BMJ Open

Automated multidimensional deep learning platform for referable diabetic retinopathy detection: a multicentre, retrospective study

Guihua Zhang,[1] Jian-Wei Lin,[1] Ji Wang ,[1] Jie Ji,[2] Ling-Ping Cen,[1] Weiqi Chen ,[1] Peiwen Xie,[1] Yi Zheng,[1] Yongqun Xiong,[1] Hanfu Wu,[1] Dongjie Li,[1] Tsz Kin Ng ,[1] Chi Pui Pang,[1,3] Mingzhi Zhang [1]

To cite: Zhang G, Lin J-W, Wang J, *et al*. Automated multidimensional deep learning platform for referable diabetic retinopathy detection: a multicentre, retrospective study. *BMJ Open* 2022;**12**:e060155. doi:10.1136/bmjopen-2021-060155

- ► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-060155).

GZ, J-WL, JW and JJ contributed equally.

Received 28 December 2021
Accepted 20 June 2022

Check for updates

[1]Joint Shantou International Eye Center of Shantou University and The Chinese University of Hong Kong, Shantou, China
[2]The big data center, Shantou University Medical College, Shantou, China
[3]Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong, China

Correspondence to
Dr Mingzhi Zhang;
zmz@jsiec.org

ABSTRACT

Objective To develop and validate a real-world screening, guideline-based deep learning (DL) system for referable diabetic retinopathy (DR) detection.

Design This is a multicentre platform development study based on retrospective, cross-sectional data sets. Images were labelled by two-level certificated graders as the ground truth. According to the UK DR screening guideline, a DL model based on colour retinal images with five-dimensional classifiers, namely image quality, retinopathy, maculopathy gradability, maculopathy and photocoagulation, was developed. Referable decisions were generated by integrating the output of all classifiers and reported at the image, eye and patient level. The performance of the DL was compared with DR experts.

Setting DR screening programmes from three hospitals and the Lifeline Express Diabetic Retinopathy Screening Program in China.

Participants 83 465 images of 39 836 eyes from 21 716 patients were annotated, of which 53 211 images were used as the development set and 30 254 images were used as the external validation set, split based on centre and period.

Main outcomes Accuracy, F1 score, sensitivity, specificity, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), Cohen's unweighted κ and Gwet's AC1 were calculated to evaluate the performance of the DL algorithm.

Results In the external validation set, the five classifiers achieved an accuracy of 0.915–0.980, F1 score of 0.682–0.966, sensitivity of 0.917–0.978, specificity of 0.907–0.981, AUROC of 0.9639–0.9944 and AUPRC of 0.7504–0.9949. Referable DR at three levels was detected with an accuracy of 0.918–0.967, F1 score of 0.822–0.918, sensitivity of 0.970–0.971, specificity of 0.905–0.967, AUROC of 0.9848–0.9931 and AUPRC of 0.9527–0.9760. With reference to the ground truth, the DL system showed comparable performance (Cohen's κ: 0.86–0.93; Gwet's AC1: 0.89–0.94) with three DR experts (Cohen's κ: 0.89–0.96; Gwet's AC1: 0.91–0.97) in detecting referable lesions.

Conclusions The automatic DL system for detection of referable DR based on the UK guideline could achieve high accuracy in multidimensional classifications. It is suitable for large-scale, real-world DR screening.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The data set in this study was constructed from multiple centres using different devices for generalisation.
- ⇒ The five-dimensional classifiers, namely image quality, retinopathy, maculopathy gradability, maculopathy and photocoagulation, were developed according to a real-world approach to diabetic retinopathy screening.
- ⇒ The deep learning platform can automatically generate three-level (image, eye and patient level) referable diabetic retinopathy decision.
- ⇒ Evaluation of two dimensions of quality, namely image quality and maculopathy gradability, on images with diverse qualities was given full consideration, which is consistent with screening practices.
- ⇒ Diabetic macular edema could be misdiagnosed in some cases without stereoscopic images and optical coherence tomography.

INTRODUCTION

Diabetic retinopathy (DR), a common ocular complication with retinal microvascular lesions in patients with diabetes mellitus (DM), is one of the leading causes of irreversible blindness and visual impairment among working-age people worldwide.[1] It is estimated that the DM population will increase to approximately 700 million by 2045, with a quarter suffering from DR.[2 3] DR screening programmes are an important interventional strategy for early identification of referable DR and allow timely referral and treatment to prevent vision loss due to DR.[4 5] Yet huge screening demand from a large number of patients with DM and the limited amount of human resources hinder the popularity and sustainability of screening services.[6]

Deep learning (DL), a subset of artificial intelligence (AI) powered by recent advances in computation and big data, permits multilayer convolutional neural networks to be trained through back propagation techniques to minimise an error function resulting in a classifier output, which works remarkably well in computer vision (image classification tasks).[7] In recent years, multiple DL algorithms for automatic detection of DR have been proposed and have shown high sensitivity and specificity (>90%) in detecting referable DR,[8–11] throwing light to large-scale DR screening with AI assistance. However, in complex real-world screening scenarios, an appropriate decision for referral is hard based only on a few dimensions of classifications. Various features or conditions should be identified and handled simultaneously, including the image quality of the fundus photos, stage of DR, maculopathy and photocoagulation status.[12–14] Hence, the DL algorithm on multidimensional features for referable DR detection should be developed to identify multiple conditions in the complex real-world DR screening scenarios. Herein, this study aimed to develop a multidimensional DL platform for detecting referable DR with five independent classifiers (image quality, retinopathy, maculopathy gradability, maculopathy and photocoagulation) using real-world DR screening data sets. Combined heatmaps were generated to visualise and explain the predicted areas of the referable lesions. The performance of our DL platform was further compared with retinal specialists.

METHODS

Informed consent was approved to be waived for this retrospective study using de-identified retinal images for the development of a DL system. This study followed the Standards for Reporting of Diagnostic Accuracy reporting guidelines.

Data sets

The images in this study were captured during DR screening programmes using three types of cameras and were collected from three hospitals (Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong (JSIEC; camera: Top-2000, Topcon, Japan); Liuzhou City Red Cross Hospital (Liuzhou; camera: AFC-230, NIDEK, Japan); and the Second Affiliated Hospital of Shantou University Medical College (STU-2nd; camera: Top-2000, Topcon, Japan)) and one event (Lifeline Express Diabetic Retinopathy Screening Program (LEDRSP); cameras: AFC-230, NIDEK, Japan, and Canon CR-DGi, Canon, Japan) from April 2014 to June 2018. Only mydriatic retinal images with two 45° fields (macula-centred and optic disc-centred) were included. Unless coexisting with DR, images presenting other ocular diseases, such as glaucoma and age-related macular degeneration, were excluded. Non-fundus images were also excluded (online supplemental figure 1).

Patient and public involvement

Neither participants nor the public were involved in the design and conduct of the present research.

Labelling and grading

Based on the English National Health Screening (NHS) Diabetic Eye Screening Programme (online supplemental table 1),[14 15] the retinal images were assessed in four dimensions, namely (1) image quality, (2) retinopathy, (3) maculopathy and (4) photocoagulation status. The labels of the retinal images were annotated according to the following:

- ‘Image quality’ was categorised as Q0 (ungradable quality, defined as an image with >one-third of the area poorly exposed, with artefact or blur which could not be classified confidently even when any DR feature was observed in the rest of the area) and Q1 (gradable quality, with ≤1/3 poor area that the image could be classified with confidence).
- ‘Retinopathy’ was divided into four levels according to severity of lesions: R0 (no DR), R1 (background DR), R2 (preproliferative DR) and R3 (proliferative DR). R0 and R1 were further defined as non-referable retinopathy, while R2 and R3 were defined as referable retinopathy.
- ‘Maculopathy’ was classified as M0 (absence of any M1 features) and M1 (exudate within 1 disc diameter (DD) of the centre of the fovea or any microaneurysm/haemorrhage within 1 DD of the centre of the fovea only if associated with a best corrected visual acuity of ≤ 6/12). Additionally, due to the limited blur or artefact (less than 1/3 area of the whole image) on the macula, maculopathy might be ungradable. Thus, evaluation of maculopathy gradability should precede classification of maculopathy and the image which could not be graded confidently in terms of maculopathy would be annotated as maculopathy ungradable (Mu).
- ‘Photocoagulation’ was categorised as P0 (image without laser spot or scar) and P1 (image presenting laser spot or scar).

Detailed definitions are shown in online supplemental table 2.

The ground truth labels of the images were obtained from grading of two-level graders. All graders have been trained and certificated by the NHS retinal screening for DR (https://www.gregcourses.com). The workflows for grading with clinical information are as follows: (1) images were primarily graded by two junior graders (PX and YZ) independently and the consistent labels were assigned as the ground truth labels; and (2) images with inconsistent labels from primary grading were submitted for final adjudication by a senior retinal ophthalmologist (GZ). The final adjudication was assigned as the ground truth label. Images satisfying the inclusion criteria and annotated by the ground truth labels were filed as the data set. The development set was constructed using images from LEDRSP and JSIEC, and further randomly divided

Table 1 Summary of the data sets

		Development set			External validation set		
	Total	JSIEC	LEDRSP	Subtotal	Liuzhou	STU-2nd	Subtotal
Camera type	AFC-230, Canon CR-DGi, Top-2000	Top-2000	AFC-230, Canon CR-DGi	AFC-230, Canon CR-DGi, Top-2000	AFC-230	Top-2000	AFC-230, Top-2000
Periods	April 2014–June 2018	June 2016–June 2018	April 2014–April 2016	April 2014–June 2018	June 2017–June 2018	June 2016–June 2018	June 2016–June 2018
Images, n/N (%)	83 465/83 465 (100)	2567/83 465 (3.1)	52 313/83 465 (60.7)	53 211/83 465 (63.8)	12 898/83 465 (15.5)	17 356/83 465 (20.8)	30 254/83 465 (36.2)
Eyes, n/N (%)	39 836/39 836 (100)	2241/39 836 (5.6)	24 299/39 836 (61.0)	26 540/39 836 (66.6)	6572/39 836 (16.5)	6916/39 836 (17.4)	13 488/39 836 (33.9)
Patients, n/N (%)	21 716/21 716 (100)	2051/21 716 (9.4)	13 026/21 716 (60.0)	15 077/21 716 (69.4)	3298/21 716 (15.2)	3512/21 716 (16.2)	6810/21 716 (31.4)
Patients with sex available, n/N (%)	17 042/21 716 (78.5)	1804/2051 (88.0)	8685/13 026 (66.7)	10 489/15 077 (69.6)	3298/3298 (100)	3426/3512 (97.6)	6724/6810 (98.7)
Male, n/N (%)	7493/17 042 (44.0)	932/1804 (51.66)	3893/8685 (44.8)	4825/10 489 (46.0)	1284/3298 (38.9)	1465/3426 (42.8)	2749/6724 (40.9)
Patients with age available, n/N (%)	20 150/21 716 (92.5)	1804/2051 (88.0)	11 793/13 026 (90.5)	13 597/15 077 (90.2)	3298/3298 (100)	3426/3512 (97.6)	6724/6810 (98.7)
Age, mean (SD), years	60.0 (12.9)	44.6 (18.8)	61.4 (10.1)	59.2 (13.0)	65.2 (9.7)	57.7 (13.9)	61.4 (12.6)

JSIEC, Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong; LEDRSP, Lifeline Express Diabetic Retinopathy Screening Program; Liuzhou, Liuzhou City Red Cross Hospital; STU-2nd, Second Affiliated Hospital of Shantou University Medical College.

into training, validation and test data sets by 75:10:15 ratio at the patient level, while images from Liuzhou and STU-2nd were used as the external validation set.

DL algorithm development

The pipeline of the DR screening system is shown in online supplemental figure 2. Briefly, image evaluation was initiated with assessment of image quality, where gradable images were inputted into the main pipeline, whereas ungradable ones were recommended for 'rephotography'. To construct the main structure of the system, we proposed four-dimensional independent classifiers (retinopathy, maculopathy gradability, maculopathy and photocoagulation) for any given images and each of the classifiers was binary. Three different kinds of neural networks (Google Inception-V3, Xception and Inceptin-ReNet-V2) were used as the base model and unweighted average was used as the model ensemble method. We also adopted a postprocessing method to integrate all single-dimension results as the image-level referable results, and further integrated the image-level results as the eye-level or patient-level results. The details of the methods are shown in online supplemental method 1.

The t-distributed stochastic neighbour embedding (t-SNE) heatmaps were used to visualise the features extracted by the neural networks. SHAP-CAM heatmap, combining Class Activation Mapping (CAM)[16 17] and DeepSHAP,[18] was used to highlight the important regions that the neural networks were used for making the predictions (online supplemental method 2).

Various recommendations were automatically generated by the system to respond the classifications of different classifiers: (1) patients with more serious lesions (R2, R3 or M1), defined as 'referable DR' by the English NHS Diabetic Eye Screening Programme (online supplemental table 1), were advised for referral in the study, whereas those with R0, R1 or M0 were advised for follow-up; (2) images with ungradable maculopathy were generally advised for rephotography, unless R2 or R3 was detected on the same image, or any referable DR was found on other field images of the same fundus; and (3) any laser spot or scar recognised on the image would remind of 'photocoagulation therapy once', suggesting previous consultation with an ophthalmologist. The following is the order of priority of various recommendations: 'refer to previous ophthalmologist' > 'referable' > 'rephotography' > 'follow-up'. Referable decision was automatically generated by the system that the image-level decision was integrated from multiple classifiers, and any dimensional positive prediction of referable lesion would recommend a referable decision. The referable image would further provide referable recommendation at the eye and patient level.

Statistical analysis

The performance of the classifiers was evaluated by true negative, false positive, false negative, true positive, F1 score, sensitivity, specificity, area under the receiver

Table 2 Performance of the five classifiers of the system

		n					F1			
Classifiers	**Data set**	**TN**	**FP**	**FN**	**TP**	**Accuracy**	**score**	**Sensitivity**	**Specificity**	**AUROC (95% CI)**
Image quality	Training	5818	180	1741	32 145	0.952	0.971	0.949	0.97	0.9882 (0.9874 to 0.9891)
	Validation	804	64	231	4165	0.944	0.966	0.947	0.926	0.9812 (0.9780 to 0.9844)
	Test	1095	103	418	6447	0.935	0.961	0.939	0.914	0.9768 (0.9737 to 0.9798)
	External validation	4706	243	1429	23 876	0.945	0.966	0.944	0.951	0.9751 (0.9732 to 0.9770)
Retinopathy	Training	30 118	544	9	3215	0.984	0.921	0.997	0.982	0.9992 (0.9991 to 0.9994)
	Validation	3916	107	15	358	0.972	0.854	0.96	0.973	0.9956 (0.9941 to 0.9970)
	Test	6031	178	16	640	0.972	0.868	0.976	0.971	0.9962 (0.9951 to 0.9972)
	External validation	21 609	784	65	2847	0.966	0.87	0.978	0.965	0.9944 (0.9936 to 0.9952)
Maculopathy gradability	Training	5068	148	1269	27 401	0.958	0.975	0.956	0.972	0.9934 (0.9928 to 0.9940)
	Validation	617	43	174	3562	0.951	0.97	0.953	0.935	0.9896 (0.9873 to 0.9918)
	Test	970	49	302	5544	0.949	0.969	0.948	0.952	0.9890 (0.9871 to 0.9910)
	External validation	4374	451	1704	18 776	0.915	0.946	0.917	0.907	0.9639 (0.9617 to 0.9660)
Maculopathy	Training	24 301	470	170	3729	0.978	0.921	0.956	0.981	0.9962 (0.9957 to 0.9967)
	Validation	3186	104	29	417	0.964	0.862	0.935	0.968	0.9906 (0.9864 to 0.9948)
	Test	4900	130	61	755	0.967	0.888	0.925	0.974	0.9928 (0.9912 to 0.9944)
	External validation	16 987	572	150	2771	0.965	0.885	0.949	0.967	0.9904 (0.9888 to 0.9919)
Photocoagulation	Training	32 526	105	0	1255	0.997	0.96	1.000	0.997	1.0000 (0.9999 to 1.0000)
	Validation	4252	27	5	112	0.993	0.875	0.957	0.994	0.9924 (0.9794 to 1.0000)
	Test	6589	33	8	235	0.994	0.92	0.967	0.995	0.9979 (0.9958 to 1.0000)
	External validation	24 277	467	29	532	0.98	0.682	0.948	0.981	0.9904 (0.9869 to 0.9940)
Referable DR*										
Images	Training	23 888	362	147	4890	0.983	0.951	0.971	0.985	0.9980 (0.9977 to 0.9983)
	Validation	3138	89	28	544	0.969	0.903	0.951	0.972	0.9932 (0.9899 to 0.9965)
	Test	4838	114	55	953	0.972	0.919	0.945	0.977	0.9952 (0.9940 to 0.9964)
	External validation	16 667	575	117	3859	0.967	0.918	0.971	0.967	0.9931 (0.9920 to 0.9942)
Eyes	Training	14 764	411	110	3136	0.972	0.923	0.966	0.973	0.9961 (0.9955 to 0.9967)
	Validation	1986	87	15	342	0.958	0.87	0.958	0.958	0.9906 (0.9850 to 0.9961)
	Test	2949	117	36	608	0.959	0.888	0.944	0.962	0.9923 (0.9901 to 0.9946)
	External validation	9429	624	59	1876	0.943	0.846	0.97	0.938	0.9884 (0.9863 to 0.9905)
Patients	Training	8415	291	74	2141	0.967	0.921	0.967	0.967	0.9956 (0.9949 to 0.9964)
	Validation	1138	64	10	237	0.949	0.865	0.96	0.947	0.9894 (0.9837 to 0.9951)
	Test	1669	80	25	407	0.952	0.886	0.942	0.954	0.9914 (0.9884 to 0.9943)
	External validation	4683	492	37	1219	0.918	0.822	0.971	0.905	0.9848 (0.9819 to 0.9877)

*By integrating the prediction of referable retinopathy and maculopathy on an image, referable DR decisions were given by the system when any referable lesion was detected, and the accuracies were based on the image, eye and patient levels.
AUROC, area under the receiver operating characteristic curve; DR, diabetic retinopathy; FN, false negative; FP, false positive; TN, true negative; TP, true positive.

operating characteristic curve (AUROC) with 95% CI and area under the precision-recall curve (AUPRC).[19] The open source package pROC (V.1.14.0; Xavier Robin) was used to calculate two-sided 95% CI with the DeLong method for AUROC. Data were analysed from 1 May 2019 to 12 June 2021.

An extra independent data set of 253 images from JSIEC and STU-2nd between 1 January 2019 and 31 December 2020 was used in the human–machine comparison with three experienced retinal ophthalmologists for further validation. The ground truth labelled by two-level graders was considered as the criterion standard. For the human–system comparison, the consistency between the graders (three experienced retinal ophthalmologists) and the DL system and the criterion standard were calculated by the Cohen's unweighted κ and Gwet's AC1.[20 21] Both of them were further graded using the following scale: 0.2 or less was considered slight agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as strong and 0.81–1.0 as near-complete agreement.

RESULTS

A total of 85 977 retinal images were collected and 2512 (2.9%) were excluded due to a non-fundus view or diseases other than DR, which would reduce the classification performance of the DL system if included in the data sets. Subsequently, a total of 83 465 images of 39 836 eyes from 21 716 patients (mean age of 20 150 patients with available age: 60.0±12.9 years; 7493 of 17 042 (44.0%) patients with known sex as male) were eventually annotated and included in the data sets. The development set compiled from JSIEC and LEDRSP included 53 211 images (63.8% of 83 465), and the external test set from Liuzhou and STU-2nd included 30254 images (36.2%). The distribution of the data is shown in table 1 and online supplemental table 3.

System performance

For the test set at the image level, the performance of all classifiers achieved an accuracy of 0.935–0.994, F1 score of 0.868–0.969, sensitivity of 0.925–0.976, specificity of 0.914–0.995, AUROC from 0.9768 (95% CI 0.9737 to 0.9798) to 0.9979 (95% CI 0.9958 to 1.0000) and AUPRC of 0.9578–0.9981 (table 2, figure 1 and online supplemental figures 3 and 4). The retinopathy classifier achieved an accuracy of 0.972, F1 score of 0.868, sensitivity of 0.976, specificity of 0.971, AUROC of 0.9962 (95% CI 0.9951 to 0.9972) and AUPRC of 0.9687, whereas the maculopathy classifier achieved an accuracy of 0.967, F1 score of 0.888, sensitivity of 0.925, specificity of 0.974, AUROC of 0.9928 (95% CI 0.9912 to 0.9944) and AUPRC of 0.9578.

For the external validation set at the image level, the performance of all classifiers achieved an accuracy of 0.915–0.980, F1 score of 0.682–0.966, sensitivity of 0.917–0.978, specificity of 0.907–0.981, AUROC from 0.9639 (95% CI 0.9617 to 0.9660) to 0.9944 (95% CI 0.9936 to 0.9952) and AUPRC of 0.7504–0.9949 (table 2, figure 1 and online supplemental figures 3 and 4). The retinopathy classifier achieved an accuracy of 0.966, F1 score of 0.870, sensitivity of 0.978, specificity of 0.965, AUROC of 0.9944 (95% CI 0.9936 to 0.9952) and AUPRC of 0.9617, whereas the maculopathy classifier achieved an accuracy of 0.965, F1 score of 0.885, sensitivity of 0.949, specificity of 0.967, AUROC of 0.9904 (95% CI 0.9888 to 0.9919) and AUPRC of 0.9551.

The performance of the three-level (image, eye and patient level) referable DR detection achieved an accuracy of 0.952–0.972, F1 score of 0.886–0.919, sensitivity of 0.942–0.945, specificity of 0.954–0.977, AUROC from 0.9914 (95% CI 0.9884 to 0.9943) to 0.9952 (95% CI 0.9940 to 0.9964) and AUPRC of 0.9679–0.9773 in the test set, and an accuracy of 0.918–0.967, F1 score of 0.822–0.918, sensitivity of 0.970–0.971, specificity of 0.905–0.967, AUROC from 0.9848 (95% CI 0.9819 to 0.9877) to 0.9931 (95% CI 0.9920 to 0.9942) and AUPRC of 0.9527–0.9760 in the external validation set (table 2, figure 1 and online supplemental figures 3 and 4).

Visualisation

The t-SNE helped in the reduction of high-dimensional data extraction from the neural network and structure visualisation on a two-dimensional map. Well-identified binary classes of each classifier are shown in online supplemental figure 5.

In the SHAP-CAM heatmap, the predictive referable lesion visualisation not only showed their located domain, but also the shape of the lesions, which were more fined-discriminative than the CAM heatmaps and with less noise than the DeepSHAP (figure 2).

Human–system comparison

Further validation was conducted on the detection of referable DR lesions between our DL algorithm and three experienced retinal ophthalmologists. Higher sensitivity was found for the DL algorithm (1.000 in retinopathy, 0.949 in maculopathy and 0.953 in referable DR) as compared with that of the retinal ophthalmologists (average (range): 0.935 (0.910–0.970) in referable retinopathy, 0.936 (0.910–0.949) in referable maculopathy and 0.933 (0.918–0.953) in referable DR; table 3). Confusion matrices showed near-complete agreement (Cohen's κ: 0.86–0.93; Gwet's AC1: 0.89–0.94) between the DL algorithm and the ground truth label (online supplemental figure 6), which was comparable with the retinal ophthalmologists (Cohen's κ: 0.89–0.96; Gwet's AC1: 0.91–0.97).

False prediction analysis

The false predictions in the external validation set were analysed by visualisation of heatmaps. Most of the false positives were due to the non-referable DR lesions, including the background DR predicted as referable retinopathy (646 of 784, 82.4%) and the haemorrhage/microaneurysm in the macula with best corrected visual acuity >0.5 as referable maculopathy (178 of 572, 31.1%). Meanwhile, the artefacts were the common interference factor in false positive classifications (7.4% in referable retinopathy and 20.6% in referable maculopathy). Limited blurred images were observed in the false negative predictions for both referable lesions (online supplemental table 4).

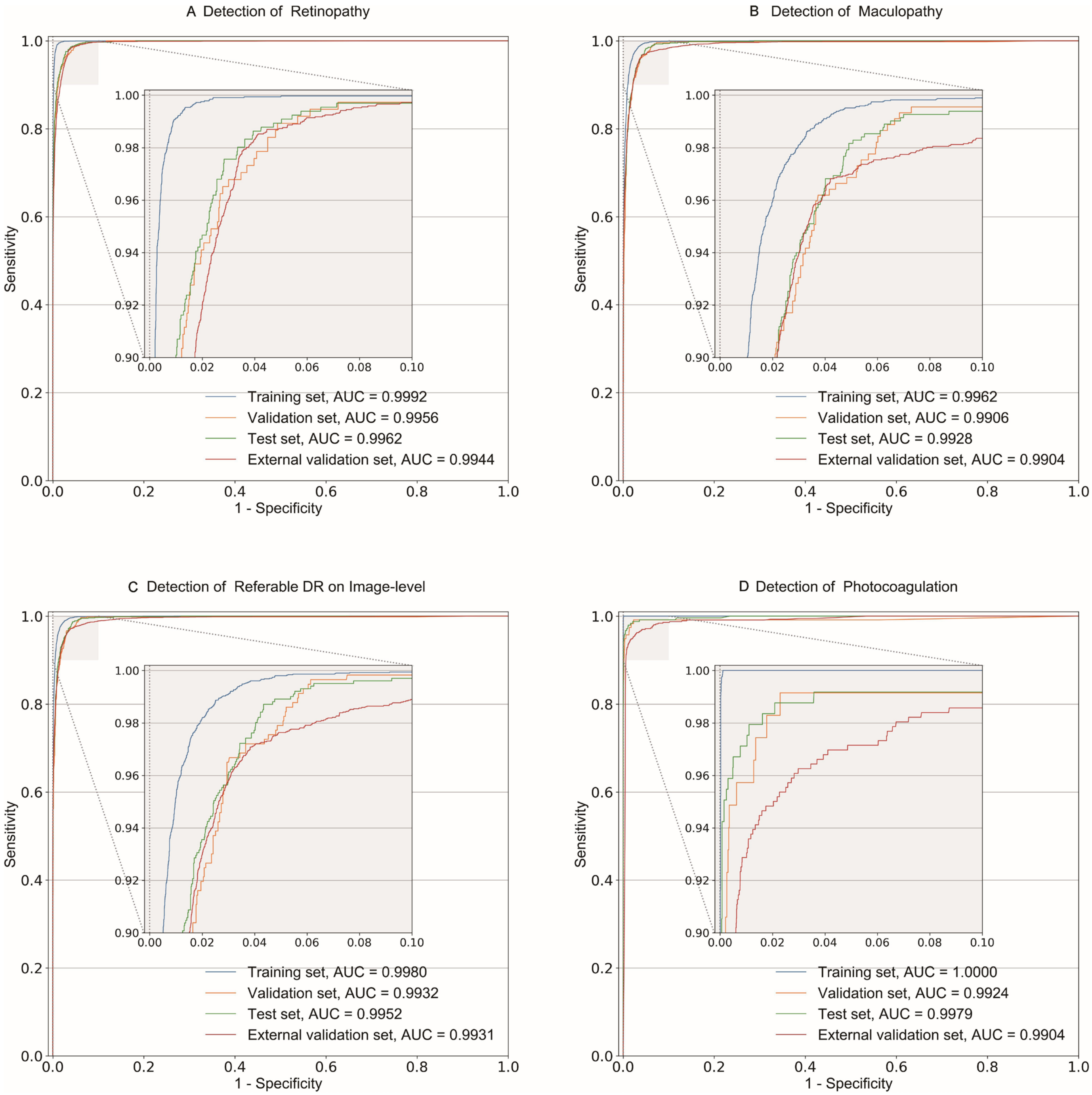


Figure 1 Receiver operating characteristic curves of the main dimensional classifiers and referable DR detection. The classification performances of four subsets (training, validation, test and external validation) are shown as receiver operating characteristic curves and AUC for detection of referable retinopathy (A), referable maculopathy (B), image-level referable DR (C) and photocoagulation (D). Notably, the detection of referable DR on an image (D) was automatically generated by integrating the results of referable retinopathy and referable maculopathy. AUC, area under receiver operating characteristic curve; DR, diabetic retinopathy.

DISCUSSION

In this study, we developed a multidimensional DL platform for DR screening based on a real-world screening guideline. Our results demonstrated that (1) the five-dimensional classifiers (image quality, retinopathy, maculopathy gradability, maculopathy and photocoagulation) achieved high accuracy in each classification; (2) a three-level referable DR decision (image, eye and patient level) could be automatically generated by the DL platform; and (3) visualisation by the SHAP-CAM heatmaps provided the explainability for the referable lesion prediction from the platform.

In this study, multiple dimensional classifications were based on the NHS DR classification guideline (NHSDRCG) rather than the International Clinical Diabetic Retinopathy Severity Scale (ICDRSS).[22] In

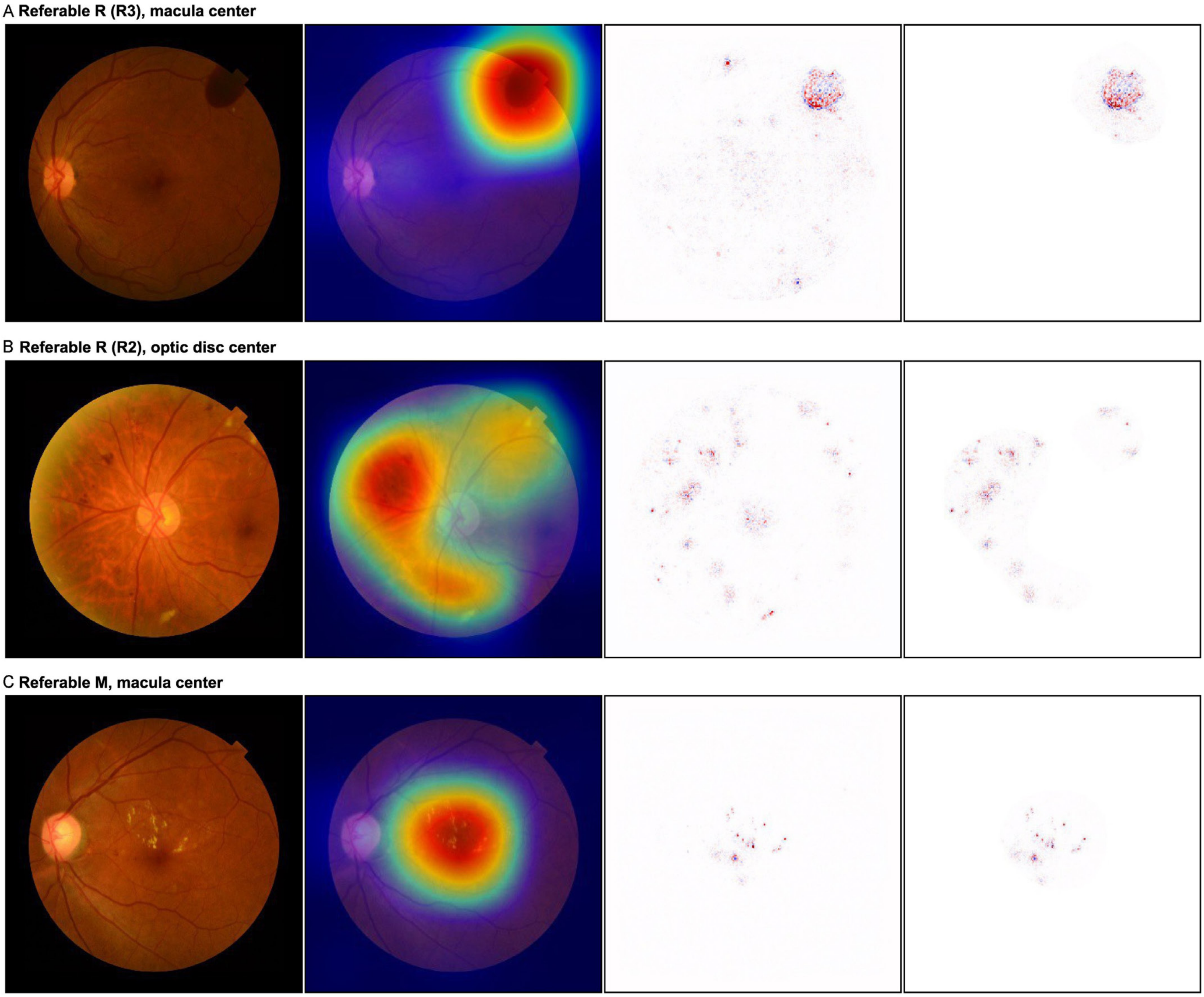

A Referable R (R3), macula center
B Referable R (R2), optic disc center
C Referable M, macula center

Figure 2 Visualisation by the SHAP-CAM heatmap technique for referable DR lesions. The original images are displayed in the first column, the combined heatmaps generated by SHAP-CAM are shown in the last column, and the heatmaps by CAM and DeepSHAP are shown in the second and third columns for comparison, respectively. (A) Vitreous haemorrhage located on the temporal-superior retina of the original image with the centred macula, suggesting the R3 degree of DR. The CAM heatmap showed a rough location as a wide red-cyan area for the lesion, while the DeepSHAP heatmap demonstrated dispersed dots with some irrelevant noises. The SHAP-CAM heatmap retained an light pink background area, with similar size as that of CAM, and depicted a deeper red clear lesion, same as that of DeepSHAP, in the background. The residue area was masked by CAM as white to reduce inference of redundant information. (B) Retinopathy of R2, including venous beading, intraretinal microvascular abnormality and multiple blot haemorrhages, located around the optic disc on original images. The CAM heatmap showed a rough area for detection, whereas the DeepSHAP heatmap indicated the optic disc as a lesion. For the SHAP-CAM heatmap, all key lesions are depicted in the accurate light pink area without involving the optic disc and macula. (C) The original image showed a referable maculopathy with multiple exudates involving the centre of the fovea. The SHAP-CAM heatmap accurately predicted the shape/outline of the lesions in the macula area, whereas CAM only visualised the lesions by a wide red-cyan circle area and the DeepSHAP showed several light noises out of the macula. DR, diabetic retinopathy.

previous studies, referable DR was defined as moderate and worse DR and diabetic macular edema (DME) or both, where patients with retinopathy that is more severe than mild (defined as the presence of microaneurysms only) would be referred to ophthalmologists. Yet there is still no effective management currently available for patients with an early stage of DR. These patients could only be monitored annually, but not referred to retinal specialists.[23 24] Over-referral could result when adopting the clinical criteria for referable DR screening, increasing the workload on eye care services and the financial burden associated with the DR screening programmes. The ICDRSS is based on the clinical fundus examination of each quadrant.[22] However, only one or two 45° fields of retinal images were taken for DR grading during the DR screening programme.[12 15 25 26] This would lead

Table 3 Performance comparison between the system and three DR experts

Dimension	Reader	n				Accuracy	F1 score	Sensitivity	Specificity	AUROC (95% CI)
		TN	FP	FN	TP					
Referable retinopathy	Expert 1	181	4	5	62	0.964	0.932	0.925	0.978	NA
	Expert 2	181	4	6	61	0.960	0.924	0.910	0.978	NA
	Expert 3	183	2	2	65	0.984	0.97	0.970	0.989	NA
	Experts' average	NA	NA	NA	NA	0.970	0.942	0.935	0.982	NA
	DLA	176	9	0	67	0.964	0.937	1.000	0.951	0.9958 (0.9916 to 1.0000)
Referable maculopathy	Expert 1	171	3	7	71	0.960	0.934	0.910	0.983	NA
	Expert 2	167	7	4	74	0.956	0.931	0.949	0.960	NA
	Expert 3	166	8	4	74	0.952	0.925	0.949	0.954	NA
	Experts' average	NA	NA	NA	NA	0.956	0.93	0.936	0.966	NA
	DLA	163	11	4	74	0.940	0.908	0.949	0.937	0.9877 (0.9756 to 0.9999)
**Referable DR	Expert 1	164	3	7	78	0.960	0.94	0.918	0.982	NA
	Expert 2	161	6	6	79	0.952	0.929	0.929	0.964	NA
	Expert 3	160	7	4	81	0.956	0.936	0.953	0.958	NA
	Experts' average	NA	NA	NA	NA	0.956	0.935	0.933	0.968	NA
	DLA	163	4	4	81	0.968	0.953	0.953	0.976	0.9909 (0.9809 to 1.0000)

NA, not applicable.
*By integrating the prediction of referable retinopathy and maculopathy on an image, referable DR decisions were given by the system when any referable lesion is detected.
AUROC, area under the receiver operating characteristic curve; DLA, deep learning algorithm; DR, diabetic retinopathy; FN, false negative; FP, false positive; TN, true negative; TP, true positive.

to inaccurate classification of DR or confuse the graders when the grading is based only on one or two fundus photos. In contrast, the NHSDRCG was specially developed for DR screening and has been used for years in different national DR screening programmes, including the Lifeline Express DR programme in China. For the NHSDRCG, the classification is based on multidimensional features of DR lesions, rather than on the most severe DR lesion. Moreover, our system could provide a referable decision at the eye-level by intergrating all image-level decisions of one eye, as well as provide the patient-level decision by combining the results of the two eyes. Multiple DL algorithms have been developed to detect referable or vision-threatening DR, with robust performance in previous studies.[8–10] Although these studies achieved high accuracy, they were designed predominately focusing on a general classification of referable DR. In daily DR screening practice, complex conditions can be found and need to be handled. The NHSDRCG should be more suitable to support the development of a multidimensional system.

The two dimensions of quality evaluation in our system, namely image quality and maculopathy gradability, are more consistent with screening practices. First, when the fundus photos are sent to the reading centre, the image quality evaluation should precede the classification of severity of DR. Poor image quality could be due to the opacity of the refractive media, artefacts, poor contrast, defocus or small pupil.[27] Previous studies assigned these ungradable pictures to referable DR,[9 10 28–30] which can cause unnecessary worries to patients and confuse the graders on their judgement of referable DR or rephotography. Second, maculopathy gradability should be evaluated before grading the maculopathy. Although some fundus image qualities meet the gradable criteria, the macular area might not be seen due to blur or opacity in the area. Third, our platform could provide the grading outcome of maculopathy alone instead of combining the results of retinopathy and maculopathy. Therefore, we could obtain the basis of referral suggestion, caused by retinopathy or by maculopathy. Since DME could now be treated in most primary medical units or hospitals with antivascular endothelial growth factor, referral to senior special hospitals to receive vitrectomy or photocoagulation therapy might not be necessary.[31 32]

Photocoagulation status on the retinal images received attention by the NHSDRCG. Its corresponding model was also established in our system, which could judge whether patients have ever received photocoagulation therapy by detecting laser spots on the fundus photos. Laser spots indicate patients have received photocoagulation therapy before screening and the treatment suggested to these patients would be different from other cases.

The SHAP-CAM heatmap highlights the predictive referable DR lesions on retinal fundus pictures. Generally, CAM could indicate the proper size but a less precise domain for lesion identification. In contrast, DeepSHAP could depict specific fine lesions,[33] but more dispersed. The combination of the two techniques can provide a heatmap of lesions in specific domains, meeting the requirement for distinguishing maculopathy from retinopathy. These visualisations provide explainability and improve the accuracy of and confidence in DR grading.[34 35]

Limitations

There were several limitations to this study. First, similar to other studies, DME was graded on non-stereoscopic images according to the presence of hard exudate, microaneurysm or haemorrhage in the macular area. This could be misdiagnosed in some cases without stereoscopic images and optical coherence tomography.[36] Second, the lesions, which could be tiny or less frequent, such as intraretinal microvascular abnormality and vein beading, might not be detected well on the images. More data presenting these lesions need to be trained if fine-grained classification than basic screening is required. Third, only two classes (referable and non-referable) and relevant indicators were adopted in the study. Besides, limited by the retrospective data with some missing information, stratified analysis of the classification performance of the DL system based on age, duration of diabetes and various devices, which could influence image quality, was not conducted. A prospective study of multiclass classification (ie, DR 0–5) and multifactor analyses could be carried out in the future. Additional indicators suitable for multiclass classification (ie, weighted kappa) could also be applied.

CONCLUSIONS

This study demonstrated that our DL platform based on a real-world DR screening guideline achieved high sensitivity and specificity with multidimensional classifiers, indicating that AI tools could assist in large-scale screening of referable DR in primary medical units.

Contributors GZ and MZ proposed and designed the study. GZ, JJ, J-WL, WC, PX, YZ, YX, HW and DL collected the data and/or executed the research. GZ, J-WL and JW analysed the data. GZ, JW, J-WL, JJ and TKN prepared the manuscript. L-PC and CPP gave critical revision. MZ is responsible for the overall content as guarantor.

Funding This work was supported by the Shantou Medical Health, Science and Technology Project Fund (project code: 190716155262406) and the Grant for Key Disciplinary Project of Clinical Medicine under the Guangdong High-Level University Development Program, China. The funding bodies were not involved in the study, including in the collection, analysis and interpretation of data and in writing the manuscript.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval This study was approved by the Human Medical Ethics Committee of the Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong (EC20190612(3)-P10), which is in accordance with the Declaration of Helsinki.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Both the main text and supplementary materials are published.

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
Ji Wang http://orcid.org/0000-0001-6997-3837
Weiqi Chen http://orcid.org/0000-0003-2852-8192
Tsz Kin Ng http://orcid.org/0000-0001-7863-7229
Mingzhi Zhang http://orcid.org/0000-0001-9032-7274

REFERENCES

1 Flaxman SR, Bourne RRA, Resnikoff S, *et al*. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. *Lancet Glob Health* 2017;5:e1221–34.
2 Saeedi P, Petersohn I, Salpea P, *et al*. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res Clin Pract* 2019;157:107843.
3 Teo ZL, Tham Y-C, Yu M, *et al*. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology* 2021;128:1580–91.
4 Bachmann MO, Nelson SJ. Impact of diabetic retinopathy screening on a British district population: case detection and blindness prevention in an evidence-based model. *J Epidemiol Community Health* 1998;52:45–52.
5 Wormald R, Courtney P. Prevention of blindness by screening for diabetic retinopathy. *BMJ* 1989;299:1528.
6 Wang LZ, Cheung CY, Tapp RJ, *et al*. Availability and variability in guidelines on diabetic retinopathy screening in Asian countries. *Br J Ophthalmol* 2017;101:1352–60.
7 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
8 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.
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 Cen L-P, Ji J, Lin J-W, *et al*. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun* 2021;12:4828.
12 Zhang G, Chen H, Chen W, *et al*. Prevalence and risk factors for diabetic retinopathy in China: a multi-hospital-based cross-sectional study. *Br J Ophthalmol* 2017;101:1591–5.
13 Lian JX, Gangwani RA, McGhee SM, *et al*. Systematic screening for diabetic retinopathy (DR) in Hong Kong: prevalence of DR and visual impairment among diabetic population. *Br J Ophthalmol* 2016;100:151–5.
14 Harding S, Greenwood R, Aldington S, *et al*. Grading and disease management in national screening for diabetic retinopathy in England and Wales. *Diabet Med* 2003;20:965–71.
15 Scanlon PH. The English national screening programme for diabetic retinopathy 2003-2016. *Acta Diabetol* 2017;54:515–25.
16 Zhou B, Khosla A, Lapedriza A. Learning deep features for discriminative localization, 2015. Available: https://arxiv.org/abs/1512.04150 [Accessed 27 Feb 2021].
17 Selvaraju RR, Cogswell M, Das A. Grad-CAM: visual explanations from deep networks via gradient-based localization, 2016. Available: https://arxiv.org/abs/1610.02391 [Accessed 5 Apr 2020].
18 Lundberg S, Lee S-I. A unified approach to interpreting model predictions, 2017. Available: https://arxiv.org/abs/1705.07874 [Accessed 3 Jan 2020].

19 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.

20 Wongpakaran N, Wongpakaran T, Wedding D, *et al*. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol* 2013;13:61.

21 Methods and applications of statistics in clinical trials, Volume 2: planning, analysis, and inferential methods. Hoboken, New Jersey. John Wiley & Sons, Inc 2014.

22 Wilkinson CP, Ferris FL, Klein RE, *et al*. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003;110:1677–82.

23 Scanlon PH. Screening intervals for diabetic retinopathy and implications for care. *Curr Diab Rep* 2017;17:96.

24 Group ETDRSR. Fundus photographic risk factors for progression of diabetic retinopathy. *Ophthalmology* 1991;98:823–33.

25 Zheng Y, Lamoureux EL, Lavanya R, *et al*. Prevalence and risk factors of diabetic retinopathy in migrant Indians in an urbanized society in Asia: the Singapore Indian eye study. *Ophthalmology* 2012;119:2119–24.

26 Pandey R, Morgan MM, Murphy C, *et al*. Irish national diabetic RetinaScreen programme: report on five rounds of retinopathy screening and screen-positive referrals. (INDEAR study report No. 1). *Br J Ophthalmol* 2022;106:409–14.

27 Scanlon PH, Malhotra R, Thomas G, *et al*. The effectiveness of screening for diabetic retinopathy by digital imaging photography and technician ophthalmoscopy. *Diabet Med* 2003;20:467–74.

28 Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 2017;124:962–9.

29 Rêgo S, Dutra-Medeiros M, Soares F, *et al*. Screening for diabetic retinopathy using an automated diagnostic system based on deep learning: diagnostic accuracy assessment. *Ophthalmologica* 2021;244:250–7.

30 Wang Y, Yu M, Hu B, *et al*. Deep learning-based detection and stage grading for optimising diagnosis of diabetic retinopathy. *Diabetes Metab Res Rev* 2021;37:e3445.

31 Diabetic Retinopathy Clinical Research Network, Wells JA, Glassman AR, *et al*. Aflibercept, bevacizumab, or ranibizumab for diabetic macular edema. *N Engl J Med* 2015;372:1193–203.

32 Patrao NV, Antao S, Egan C, *et al*. Real-World outcomes of ranibizumab treatment for diabetic macular edema in a United Kingdom National health service setting. *Am J Ophthalmol* 2016;172:51–7.

33 Wang J, Ji J, Zhang M, *et al*. Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening. *JAMA Netw Open* 2021;4:e218758.

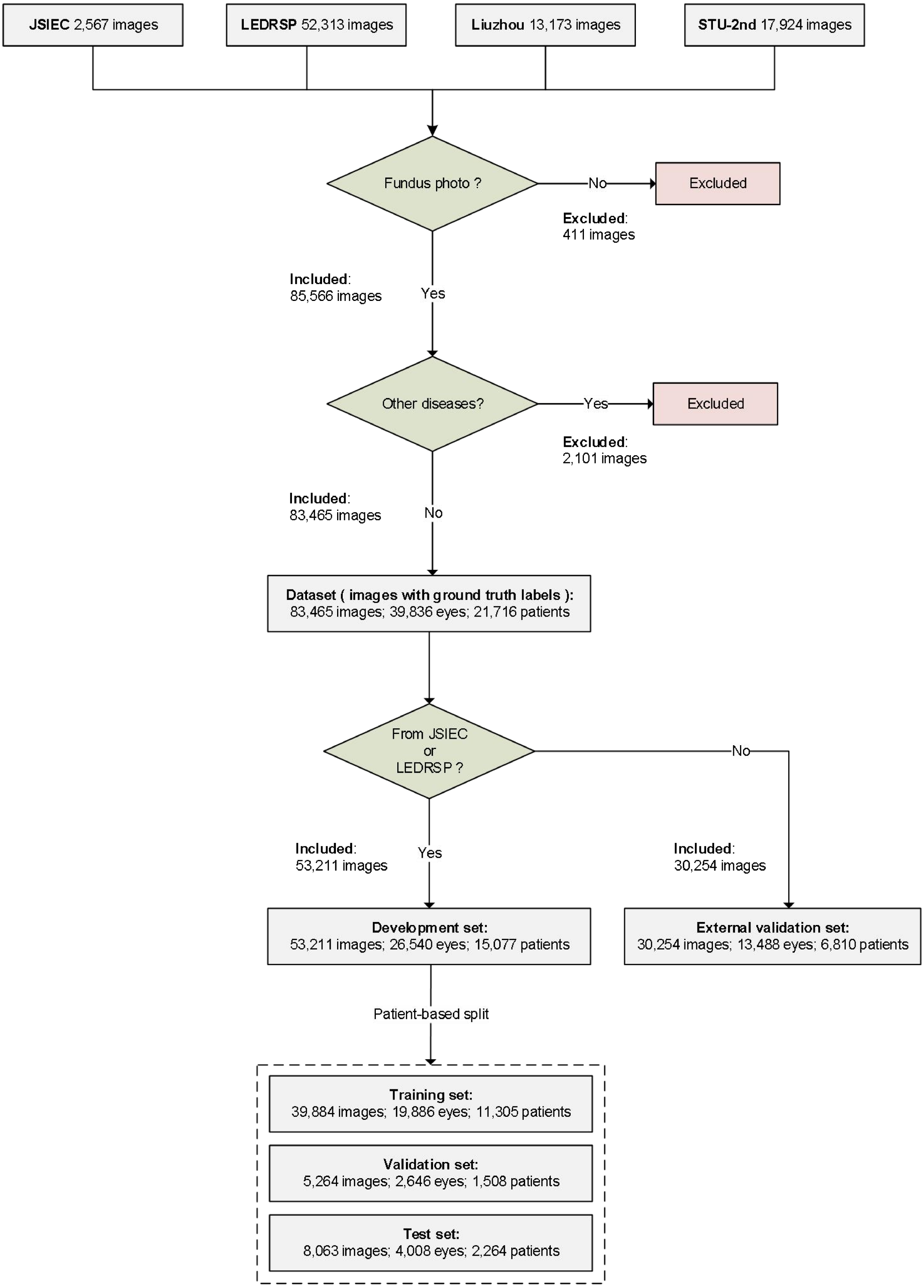
34 Keel S, Wu J, Lee PY. Visualizing deep learning models for the detection of referable diabetic retinopathy and glaucoma. *JAMA Ophthalmol* 2018.

35 Sayres R, Taly A, Rahimy E, *et al*. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* 2019;126:552–64.

36 Wong RL, Tsang CW, Wong DS, *et al*. Are we making good use of our public resources? the false-positive rate of screening by fundus photography for diabetic macular oedema. *Hong Kong Med J* 2017;23:356–64.

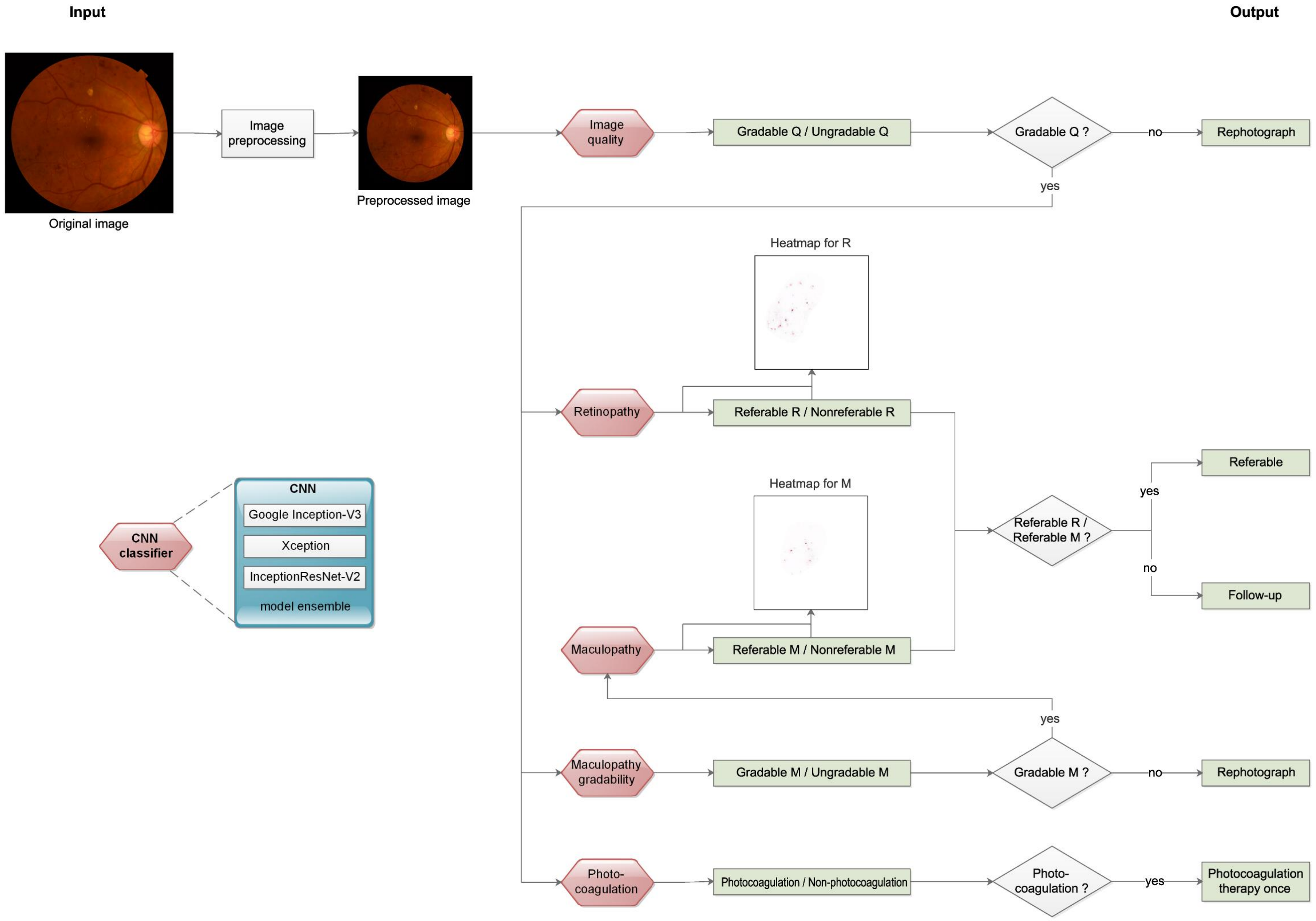
Supplementary materials

Online Content


JSIEC 2,567 images
LEDRSP 52,313 images
Liuzhou 13,173 images
STU-2nd 17,924 images
Fundus photo ?
No
Excluded
Excluded: 411 images
Included: 85,566 images
Yes
Other diseases?
Yes
Excluded
Excluded: 2,101 images
Included: 83,465 images
No
Dataset (images with ground truth labels): 83,465 images; 39,836 eyes; 21,716 patients
From JSIEC or LEDRSP ?
No
Included: 53,211 images
Yes
Included: 30,254 images
Development set: 53,211 images; 26,540 eyes; 15,077 patients
External validation set: 30,254 images; 13,488 eyes; 6,810 patients
Patient-based split
Training set: 39,884 images; 19,886 eyes; 11,305 patients
Validation set: 5,264 images; 2,646 eyes; 1,508 patients
Test set: 8,063 images; 4,008 eyes; 2,264 patients

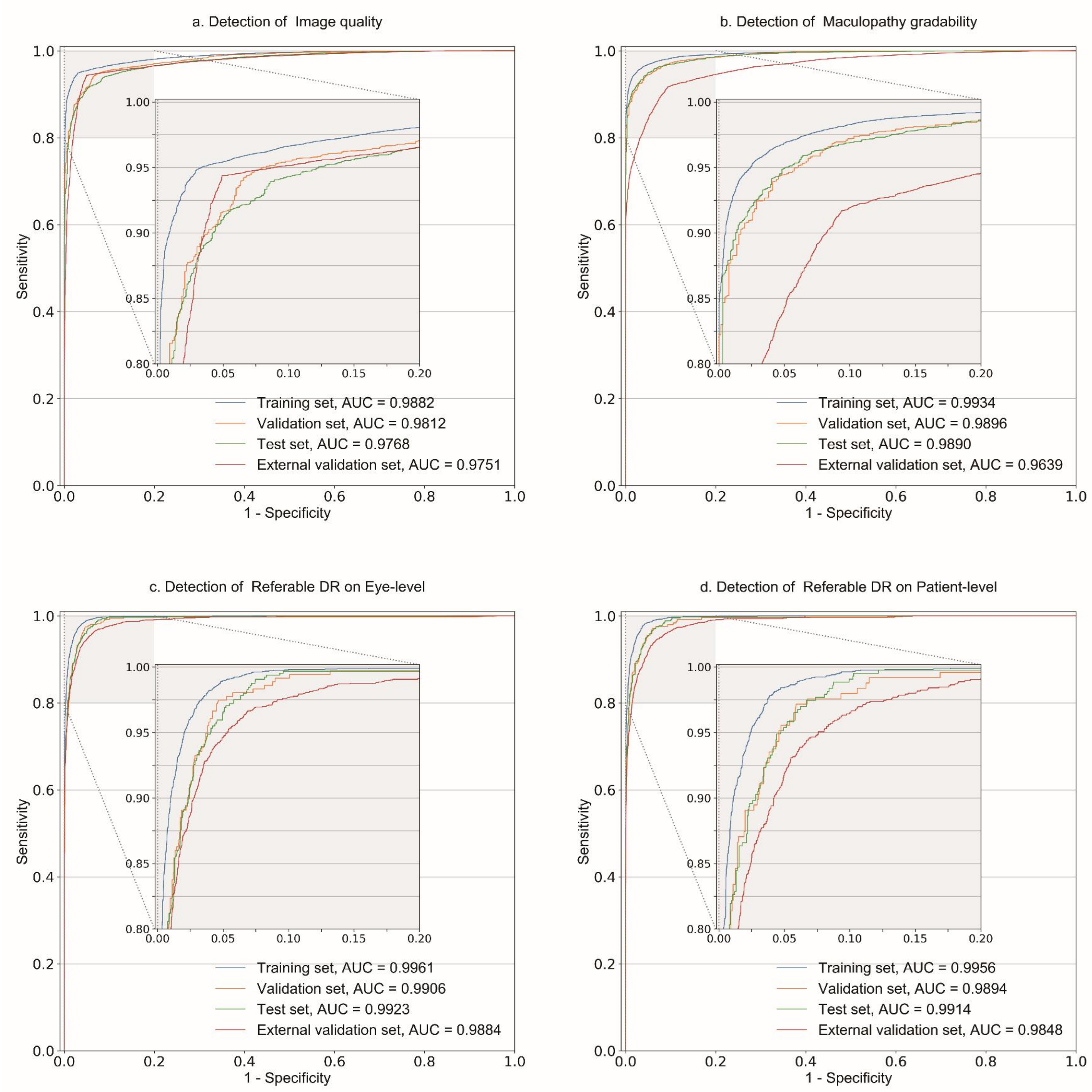
Supplemental figure 1. Workflow of retinal images database construction

85977 retinal images from 4 centers were collected initially. After cleaning, 83465 images were included and annotated as final database. Of them, 53211 (63.8%) images were used as development set and further split into training, validation and test subsets, whereas 30254 ones (36.2%) were as external validation set.


Input
Output
Original image
Image preprocessing
Preprocessed image
Image quality
Gradable Q / Ungradable Q
Gradable Q ?
no
Rephotograph
yes
Heatmap for R
Retinopathy
Referable R / Nonreferable R
Referable
yes
Heatmap for M
Referable R / Referable M ?
no
Follow-up
CNN classifier
CNN
Google Inception-V3
Xception
InceptionResNet-V2
model ensemble
Maculopathy
Referable M / Nonreferable M
yes
Maculopathy gradability
Gradable M / Ungradable M
Gradable M ?
no
Rephotograph
Photo-coagulation
Photocoagulation / Non-photocoagulation
Photo-coagulation ?
yes
Photocoagulation therapy once

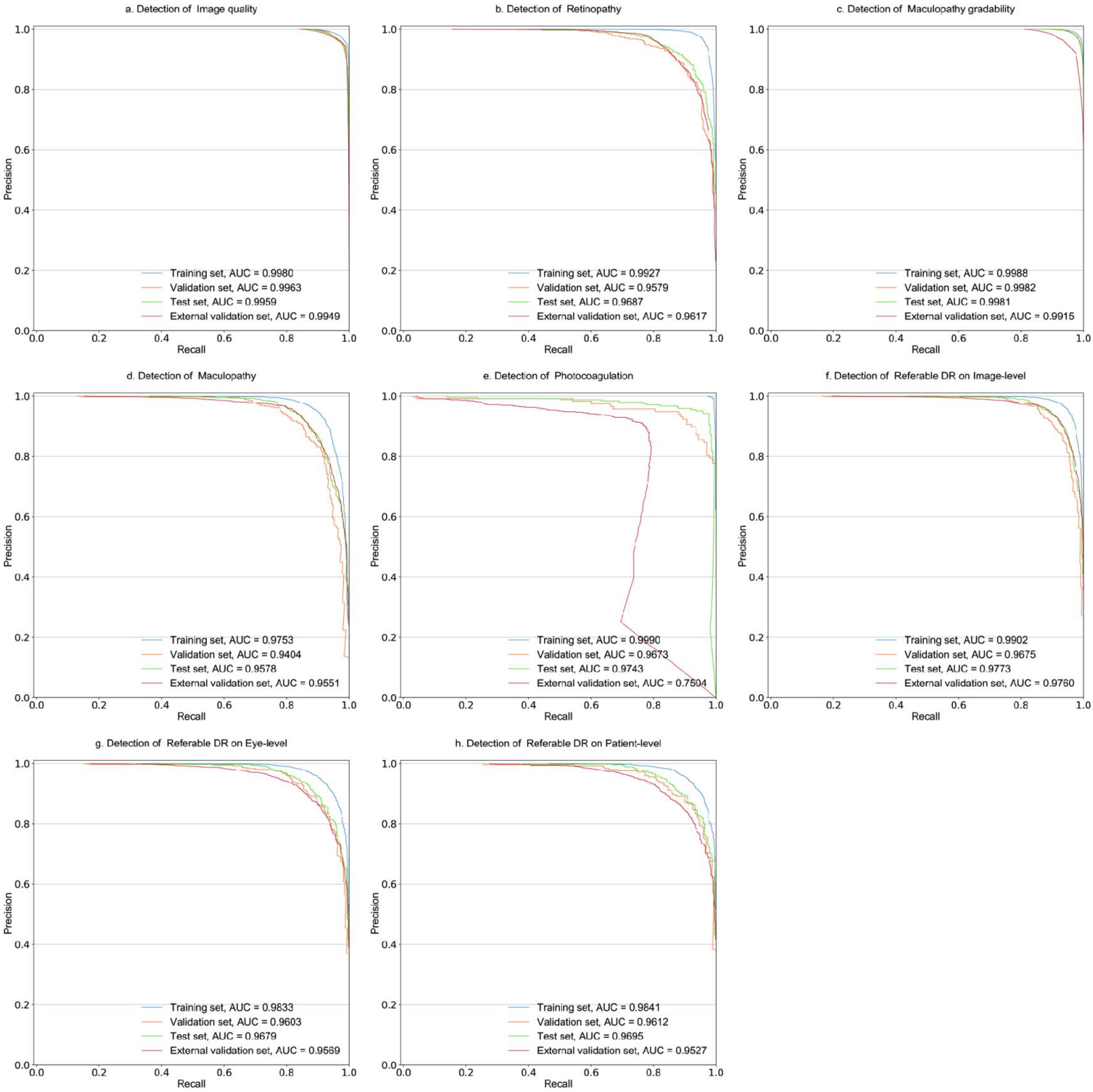
Supplemental figure 2. The pipeline of referable diabetic retinopathy screening system

A deep learning ensemble model of three single models, including Google Inception-V3, Xception and InceptionResNet-V2, was developed. In the model, there are 5 independent classifiers (image quality, retinopathy, maculopathy grabability, maculopathy and photocoagulation) to identify 5 dimensions of a given retinal image, respectively. The image quality is the first evaluated dimension, and the gradable quality images would be transmitted to next classifiers. For decreasing the false classifications due to limited blur and artifacts on macula, the maculopathy gradability should be processed before prediction of referable maculopathy, Any predictived referable lesion, such as referable retinopathy and referable maculopathy will results in the automated recommendation of “referable”. Any laser spot scar on retina suggested the previous photocoagulation therapy, and the corresponding patient would be recommended refer to previous ophthalmologist. The image of ungradable quality or ungradable maculopathy should be rephotographed. The heatmaps generated by SHAP-CAM, combining Class Activation Mapping (CAM) [1,2]and DeepSHAP, would be provided for any positive prediction of retinopathy or maculopathy. Abbreviation: M, maculopathy; R, retinopathy; Q, quality.



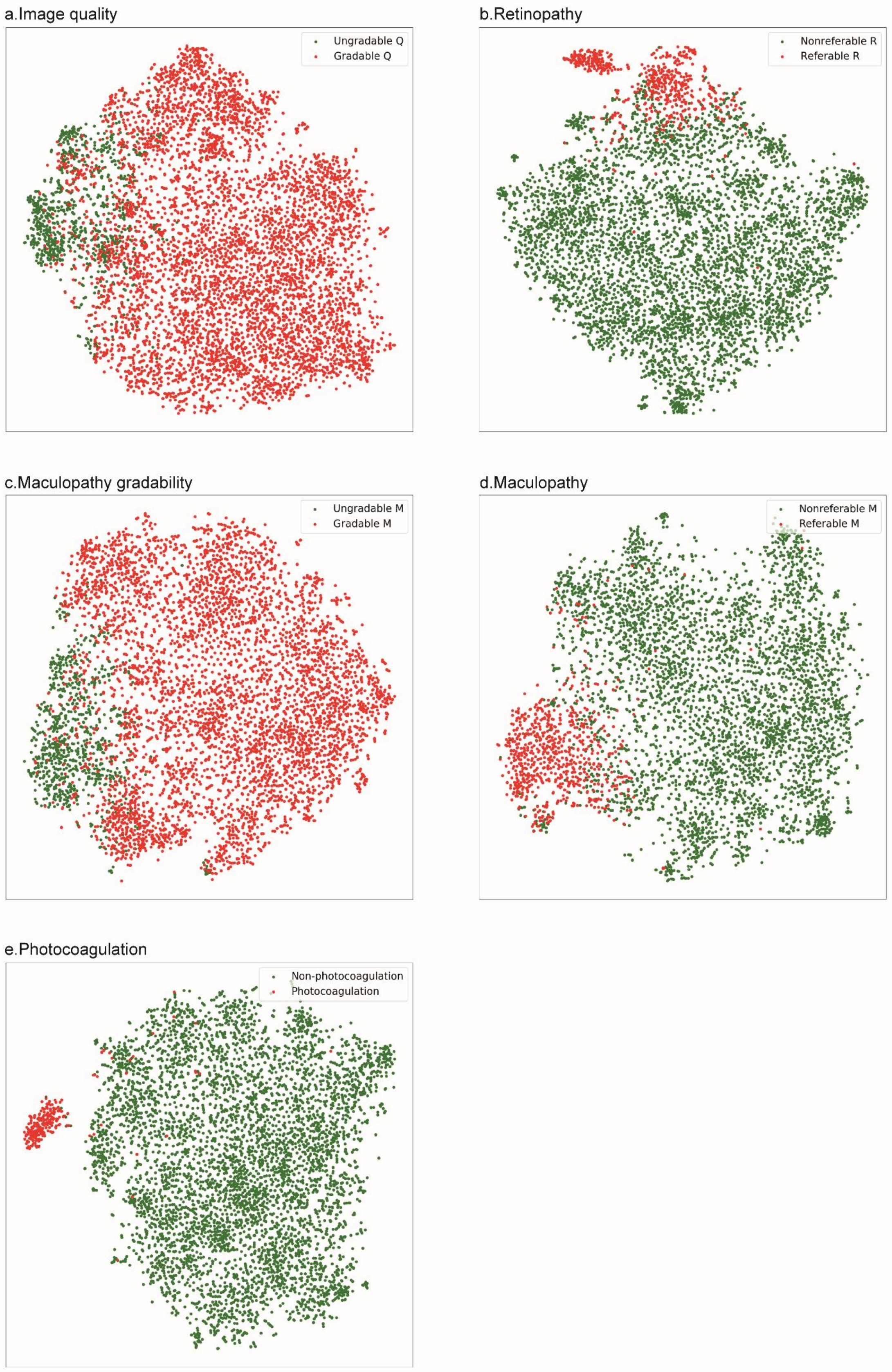
Supplemental figure 3. The receive operating characteristic (ROC) curves for system performance

The ROC and area under curve (AUC) for detecting gradable image quality (upper left) and maculopathy gradability (upper right) was shown in each set. The referable diabetic retinopathy (DR) detection on image-level was automate generated from integrating the multi-dimension classifications of an image by deep learning system. The referable DR on eye- and patient-level were automatically generated from integrating the results of all the images per eye and per patient, respectively. The ROC and AUC on eye- (lower left) and patient-level (lower right) of each set were plotted accordingly.



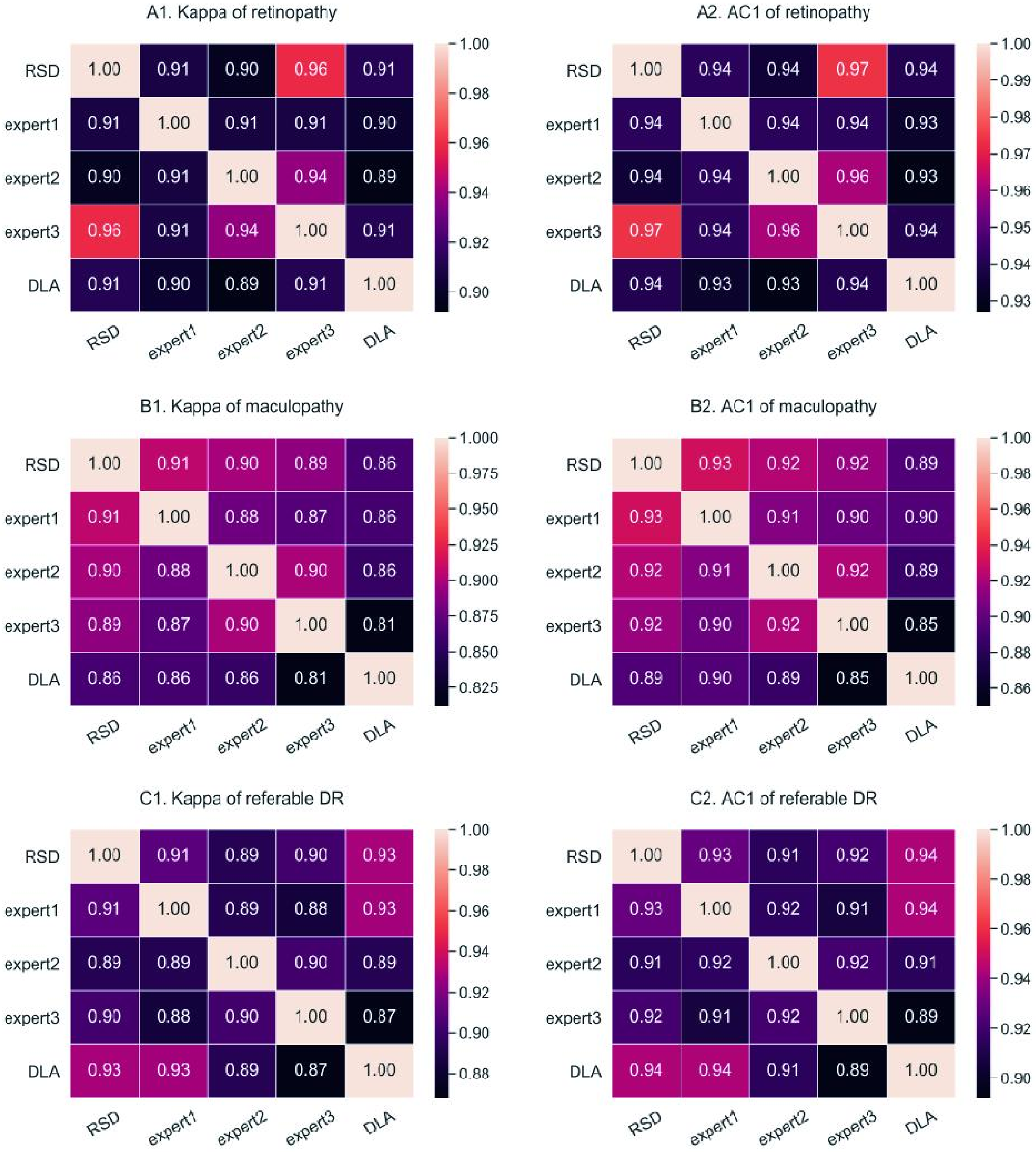
Supplemental figure 4. The precision-recall curve (PRC) curves for system performance

The PRC and the area under curve (AUC) for detecting 5 dimensions were plotted (a-e). The referable diabetic retinopathy (DR) detection on image-level was automate generated from integrating the multi-dimension classifications of an image by deep learning system. The referable DR on eye- and patient-level were automatically generated from integrating the results of all the images per eye and per patient, respectively. The PRC and AUC on image-, eye- and patient-level of each set were plotted accordingly (f-h).



Supplemental figure 5. Visualization by the t-distributed stochastic neighbor embedding (t-SNE) of 5 classifiers

On a t-SNE map, each point represents a sample, when different colors represent different classes. Well separation between binary classes of each classifier was shown in t-SNE map, which visualizing the potential pattern of features extraction from neural networks.



Supplemental figure 6. The consistency heat-map for human-machine comparison

The Cohen's unweighted K values (left column) and Gwet's AC1(right column) were calculated for evaluating the consistency of graders with reference standard diagnosis. Three dimensional detection, including referable retinopathy **(A)**, referable maculopathy **(B)** and referable diabetic retinopathy **(C)**, were involved in the comparison. Deep learning algorithm showed the comparable performance with three human experts. Abbreviations: RSD, reference standard diagnosis; DLA, deep learning algorithm.

Supplemental table 1. Summary of DR grading protocol in National Guidelines on Screening and the management of cases post-grading[1,2]

Dimension (abbreviation)	Level/scale	Definition	Recommendation/Management
Retinopathy (R)	R0	No any DR	Annual screening
	R1	Background phase of DR, including microaneurysm(s), retinal haemorrhage(s), venous loop(s), or any above feature coexisting with the presence of any exudate or any number of cotton wool spots	Annual screening
	R2	Preproliferative phase of DR, including venous beading, venous reduplication, multiple blot hemorrhages or IRMA	Refer to hospital eye service
	R3	Proliferative phase of DR, including the feature of new vessels on disc, new vessels elsewhere, pre-retinal or vitreous hemorrhage, or pre-retinal fibrosis with/without tractional retinal detachment	Fast-track referral to hospital eye service
Maculopathy (M)	M0	Absence of any M1 features	Annual screening (R0M0 or R1M0)
	M1	Exudate within 1 disc diameter (DD) of the centre of the fovea; Circinate or group of exudates within the macula; Retinal thickening within 1 DD of the centre of the fovea (if stereo available); Any microaneurysm or haemorrhage within 1 DD of the centre of the fovea only if associated with a best VA of ≤ (if no stereo) 6/12	Refer hospital eye service
Photocoagulation (P)	P0	No evidence of previous photocoagulation	/
	P1	Focal/grid to macula	New screenee→refer hospital eye service Quiescent post treatment→annual screening
Other lesions (OL)	/	The lesions other than DR (e.g., cataract, glaucoma or age-related macular degeneration)	Refer to hospital eye service or inform primary physician
Ungradable/unobtainable (U)	/	An image set that is inadequate for grading*	Poor view but gradable on biomicroscopy→refer hospital eye service; Unscreenable→discharge, inform general practitioner (option to recall for further photos if purely technical failure)

*Ungradable/unobtainable images in photography (usually due to media opacity such as cataract or occasionally severe asteroid hyalosis; no clinical examination in optometry-based programmes) should be referred directly for secondary assessment and classified as U.Abbreviations: DR, diabetic retinopathy.

Supplemental table 2. Definitions of dimensions/labels and corresponding recommendation/management in the study

Dimension (abbreviation)	Image field center	Level/scale	Definition	Recommendation/Management
Image quality (Q)	Optic disc or macula	Q0	Ungradable image quality: more than 1/3 area of the image due to poor exposure, artifact or blur cannot be classified confidently, even if any DR feature is observed in other area	Rephotograph
		Q1	Gradable image quality: image is classifiable with confidence	Step into the main classification pipeline
Retinopathy (R)	Optic disc	R0	No any DR	Follow-up
		R1	Background phase of DR, including microaneurysm(s), retinal haemorrhage(s), venous loop(s), or any above feature coexisting with the presence of any exudate or any number of cotton wool spots	Follow-up
		R2	Preproliferative phase of DR, including venous beading, venous reduplication, multiple blot hemorrhages or IRMA	Referable
		R3	Proliferative phase of DR, including the feature of new vessels on disc, new vessels elsewhere, pre-retinal or vitreous hemorrhage, or pre-retinal fibrosis with/without tractional retinal detachment	Referable
Maculopathy (M)	Macula	Mu	Maculopathy ungradable due to the limited blur or artifact	Referrable (if the severity of retinopathy requires referral)
				Maculopathy ungradable (No other evidences support the referral)
		M0	Absence of any M1 features	Follow-up
		M1	Exudate within 1 disc diameter (DD) of the centre of the fovea; any microaneurysm or haemorrhage within 1DD of the centre of the fovea only if associated with a best VA of ≤6/12	Referable
Photocoagulation (P)	Optic disc or macula	P0	No scar of laser spot observed	No recommendation
		P1	Presenting laser spot or scar	Refer to previous ophthalmologist

Abbreviations: DR, diabetic retinopathy; intraretinal microvascular abnormality, IRMA; VA, visual acuity.

Supplemental table 3. The distributions of images with various labels and conditions

					Images, No.				
Condition	Q	R	M	P	JSIEC	LEDRSP	Liuzhou	STU-2nd	Total
Ungradable Q	Q0	-	-	-	13	8,051	1,474	3,475	13,013
Nonreferable R, Ungradable M	Q1	R0	Mu	P0	83	4,530	2,617	390	7,620
	Q1	R0	Mu	P1	0	7	4	1	12
	Q1	R1	Mu	P0	0	1,480	336	738	2,554
	Q1	R1	Mu	P1	0	1	1	0	2
Referable R, Ungradable M	Q1	R2	Mu	P0	1	235	76	331	643
	Q1	R2	Mu	P1	0	12	7	1	20
	Q1	R3	Mu	P0	0	250	6	164	420
	Q1	R3	Mu	P1	0	296	30	123	449
Nonreferable R, Nonreferable M	Q1	R0	M0	P0	2,453	25,239	6,848	7,616	42,156
	Q1	R0	M0	P1	0	28	4	0	32
	Q1	R1	M0	P0	2	4,704	737	2,032	7,475
	Q1	R1	M0	P1	0	3	0	5	8
Nonreferable R, Referable M	Q1	R0	M1	P0	14	139	5	17	175
	Q1	R0	M1	P1	0	5	0	0	5
	Q1	R1	M1	P0	1	2,205	359	680	3,245
	Q1	R1	M1	P1	0	0	1	2	3
Referable R, Nonreferable M	Q1	R2	M0	P0	0	149	22	93	264
	Q1	R2	M0	P1	0	68	49	0	117
	Q1	R3	M0	P0	0	33	1	59	93
	Q1	R3	M0	P1	0	412	21	72	505
Referable R, Referable M	Q1	R2	M1	P0	0	1,666	209	1,090	2,965
	Q1	R2	M1	P1	0	14	77	7	98
	Q1	R3	M1	P0	0	348	13	305	666
	Q1	R3	M1	P1	0	769	1	155	925

Abbreviations: Q, image quality; R, retinopathy ; M, maculopathy; P, photocoagulation

Supplemental table 4. The possible reasons of the false prediction by system in external validation set

Referable retinopathy	n (%)
False positive	
Total	784 (100)
Background DR	646 (82.4)
Artifacts	58 (7.4)
Changes of fundus pigment	19 (2.4)
Retinopathy other than DR	61 (7.8)
AMD	17 (2.2)
RVO	2 (0.3)
Retinal detachment	2 (0.3)
Others	40 (5.1)
False negative	
Total	65 (100)
Limited blurred images	25 (38.5)
IRMA	15 (23.1)
Blot hemorrhage	12 (18.5)
Venous beading	5 (7.7)
Small preretinal hemorrhage	4 (6.2)
Questionable new vessels	2 (3.1)
Small membrane	2 (3.1)

Referabe maculopathy	n (%)
False positive	
Total	572 (100)
H/M in macula with BCVA>0.5	178 (31.1)
Drusens	154 (26.9)
Artifacts	118 (20.6)
AMD	59 (10.3)
DR lesion located outside 1DD of fovea	55 (9.6)
mERM	8 (1.4)
False negative	150 (100)
Tiny H/M	70 (46.7)
Tiny hard exudates	47 (31.3)
Limited blurred images	33 (22.0)

Abbreviations: AMD, age-related macular degeneration; DR, diabetic retinopathy; H/M, hemorrhage or icroaneurysm; IRMA, Intraretinal microvascular abnormalities; mERM, macular epiretinal membrane.

Supplemental method 1. Deep learning algorithm development

1. Image preprocessing

Image preprocessing is the first step because the image resolution of the original image is different and too large to load into neural networks, and the original image usually contains large black areas. The black background areas were cropped using a threshold method, followed by converting the image into square by adding black paddings. To avoid deleting meaningful areas during the image augmentation process, some black areas (5% of the side length of the image square) were added to the borders of the fundus images. After that, the image was resized to 384*384 pixels

2. Neural network models

Even though there exist only two classes for every dimension, the multi-class classification was used instead of the binary classification because in the future we will add more classes for DR and DME. So softmax was used as the last layer's activation function, and categorical cross-entropy as the loss function. Ensemble learning was best suited for models that are high accurate and different, so different kind of neural networks were used. A simple unweighted average (a kind of soft voting method) was used to combine results of multiple models, and it will be discussed in detail in the **Prediction process section.** Inception-V3[3], Xception[4] and InceptionResNet-V2[5] were used as base models. It is not only because these models were widely used in medical image analysis but also because in our pre-experiments they performed no worse than other more advanced models such as EfficientNet-V2, Regnet and Vision Transformer.

3. Real-time Image Augmentation

In order to enlarge the samples size and improve the generalization ability of the model, image augmentation was used during training[6]. Compared with image augmentation before training, the real time implementation not only save time but also is more flexible. Both geometry transformations and lightness and color transformations were used in image augmentation. Specifically, the images were randomly rotated (range: [-15◦, 15◦]), translated (range:[-10%,10%]), scaled (range: [95%,105%]), horizontally and vertically flipped, and image contrast were modified (multiplicative factor range:[90%,110%]).

4. Training

The dynamic data re-sampling[7,8] was used to tackle the class imbalance problem. These models were initialized using the corresponding ImageNet models[9], and then all layers were fine-tuned. Adam[10] was used as the optimizer. The number of epochs was set to 15. The initial learning rate was set to 0.001, and multiplied by a factor of gamma=0.3 after every 2 epoch. During every training, the model with the minimum validation loss was chosen as the best model. During experiments, performances were not sensitive to these hyper-parameters.

5. Prediction process

Given an image, it will be classified by 5 classifiers independently and every classifier contains 3 models. Unweighted average (a kind of soft voting method) was used to combine the results of multiple models. The ensemble learning would generate a more accurate prediction than single model.[11]

The formulas of the unweighted average algorithm are as follows:

$$\text{probs}_j = \frac{\sum_{i=1}^{N} (W_i \times p_{ij})}{\sum_{i=1}^{N} W_i}$$

pred_class = probs.argmax(axis=-1)

The number of base models is denoted by N, and Wi is the weight of the model No. i. p_{ij} is the predicted probability of model i for class number j. For simplicity, instead of being learned by a meta-learner[12], Wi is set to 1 for all models(unweighted ensemble). probs_j is the predicted probability for class i after model ensemble and pred_class is the final predicted class.

For an image, if it is predicted as positive for at least one class of DR, DME, the image will receive a referral result. If at least one image of a eye is referral, the result of the eye is referral. Likewise, If at least one eye of a patient is referral, the result of the patient is referral.

Supplemental method 2. Visualizing and explaining CNNs

t-SNE[13], which is a non-linear dimensionality reduction technique, was used to show the discrimination of neural networks by visualizing the distribution of features extracted by the neural network. High dimensional features were converted to two dimensional data and then a scatter plot was drawn using it. In the t-SNE map, every point stands for a sample in the dataset. The Sklearn.manifold.t-SNE library was used to process the data, and the Matplotlib library was used to draw scatter plot images.

The explainability of neural networks was very important, unfortunately， all current explanation methods were fragile[14] and no one technique was perfect. SHAP-CAM heat-maps, which combines Class Activation Maps (CAMs)[15] and DeepShap[16] (DeepExplainer), were used to explain decisions made by neural networks. CAMs were class discriminative and faithful to predicted values， but with low resolution. DeepExplainer was a combination of Deeplift[17] and Shapley value, which could generate fine-grained heat-maps but sometimes its heat-maps contain irrelevant noises. The design instinct of SHAP-CAM was similar to that of Guided Grad-CAM[18]. Given an image, a CAM and a DeepShap heat-map were generated independently, SHAP-CAM was generated by normalize the CAM heat-map to value 0-1 and multiply by the Deepshap heat-map.

References

1. Harding S, Greenwood R, Aldington S*, et al.* Grading and disease management in national screening for diabetic retinopathy in England and Wales. ***Diabetic Medicine*** 2003;20:965-971.

2. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003-2016. ***Acta Diabetol*** 2017;54(6):515-525.

3. Chao YW, Vijayanarasimhan S, Seybold B*, et al.* Rethinking the Inception Architecture for Computer Vision. ***2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition*** 2016.

4. Chollet F. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. Available at: https://arxiv.org/abs/1610.02357. Accessed October 1, 2020.

5. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. ***Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*** 2016.

6. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. ***Journal of Big Data*** 2019;6(1).

7. Wang J, Ji J, Zhang M*, et al.* Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening. ***JAMA Netw Open*** 2021;4(5):e218758.

8. Kaggle Diabetic Retinopathy Detection. 2015. Team o_O solution for the Kaggle Diabetic Retinopathy etection Challenge. Available at: https://github.com/sveitser/kaggle_diabetic. Accessed May 18, 2019.

9. Raghu M, Zhang C, Kleinberg J, Bengio S. 2019. Transfusion: Understanding Transfer Learning for Medical Imaging. Available at: https://arxiv.org/abs/1902.07208. Accessed July 16, 2020.

10. Kingma DP, Ba JL. 2015. ADAM: a method for stochastic potimization. Available at: https://arxiv.org/abs/1412.6980. Accessed July 15, 2020.

11. Minaee S, Boykov Y, Porikli F*, et al.* 2020. Image Segmentation Using Deep Learning: A Survey. Available at: https://arxiv.org/abs/2001.05566. Accessed May 6, 2020.

12. Ju C, Bibaut A, van der Laan M. The Relative Performance of Ensemble Methods with Deep Convolutional Neural Networks for Image Classification. ***J Appl Stat*** 2018;45(15):2800-2818.

13. van der Maaten L, Hinton G. Visualizing Data using t-SNE. ***Journal of Machine Learning Research*** 2008;9:2579--2605.

14. Amirata Ghorbani AA. Interpretation of Neural Networks Is Fragile. AAAI 2019; 2019.

15. Zhou B, Khosla A, Lapedriza A*, et al.* Learning Deep Features for Discriminative Localization. *ArXiv e-prints.* 2015;1512. http://adsabs.harvard.edu/abs/2015arXiv151204150Z. Accessed December 1, 2015.

16. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *arXiv e-prints.* 2017. https://ui.adsabs.harvard.edu/\#abs/2017arXiv170507874L. Accessed May 01, 2017.

17. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation

Differences. *arXiv e-prints.* 2017. https://ui.adsabs.harvard.edu/\#abs/2017arXiv170402685S. Accessed April 01, 2017.

18. Selvaraju RR, Cogswell M, Das A*, et al.* 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Available at: http://adsabs.harvard.edu/abs/2016arXiv161002391S. Accessed September 20, 2021.