres



International
Diabetes
Federation



Efficacy of artificial intelligence-based screening for diabetic retinopathy in type 2 diabetes mellitus patients

Xiaoting Pei^a, Xi Yao^a, Yingrui Yang^a, Hongmei Zhang^b, Mengting Xia^a, Ranran Huang^a, Yuming Wang^c, Zhijie Li^{a,*}

^aHenan Eye Institute, Henan Eye Hospital, and Henan Key Laboratory of Ophthalmology and Visual Science, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, People's Hospital of Henan University, Zhengzhou, China

^bNursing Department, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, People's Hospital of Henan University, Zhengzhou, China

^cDepartments of Science and Technology Administration, Henan Provincial People's Hospital, Henan University People's Hospital, Zhengzhou University People's Hospital, Zhengzhou, China

ARTICLE INFO

Article history:

Received 4 November 2020

Received in revised form

14 August 2021

Accepted 4 January 2022

Available online 11 January 2022

Keywords:

Diabetic Retinopathy

Artificial Intelligence

Screening

Sensitivity

Specificity

ABSTRACT

Aim: To explore the efficacy of artificial intelligence (AI)-based screening for diabetic retinopathy (DR) in type 2 diabetes mellitus (T2DM) patients.

Methods: Data were obtained from 549 T2DM patients who visited the Fundus Disease Center at Henan Provincial People's Hospital from 2018/10–2020/09. DR identification and grading were conducted by two retina specialists, EyeWisdom®DSS and EyeWisdom®MCS, with ophthalmologist grading as reference standard, efficacy of EyeWisdom was evaluated according to sensitivity, specificity, positive predictive value, and negative predictive value. **Results:** Ophthalmologists detected 324 DR cases. Among them, there were 43 of mild non-proliferative DR (NPDR), 79 of moderate NPDR, 61 of severe NPDR, and 141 of proliferative DR (PDR). EyeWisdom®DSS detected 337 DR and EyeWisdom®MCS detected 264 DR. Sensitivity and specificity of EyeWisdom®DSS were 91.0%(95 %CI: 87.3%–93.8%) and 81.3% (95 %CI: 75.5%–86.1%), while EyeWisdom®MCS correctly identified 76.2%(95 %CI: 71.1%–80.7%) of patients with DR and 92.4%(95 %CI: 87.9%–95.4%) of patients without DR. EyeWisdom®DSS showed 76.5%(95 %CI: 69.6%–82.3%) sensitivity and 78.4%(95 %CI: 73.7%–82.5%) specificity for detecting NPDR and 64.5%(95 %CI: 56.0%–72.3%) sensitivity and 93.1%(95 %CI: 90.1%–95.3%) specificity for diagnosing PDR.

Conclusion: EyeWisdom®DSS is effective in screening for DR, and the accuracy of EyeWisdom®MCS was higher for identifying patients without DR. It is valuable to carry out AI-based DR screening in poorer regions.

© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

* Corresponding author. at: Henan Eye Institute & Henan Eye Hospital, Henan Provincial People's Hospital, Zhengzhou City 450000, China.

E-mail address: zhijielee@vip.163.com (Z. Li).

<https://doi.org/10.1016/j.diabres.2022.109190>

0168-8227/© 2022 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

According to data from the International Diabetes Federation, there were about 463 million diabetes mellitus (DM) patients worldwide in 2019, and it is predicted that this number will increase to 700 million by 2045 [1]. Diabetic retinopathy (DR) is one of the most common and serious complications of DM. It is the primary eye disease that causes blindness in the working population [2–4]. The reported prevalence of DR varies from 10% to 61% in people with DM in different countries [5]. However, research has shown that reasonable intervention and treatment of DR in the early stage can achieve good results in preventing the development of the disease and significantly reduce the blindness rate [6–8].

Unfortunately, people who live in counties, townships, villages, and marginal areas, often lack sufficient health knowledge. By the time DR is found in these individuals, it has often developed to a serious stage and caused irreversible visual impairment, which not only affects patients' quality of life but also increases the economic burden on society and the family, often leading to poverty in the latter case. Therefore, expanding the screening area of DR and carrying out targeted prevention and treatment of blindness can greatly reduce curable blindness. Regrettably, the number of ophthalmologists in China is insufficient, especially in primary medical institutions. According to data from the 21st National Ophthalmology Conference of the Chinese Medical Association held in 2016, 20% of county hospitals in China do not have an ophthalmology department. Even in institutions with such a department, there are few specialists in fundus diseases. The ratio of ophthalmologists to patients is 1:3000 in these areas [9], which is extremely unbalanced. Therefore, increasing the screening of DM patients to improve awareness, the treatment rate, and control rate of DR has become one of the main public health challenges in China [10–12].

Auxiliary measures based on artificial intelligence (AI) are efficient, cheap, and easy to operate for DR diagnosis [13]. EyeWisdom is an auxiliary diagnosis system for fundus diseases based on an AI algorithm developed by the company Zhiyuan Huitu (Vistel) in 2017. It mainly includes EyeWisdom®DSS software, a DR-specific auxiliary diagnosis system, and EyeWisdom®MCS software, a system with ophthalmic multi-disease screening as its core function [14]. EyeWisdom can screen for nearly 20 different eye diseases, such as DR, glaucoma, and age-related macular degeneration, based on the fundus photographs and disease history of subjects using an AI algorithm. It can not only directly provide suggestions for screening results, but it can also display nine typical DR lesions, such as microvascular tumor, retinal hemorrhage, hard exudation, and cotton wool spot, to help clinicians confirm the examination results. In addition, this fundus image analysis software is a cloud-based product, which can be used for real-time telemedicine combined with internet and 5G technology. It only takes 10 s from reading an image to outputting the results. The EyeWisdom AI algorithms have been trained in clinical practice and verified by retinal images obtained from the EyePACS database [15]. However, as yet, no report on the diagnostic and grading efficacy of

EyeWisdom®DSS and EyeWisdom®MCS in patients with DM has been found.

Therefore, this study collected the disease history and fundus photographs of DM patients at the Fundus Disease Center in Henan Provincial People's Hospital from 2018/10–2020/09. The diagnostic efficacy of the EyeWisdom®DSS and EyeWisdom®MCS systems were evaluated according to the sensitivity, specificity, area under the curve (AUC), as well as the positive predictive value (PPV) and negative predictive value (NPV), with an ophthalmologist's diagnosis as a reference standard. This study was designed to improve the awareness rate and treatment rates of DR and reduce the blindness rate in primary medical institutions.

2. Methods

2.1. Participants

DM patients at the Fundus Disease Center in Henan Provincial People's Hospital from 2018/10/01–2020/09/30 were invited to participate in this study. Type 2 diabetes mellitus (T2DM) patients ≥ 18 years old for whom fundus photographs could be obtained were included in this research. The exclusion criteria were as follows: (1) T2DM patients with missing key variables, such as age, gender, and disease history; (2) patients whose fundus photographs were not clear enough due to small pupils, cataracts, or vitreous opacity that prevented ophthalmologists from making a diagnosis; (3) patients who suffered from heart, liver, kidney, and/or other important organ failure; (4) and those with malignant tumors. Demographic characteristics, disease history, fundus photographs, and images from optical coherence tomography examination (if available) were collected, which was performed by two authors (M.X. and R.H.) independently and consistency check was conducted. The study was approved by the Ethics Committee of Henan Provincial People's Hospital (registration number 58/2017), and written informed consent was obtained from all participants.

The sample size was calculated using formula (1) below based on the diagnostic test [16], where $\alpha = 0.05$, allowable error $\delta = 0.08$, and p is the sensitivity or specificity of the method to be tested. According to our pre-analysis, the sensitivity and specificity were 90.0% and 80.0%, respectively, for the DR-specific system (EyeWisdom®DSS), and 75.0% and 92.0%, respectively, for the multi-disease system (EyeWisdom®MCS). The minimum sample size of the DR group was 113, and that of the non-DR group was 97. A margin of 20% was used for the sample size to account for any invalid samples. Therefore, the minimum sample size was 136 for the DR group and 117 for the non-DR group.

$$n = \left(\frac{Z_{\alpha}}{\delta} \right)^2 (1 - p)p \quad (1)$$

2.2. Acquisition of retinal images

Fundus examinations of DM patients were performed using the Zeiss non-mydratic fundus camera (VISUCAM 224, Germany) by an ophthalmologist according to the unified

standards. This camera does not require mydriasis before use and provides a 45° field of view for each eye. Five fields were captured in each eye: macula centered, temporal side, nasal side, and the upper and lower quadrant of the retina.

2.3. Definitions and diagnostic criteria

DM was determined according to the Standards of Medical Care in Diabetes set in 2018 by the American Diabetes Association [17]. The diagnosis and grading of DR were determined by two ophthalmologists (H.D. and D.Q.) with more than five years of work experience according to International Clinical Diabetic Retinopathy (ICDR) criteria [18,19]. DR was divided into five stages as follows: (1) absence of DR: no obvious retinopathy and no abnormality; (2) mild non-proliferative diabetic retinopathy (NPDR): the early stage of retinopathy with only microaneurysms; (3) moderate NPDR: some of the blood vessels that nourish the retinas are blocked; (4) severe NPDR: one or more of the following: (i) more than 20 intraretinal hemorrhages in each of the four quadrants of the retina, (ii) clear venous beading in two or more quadrants, and (iii) significant intraretinal microvascular abnormality in one or more quadrants; and (5) proliferative diabetic retinopathy (PDR): retinal signals triggering the growth of neovascularization in which the new blood vessels are abnormal and fragile.

The kappa (κ) agreement between the two ophthalmologists was 0.91. When the diagnosis or grading results were inconsistent, the fundus photographs were adjudicated by a third retinal specialist (D.W.), whose diagnosis was accepted as the final judgment for subsequent analysis. Any patient diagnosed with DR in both eyes was considered one case, with the DR grade of the more serious eye accepted as the final diagnosis according to the ICDR severity grading system.

2.4. AI-based grading

The retinal photographs and medical history (after masking the patient's identity and diagnosis) were uploaded to the EyeWisdom platform for automatic diagnosis and grading. EyeWisdom is a software-based online cloud-computing platform. It has two systems: a DR-specific diagnosis system (EyeWisdom®DSS) and eye-related multi-disease diagnosis system (EyeWisdom®MCS). It can automatically analyze retinal images in conjunction with information about the patient's age, gender, and DM history, and then it provides information about the DR diagnosis and severity by automatically detecting the type, quantity, size, and location of retinopathy. In addition to the severity of DR, this software can also report the presence/absence of retinal hemorrhage, micro angioma, neovascularization, hard exudation, and fibroproliferative membrane. Images not clear enough to be diagnosed by EyeWisdom were excluded. The diagnosis of DR by EyeWisdom®DSS and EyeWisdom®MCS and the grading results of EyeWisdom®DSS were collected.

2.5. Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics 23.0 (SPSS Inc, Chicago, IL). Qualitative data were described as frequencies. Sensitivity, specificity, PPV, and NPV were used to

evaluate the diagnostic efficacy of EyeWisdom with the diagnosis of an ophthalmologist as a reference standard. Among these terms, PPV refers to the probability of disease when the test result is positive and NPV refers to the probability of absence of disease when the test result is negative. Kappa statistics were used to quantify and evaluate the consistency between AI analysis and the ophthalmologist's grading. All P-values were two-tailed, and the level of significance was set at $\alpha = 0.05$.

3. Results

3.1. Participant characteristics

Based on the inclusion criteria, 1768 retinal images from 563 DM patients were obtained; 14 DM patients (40 retinal images) were removed due to cataract or vitreous hemorrhage. A total of 549 DM patients aged 18–97 years old, for whom there were 1728 final images, were diagnosed by doctors and EyeWisdom. Of them, 272 (49.5%) were male and 277 (50.5%) were female. The mean age was 61.2 ± 11.8 years old. According to the ICDR standards, 225 (41%) were diagnosed as not having DR by the ophthalmologist. There were 324 (59.0%) DM patients diagnosed with DR, among whom 43 (7.8%) had mild NPDR, 79 (14.4%) had moderate NPDR, 61 (11.1%) had severe NPDR, and 141 (25.6%) had PDR. Based on the ICDR standards, typical fundus photographs of DR in different stages are shown in Fig. 1.

3.2. Comparison of ophthalmologist and AI DR diagnosis

EyeWisdom®DSS software detected 337 (61.4%) cases of DR in 549 DM patients, and EyeWisdom®MCS software detected 264 (48.1%) DR cases in these participants. Based on automatic grading by EyeWisdom®DSS, 68 (12.4%) patients had mild NPDR, 79 (14.4%) had moderate NPDR, 71 (12.9%) had severe NPDR, and 119 (21.7%) had PDR. The comparison of DR grading severity between the ophthalmologist and EyeWisdom®DSS software is shown in Fig. 2. In the 324 DR patients, 295 (91.0%) were correctly diagnosed with DR by EyeWisdom®DSS and 247 (76.2%) by EyeWisdom®MCS. For the 225 DM patients without DR, 83 (81.3%) were correctly diagnosed as not having DR by EyeWisdom®DSS, and 208 (92.4%) were correctly diagnosed by EyeWisdom®MCS. Fig. 3A shows the Venn diagram of the DR identified by the ophthalmologist versus AI and the overlap of DR observed in 373 patients. Fig. 3B shows the overlap of the absence of DR identified by the ophthalmologist versus AI. Fig. 3C–D show the overlap of NPDR and PDR identified by the ophthalmologist versus EyeWisdom®DSS software.

3.3. Efficacy of AI in the screening of DR in DM patients

The sensitivity, specificity, AUC, PPV, NPV, and kappa values for detecting DR, NPDR, and PDR using EyeWisdom software are shown in Table 1, in which ophthalmologist grading was taken as the reference standard. EyeWisdom®DSS correctly identified 91.0% (95% CI: 87.3%–93.8%) of patients with DR and 81.3% (95% CI: 75.5%–86.1%) of patients without DR, and

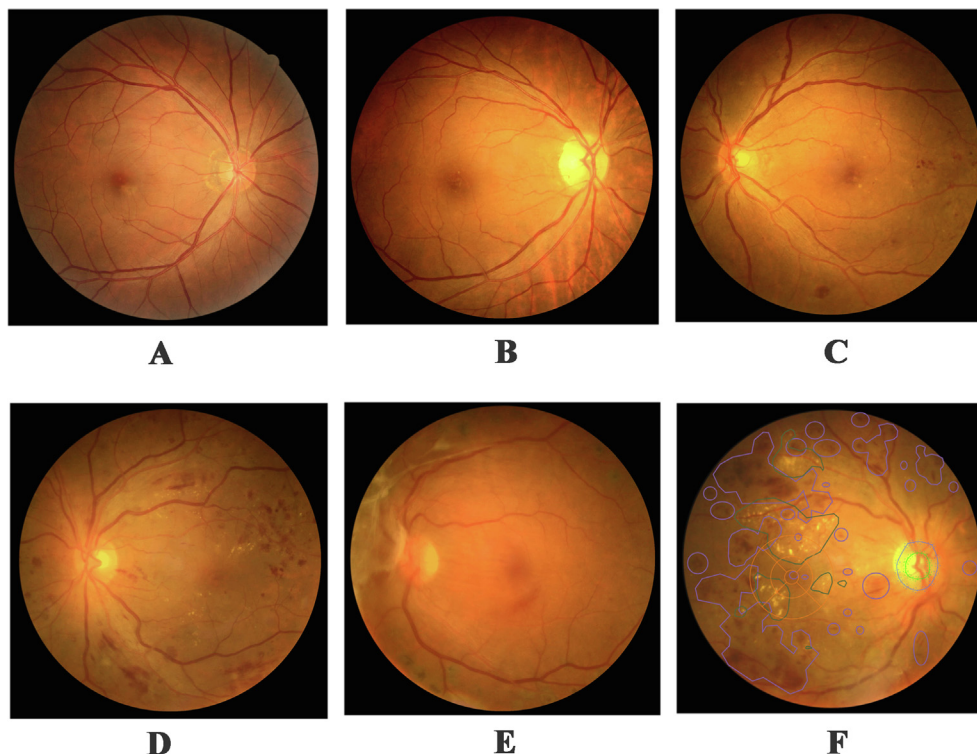


Fig. 1 – Typical fundus photographs of DR in different stages. Typical fundus photograph showing the absence of DR. (B–E) Typical fundus photographs of mild NPDR (B), moderate NPDR (C), severe NPDR (D), and PDR (E). (F) Fundus lesion markings of severe NPDR identified by EyeWisdom®DSS. The purple outline marks retinal hemorrhage, green identifies hard exudation, and yellow shows macular fovea.

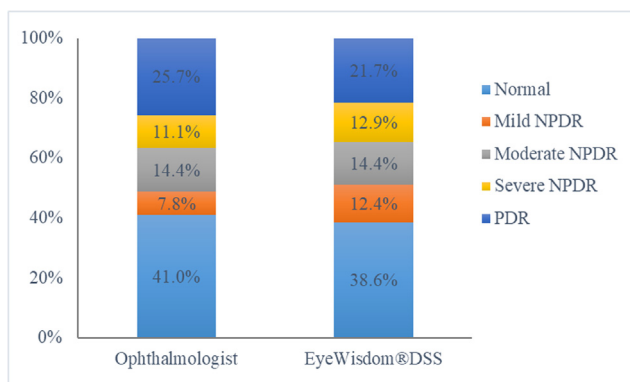


Fig. 2 – Comparison of ophthalmologist and EyeWisdom®DSS software DR severity grading.

EyeWisdom®MCS correctly identified 76.2% (95% CI: 71.1%–80.7%) of patients with DR and 92.4% (95% CI: 87.9%–95.4%) of patients without DR. EyeWisdom®DSS showed 76.5% (95% CI: 69.6%–82.3%) sensitivity and 78.4% (95% CI: 73.7%–82.5%) specificity for detecting NPDR and 64.5% (95% CI: 56.0%–72.3%) sensitivity and 93.1% (95% CI: 90.1%–95.3%) specificity in diagnosing PDR. The PPV of EyeWisdom®DSS for the detection of DR, NPDR, and PDR was 87.5% (95% CI: 83.4%–90.8%), 64.2% (95% CI: 89.7%–96.1%), and 74.5% (95% CI: 67.6%–83.6%), respectively. In addition, the NPV of EyeWisdom®DSS for the detection of DR, NPDR, and PDR was 86.3% (95% CI: 80.8%–90.5%), 86.8% (95% CI: 82.5%–90.2%), and 88.4%

(95% CI: 84.9%–91.2%), respectively. The degree of agreement between EyeWisdom®DSS and the ophthalmologist grading for DR was 0.730 ($P < 0.001$), for NPDR it was 0.527 ($P < 0.001$), and for PDR it was 0.608 ($P < 0.001$) using the kappa statistics. The kappa value between EyeWisdom®MCS and ophthalmologist grading for DR was 0.660 ($P < 0.001$).

4. Discussion

This study evaluated the accuracy of EyeWisdom AI software for DR screening. We found that EyeWisdom®DSS has higher sensitivity and EyeWisdom®MCS has greater specificity. That is, the specific disease system was good at identifying patients and the multi-disease system was good at identifying normal participants, which suggested EyeWisdom can be established as an AI-based DR-screening model to be used in community and grassroots clinics in China in the future. The combination of these two systems may improve the awareness and treatment rate of DR, including avoiding or delaying its progression.

With the aging of the global population and increased prevalence of DM, the incidence of DR is also increasing [20,21]. A meta-analysis indicated that from 1990 to 2017 in the Chinese population, the pooled prevalence of DR, NPDR, and PDR was 1.14%, 0.90%, and 0.07%, respectively, and in patients with DM, the corresponding prevalence was 18.45%, 15.06%, and 0.99%, respectively [22]. A study conducted by Ruta et al., which was based on 72 articles from 33 developing and developed countries, showed the prevalence of DR varied

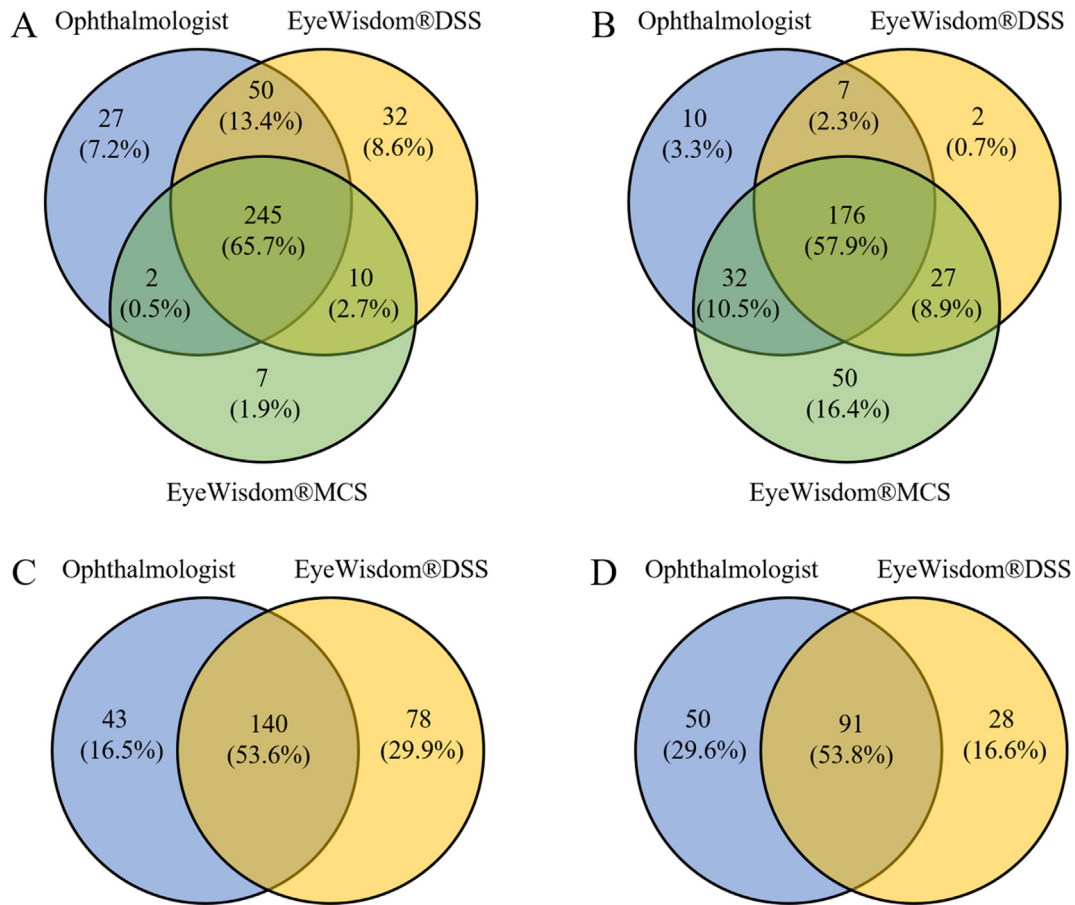


Fig. 3 – Overlap of DR and absence of DR identified by the ophthalmologist and AI. **(A)** Overlap of DR identified by the ophthalmologist versus AI (EyeWisdom@DSS and EyeWisdom@MCS). **(B)** Overlap of the absence of DR identified by the ophthalmologist versus AI (EyeWisdom@DSS and EyeWisdom@MCS). **(C–D)** Overlap of NPDR **(C)** and PDR **(D)**, identified by the ophthalmologist versus EyeWisdom@DSS software.

from 10% to 61% in people with known type 2 diabetes mellitus (T2DM) and from 1.5% to 31% in people with newly diagnosed T2DM [5]. The prevalence of DR was 33.2% in the United States and 17.6% in India [5]. In this study, the prevalence of DR, NPDR, and PDR in DM patients was 59.0%, 33.3%, and 25.7%, respectively. The reasons for the difference in DR prevalence in different regions are multifactorial and include differences in demographic characteristics, research methods, and diagnosis and classification criteria of DR. In addition, DM patients in this study were all from the Fundus Disease Center at Henan Provincial People's Hospital. They cannot represent the general DM population that from the Department of Endocrinology. Therefore, most of these individuals were DM patients who had developed ocular symptoms or fundus lesions. This may explain the higher prevalence of DR, NPDR, and PDR in this study.

It should be noted that the onset of DR is insidious, and most patients have a long asymptomatic period before visual impairment. During this time, fundus lesions can be easily identified by fundus examination or retinal photography. Early detection is necessary for a good DR prognosis [23]. Therefore, DR screening for all DM patients is cost-effective in the long run and significant in terms of public health, especially in developing countries [24]. However, in the grassroots areas of

China and other developing countries, the awareness and treatment rates of DR are very low due to a lack of medical resources, education, and experienced ophthalmologists. Existing DR-screening equipment has not been widely used due to the need for operation by professional ophthalmologists, its slow output, and the fact that it is inconvenient to move. In contrast, EyeWisdom, as an auxiliary AI diagnosis system for fundus diseases, has a number of advantages. If fundus photographs can be obtained, the system can provide diagnosis suggestions with only a computer. After simple training, the operator can complete the screening without needing professional ophthalmic knowledge. Therefore, this approach is not only suitable for large-scale screening of DR, glaucoma, and age-related macular degeneration, but it is also helpful for the long-term follow-up of DM patients. The system also enables remote guidance from ophthalmologists in tertiary hospitals, which could contribute to solving the problems related to diagnosis and treatment in poor areas.

The results of our study showed that the sensitivity of DR screening using EyeWisdom@DSS and the specificity of DR screening using EyeWisdom@MCS were high, reaching 91.0% and 92.4%, respectively. However, the efficacy of NPDR (76.5%) and PDR (64.5%) screening using EyeWisdom@DSS was relatively lower. For EyeWisdom@DSS, when the screen-

Table 1 – Efficacy of AI for detection of varying degrees of DR with ophthalmologist grading as reference standard (N = 549).

Retinopathy	Sensitivity (95% CI), %	Specificity (95% CI), %	AUC (95% CI)	PPV (95% CI), %	NPV (95% CI), %	Kappa	P
DR _{DSS}	91.0 (87.3, 93.8)	81.3 (75.5, 86.1)	0.862 (0.827, 0.897)	87.5 (83.4, 90.8)	86.3 (80.8, 90.5)	0.730	<0.001
DR _{MCS}	76.2 (71.1, 80.7)	92.4 (87.9, 95.4)	0.843 (0.809, 0.878)	93.5 (89.7, 96.1)	73.0 (67.4, 78.0)	0.660	<0.001
NPDR	76.5 (69.6, 82.3)	78.4 (73.7, 82.5)	0.776 (0.733, 0.819)	64.2 (57.4, 70.5)	86.8 (82.5, 90.2)	0.527	<0.001
PDR	64.5 (56.0, 72.3)	93.1 (90.1, 95.3)	0.788 (0.738, 0.839)	76.5 (67.6, 83.6)	88.4 (84.9, 91.2)	0.608	<0.001

DR_{DSS}: Diabetic retinopathy diagnosis by EyeWisdom®DSS; DR_{MCS}: Diabetic retinopathy diagnosis by EyeWisdom®MCS; NPDR: non-proliferative diabetic retinopathy; PDR: proliferative diabetic retinopathy; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value.

ing result was positive, the probability of DR was 87.5%, and when the screening result was negative, the probability of participants not suffering from DR was 86.3%. For EyeWisdom®MCS, when the screening result was positive, the probability of DR was 93.5%, and when the screening result was negative, the probability of participants not suffering from DR was 73.0%. Our sensitivity was similar to that found in He et al.'s study [23], in which the sensitivity of AI software (Airdoc, Beijing, China) was 90.8%. Another study by Rajalakshmi et al. reported that EyeArt smartphone-based AI software showed 95.8% sensitivity and 80.2% specificity for detecting DR. The kappa agreement between EyeArt and the ophthalmologist grading for DR was 0.78 and for PDR it was 0.53 [25]. EyeWisdom has two systems: EyeWisdom®DSS has the advantage in detecting "DR" and EyeWisdom®MCS is good at detecting "no DR". Thus, it is necessary to build a new synthetic deep-learning AI system based on the algorithms of both EyeWisdom®DSS and EyeWisdom®MCS that can detect both "DR" and "no DR" effectively.

Undoubtedly, some limitations of this study should be noted. First, although EyeWisdom can diagnose and grade DR through an AI algorithm, it is not suitable for some patients. For example, it is not possible to obtain fundus photographs from some DM patients due to their small pupils or poor image quality from the opacity of cataracts. Second, EyeWisdom completes the diagnosis and grading according to the location and number of typical fundus lesions, such as retinal hemorrhage, micro angioma, and hard exudation. For retinal hemorrhage or neovascularization caused by other diseases, EyeWisdom cannot make a differential diagnosis. Third, the study participants were DM patients at the Fundus Disease Center, most of which had developed eye-related symptoms or retinopathy. Therefore, the prevalence and diagnostic efficacy found in this study may not be representative for all DM patients. Finally, the sample size was relatively small and the patients were from a single hospital; therefore, future studies should be conducted with larger samples and field settings.

Measures can be taken to address these limitations in the future. For example, the AI algorithm can be optimized to enable the system to make a differential diagnosis by taking into account more detailed information, such as disease history indicators, duration of disease, and pathological characteristics. For patients with a smaller pupil, manual photography after mydriasis and then transmission of the image to EyeWisdom®DSS and EyeWisdom®MCS may be suitable. Where mydriasis cannot be photographed, scanning laser ophthalmoscopy is currently used for DR diagnosis.

In conclusion, the prevalence of DR, NPDR, and PDR was high in patients with DM. On the one hand, EyeWisdom®DSS is more proficient at identifying DR, but its DR-classification accuracy is relatively poor. EyeWisdom®MCS, on the other hand, is better at identifying the absence of DR. Although the classification efficacy of EyeWisdom is poor, it has the benefits of economy, simple operation, convenient image transmission, and remote guidance. With the development of the Internet and 5G technology, using AI to diagnose DR will undoubtedly save significant manpower and financial resources in countryside or rural district and help to make population screening for DR more affordable [26,27]. Therefore, the system can not only provide a diagnostic platform

for community clinics and countryside areas in developing countries, but it can also help establish a large-scale AI-based screening model for DR. It can also play an important auxiliary role in improving the awareness, treatment, and monitoring of disease progression of patients in developing countries, especially in rural areas.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Ministry of Science and Technology of the Peoples Republic of China [grant number: 2018YFC0114500 to YW and ZL], and the National Natural Science Foundation of China [grant numbers: 81770962 and 81470603 to ZL]. We thank all members of our study team for their whole-hearted cooperation and the retinal specialists (Handong Dan, Dong Qin, and Dongdong Wang) for their diagnosis.

Author contribution

Z.L. and X.P. contributed to the conception of the study. X.P. and X.Y. contributed to the data interpretation, data analysis and manuscript writing. Y.Y. and H.Z. contributed to interpretation, discussion, reviewed/edited the manuscript. M.X. and R.H. contributed to the data collection. Y.W. contributed to the grammar of the manuscript. All authors read and approved the final manuscript.

REFERENCES

- [1] International diabetes federation diabetes atlas, 2019, 9th Edition [Internet]. Accessed on 13 October 2020. Available from: <http://www.diabetesatlas.org/>.
- [2] Bawankar P, Shanbhag N, K. SS, Dhawan B, Palsule A, Kumar D, et al. Sensitivity and specificity of automated analysis of single-field non-mydratric fundus photographs by Bosch DR Algorithm-Comparison with mydratric fundus photography (ETDRS) for screening in undiagnosed diabetic retinopathy. *PLoS One* 2017;12(12):e0189854.
- [3] Lian F, Wu L, Tian J, Jin M, Zhou S, Zhao M, et al. The effectiveness and safety of a danshen-containing Chinese herbal medicine for diabetic retinopathy: a randomized, double-blind, placebo-controlled multicenter clinical trial. *J. Ethnopharmacol.* 2015;164:71–7.
- [4] Sivaprasad S, Gupta B, Crosby-Nwaobi R, Evans J. Prevalence of diabetic retinopathy in various ethnic groups: a worldwide perspective. *Surv. Ophthalmol.* 2012;57(4):347–70.
- [5] Ruta LM, Magliano DJ, Lemesurier R, Taylor HR, Zimmet PZ, Shaw JE. Prevalence of diabetic retinopathy in Type 2 diabetes in developing and developed countries. *Diabet. Med.* 2013;30:387–98.
- [6] Sadda SR. Assessing the severity of diabetic retinopathy: early treatment diabetic retinopathy study report number 10. *Ophthalmology* 2020;127(4):S97–8.

- [7] Tan F, Chen Qi, Zhuang X, Wu C, Qian Y, Wang Y, et al. Associated risk factors in the early stage of diabetic retinopathy. *Eye Vis (Lond)*. 2019;6(1):23.
- [8] Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit Med*. 2019;2:92.
- [9] Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. *Prog. Retin. Eye Res*. 2018;67:1–29.
- [10] Huang OS, Tay WT, Tai ES, Wang JJ, Saw SM, Jeganathan VS, et al. Lack of awareness amongst community patients with diabetes and diabetic retinopathy: the Singapore Malay eye study. *Ann. Acad. Med. Singap*. 2009;38:1048–55.
- [11] Teng Y, Cui H, Zhang QS, Teng YF, Su Y, Yang MM, et al. Prevalence of diabetic retinopathy among the elderly in rural southern Shuangcheng city, Heilongjiang province. *Chin. J. Epidemiol*. 2010;31:856–9.
- [12] Bakkar MM, Haddad MF, Gammoh YS. Awareness of diabetic retinopathy among patients with type 2 diabetes mellitus in Jordan. *Diabetes Metab. Syndr. Obes*. 2017;10:435–41.
- [13] Ting DSW, Cheung C-L, Lim G, Tan GSW, Quang ND, Gan A, 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. *JAMA* 2017;318(22):2211.
- [14] Zhiyuanhuitu Vistel [Internet]. Accessed on 10 October 2020. Available from: <https://vistel.cn/>.
- [15] Cuadros J, Sim I. EyePACS: an open source clinical communication system for eye care. *Stud. Health Technol. Inform*. 2004;107:207–11.
- [16] Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform*. 2014;48:193–204.
- [17] Li Y, Teng D, Shi X, Qin G, Qin Y, Quan H, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: national cross sectional study. *BMJ* 2020;369:m997.
- [18] Seino Y, Nanjo K, Tajima N, Kadowaki T, Kashiwagi A, Araki E, et al. Report of the committee on the classification and diagnostic criteria of diabetes mellitus. *J. Diabetes Investig*. 2010;1(5):212–28.
- [19] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003;110(9):1677–82.
- [20] Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res. Clin. Pract*. 2010;87(1):4–14.
- [21] Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res. Clin. Pract*. 2014;103(2):137–49.
- [22] Song P, Yu J, Chan KY, Theodoratou E, Rudan I. Prevalence, risk factors and burden of diabetic retinopathy in China: a systematic review and meta-analysis. *J. Glob. Health* 2018;8(1). <https://doi.org/10.7189/jogh.08.010803>.
- [23] He J, Cao T, Xu F, Wang S, Tao H, Wu T, et al. Artificial intelligence-based screening for diabetic retinopathy at community hospital. *Eye (Lond)*. 2020;34(3):572–6.
- [24] Rajalakshmi R, Arulmalar S, Usha M, Prathiba V, Kareemuddin KS, Anjana RM, et al. Validation of smartphone based retinal photography for diabetic retinopathy screening. *PLoS One* 2015;10(9):e0138285.
- [25] Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye (Lond)*. 2018;32(6):1138–44.
- [26] Xie Y, Gunasekeran DV, Balaskas K, Keane PA, Sim DA, Bachmann LM, et al. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. *Transl. Vis. Sci. Technol*. 2020;9(2):22.
- [27] Fenner BJ, Wong RLM, Lam W-C, Tan GSW, Cheung GCM. Advances in retinal imaging and applications in diabetic retinopathy screening: a review. *Ophthalmol. Ther*. 2018;7(2):333–46.