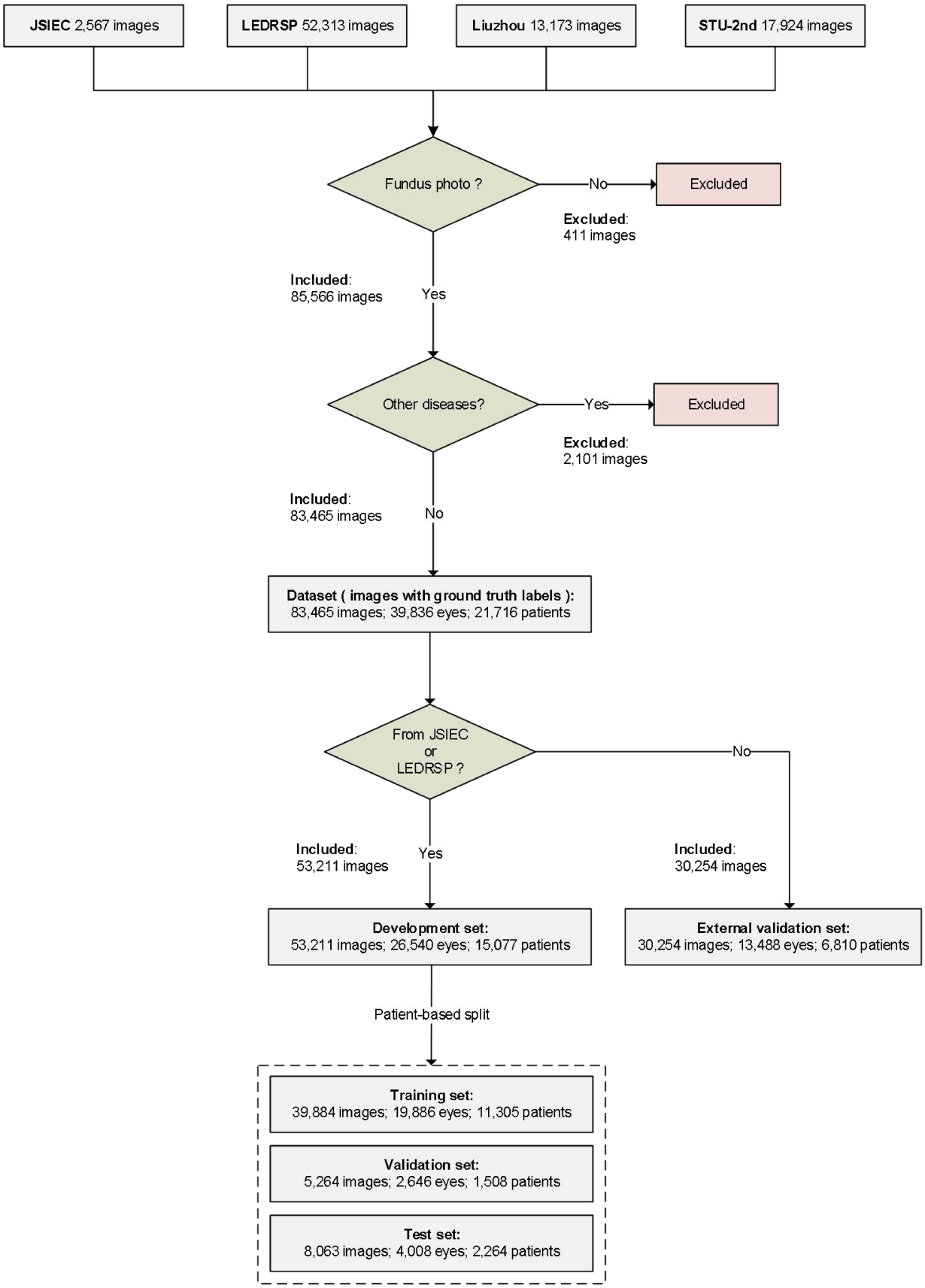
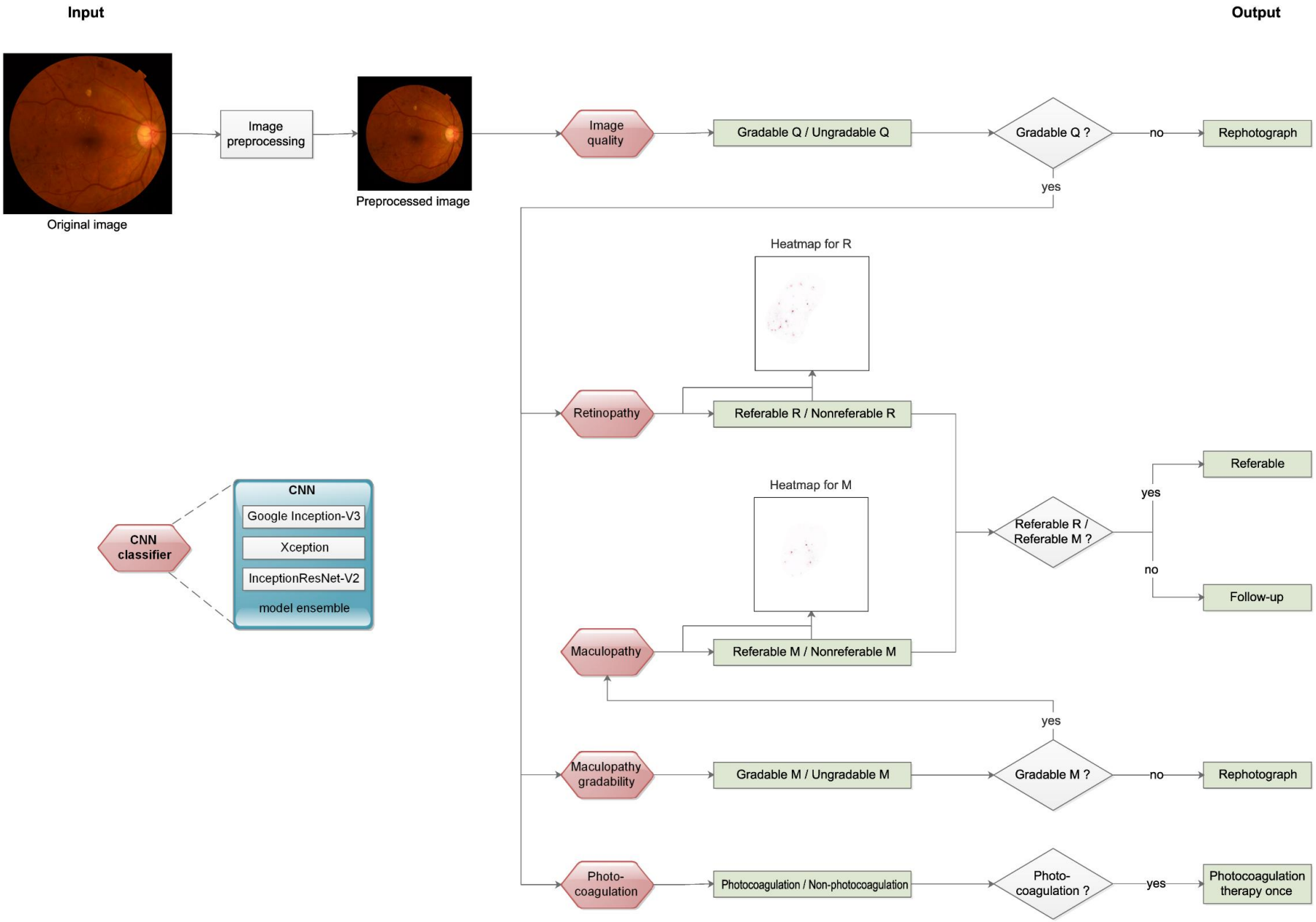
Supplementary materials

Online Content



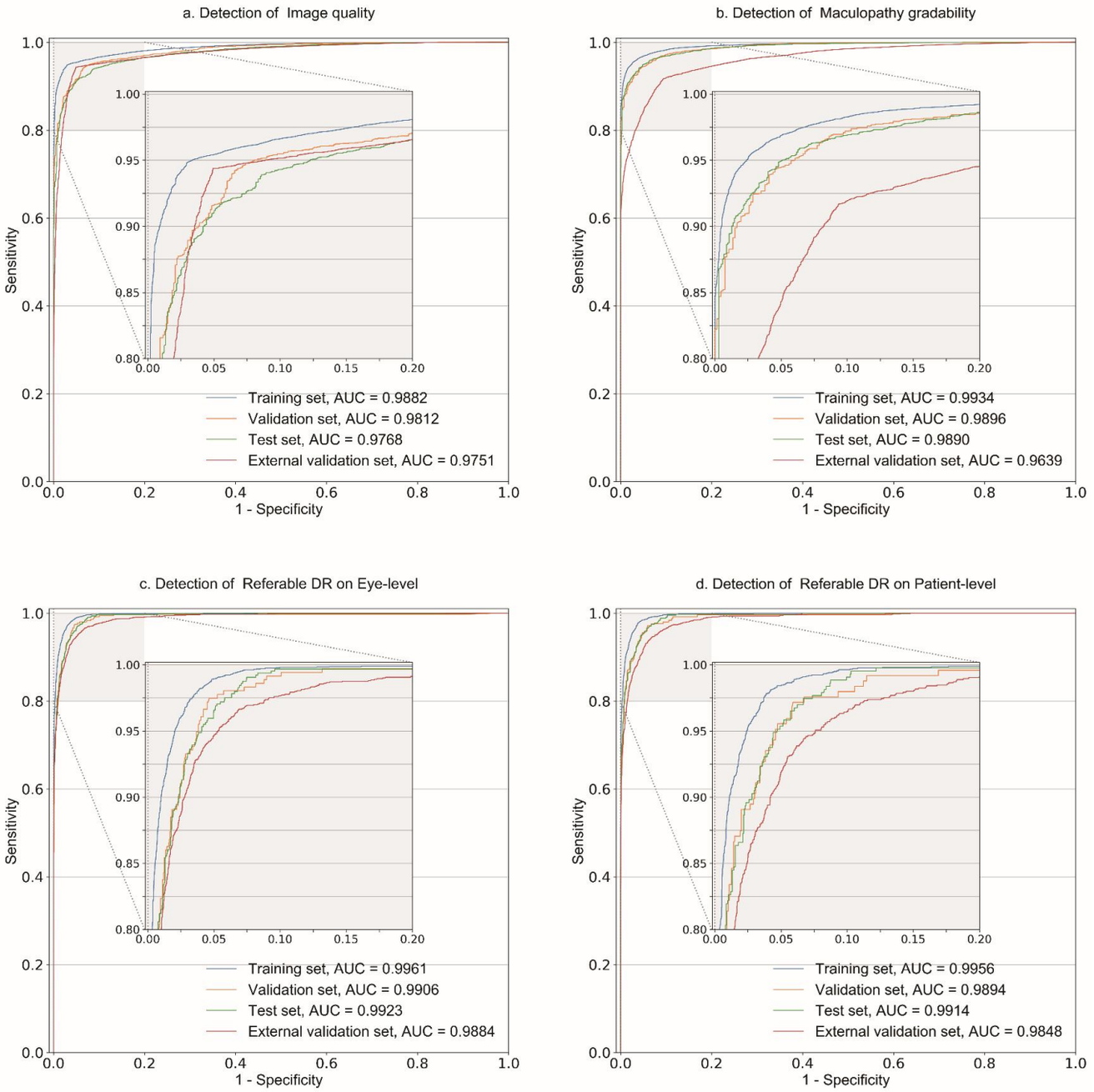
Supplemental figure 1. Workflow of retinal images database construction

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
Heatmap for M
Maculopathy
Referable M / Nonreferable M
Referable R / Referable M ?
yes
Referable
no
Follow-up
CNN classifier
CNN
Google Inception-V3
Xception
InceptionResNet-V2
model ensemble
yes
Maculopathy gradability
Gradable M / Ungradable M
Gradable M ?
no
Rephotograph
Photo-coagulation
Photocoagulation / Non-photocoagulation
Photo-coagulation ?
yes
Photocoagulation therapy once

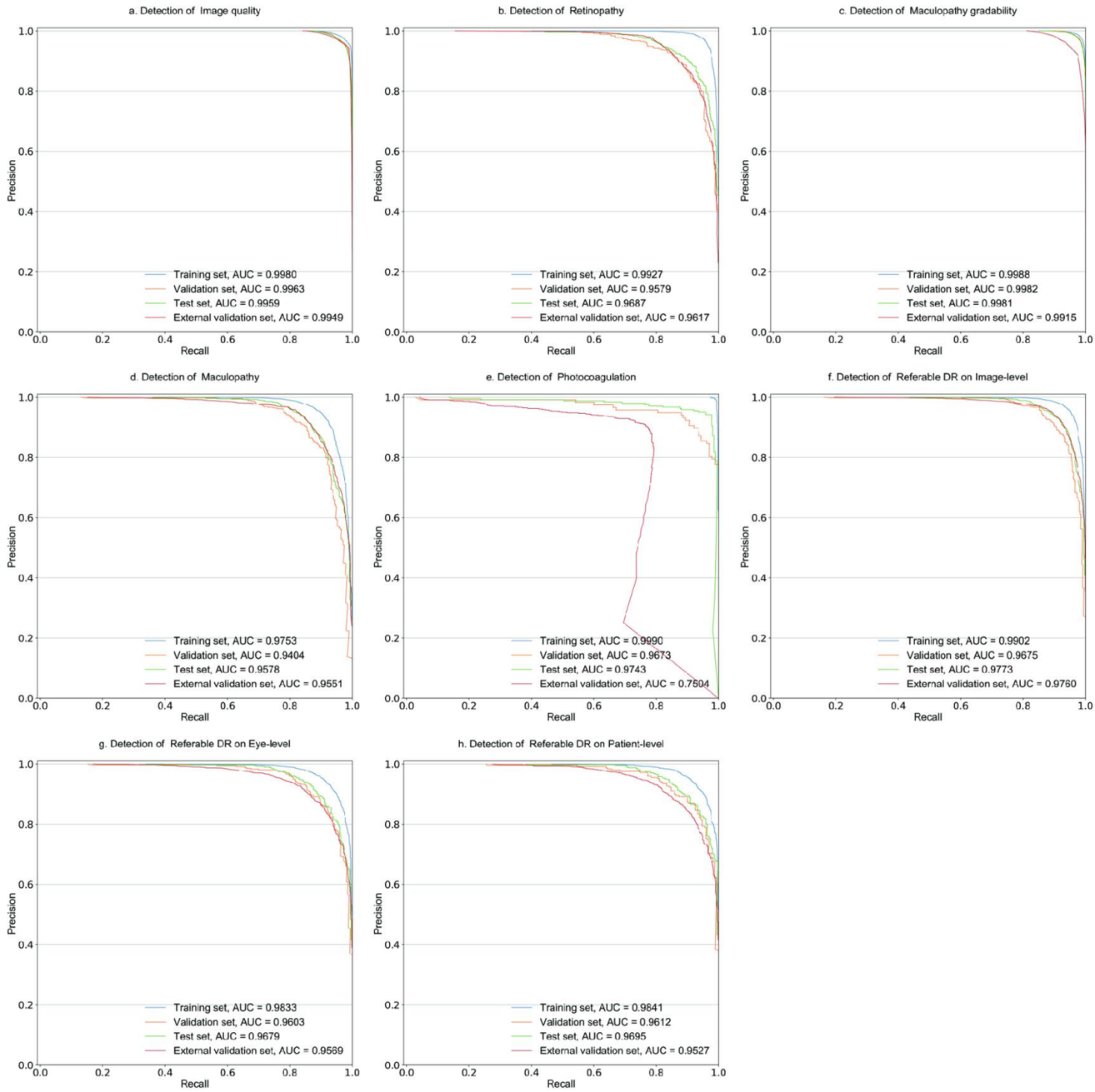
Supplemental figure 2. The pipeline of referable diabetic retinopathy screening system

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.



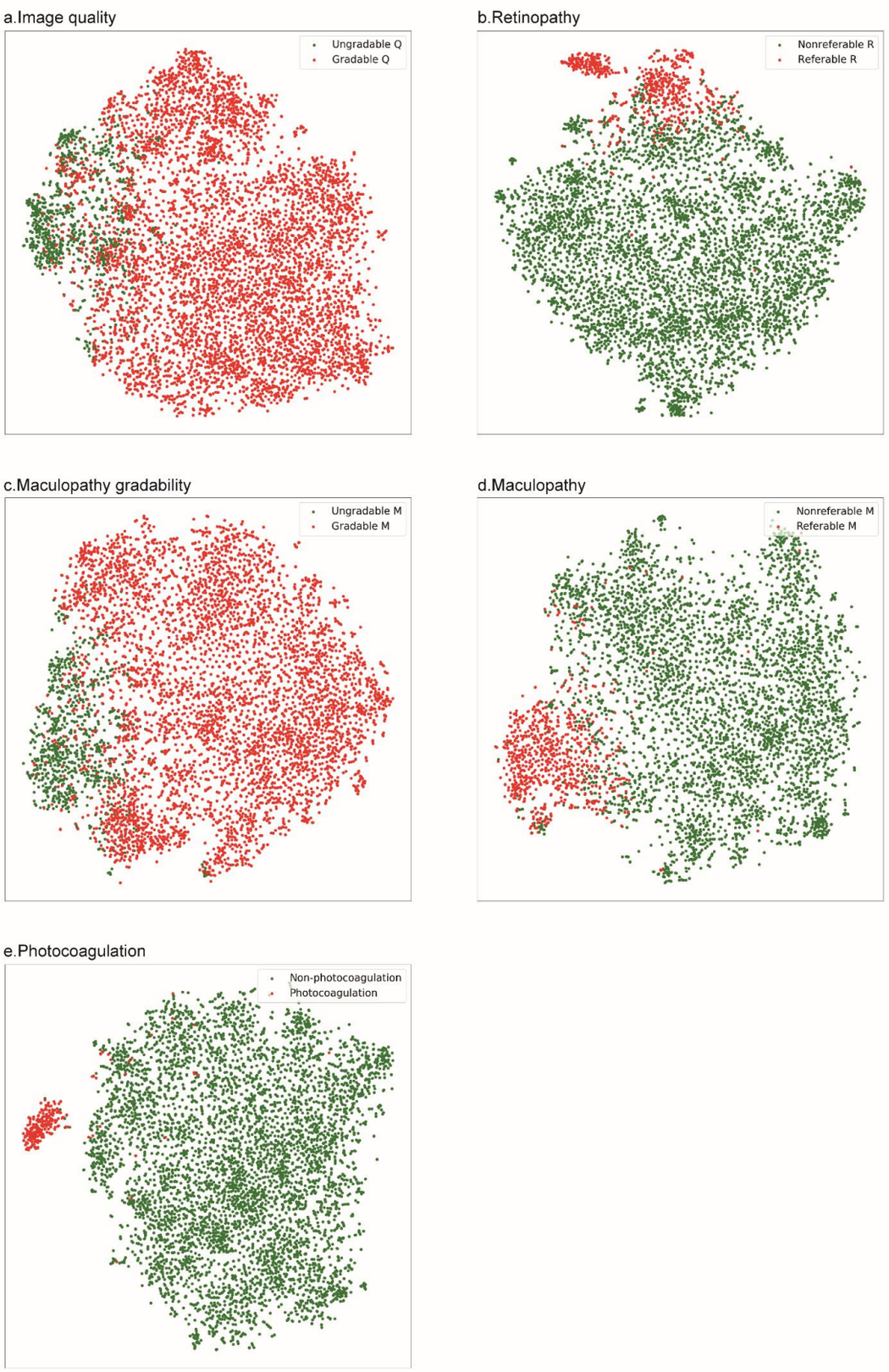
Supplemental figure 3. The receive operating characteristic (ROC) curves for system performance

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:

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

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.

Zhang G, *et al. BMJ Open* 2022; 12:e060155. doi: 10.1136/bmjopen-2021-060155

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.