Supplementary Online Content

Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. Published online September 12, 2019. doi:10.1001/jamaophthalmol.2019.3501

This supplementary material has been provided by the authors to give readers additional information about their work.

eAppendix. Supplemental Information

1. Dataset

1.1 A Brief Introduction of Chinese Glaucoma Study Alliance (CGSA)

"Chinese Glaucoma Study Alliance" is a national multi-center glaucoma research alliance in China. At present, there are 89 hospitals in the project group, and we're planning to include 100 hospitals in China, so this group is also called "100 hospitals Union". In 2009, under the leadership of Professor Ningli Wang and with the Beijing Tongren Hospital as the leading center, The Third Hospital of Handan (Han Dan City Eye Hosipital), Hebei Eye Hospital and Anyang Eye Hospital formed a multi-center research group to carry out glaucoma genetic research and to promote new technology and new projects. Since 2014, the project team has gradually expanded and included more hospitals excelling in the diagnosis and treatment of glaucoma. By then, the "glaucoma 100 Union" was officially established. The project alliance is built to further develop glaucoma genetic research, construct research platform to carry out high-quality multi-center clinical research, train clinical researchers, improve clinical research capacity, and promote new technology and new projects.

At present, the research group has set up its own EDC (electric data capture) system to achieve effective data quality control. At the same time, the group is actively preparing the voice input system, remote video consultation and rounds, expecting to further improve the glaucoma diagnosis and treatment capacity of local hospitals, help them effectively practice the diagnosis and treatment guidelines and norms from glaucoma association.

1.2 Image Quality Control and Grading Information of Development Dataset

Before training, each image went through a tiered grading system consisting of multiple layers of trained graders of increasing expertise for verification and correction of image labels. Each image imported into the database started with a label matching the most recent diagnosis of the patient.

(1) The first tier of graders consisted of five trained medical students and nonmedical undergraduates, who conducted initial quality control according to the rubric below:

a. If the image contains severe resolution reductions or artifacts significant?
b. If the image field include the entire optic nerve head and macula?
c. If the illumination is too dark, or too light?
d. If the focus good enough for grading the optic nerve head and RNFL?

(2) The second tier of graders consisted of twenty two of Chinese board-certified

ophthalmologists or postgraduate ophthalmology trainees (>2 years' experience)，who passed the pre-training test. In the process of grading, each image was assigned randomly to 2 ophthalmologists for initial grading. Each grader independently graded and recorded each image according to the criteria of GON showed in Table2.

(3) The third tier of graders consisted of two senior independent glaucoma specialists (>10 years' experience in conducting glaucoma retinopathy diagnosis), who consulted in the cases of disagreement.

1.3 Website-based dataset

We performed image searches of the search engines Google, Yahoo, Baidu and Bing throughout January 2018. In our image searches, we included keywords, such as 'color fundus images', 'color fundus photographs', 'glaucoma fundus', 'glaucoma optic nerve head', 'glaucomatous optic neuropathy', 'healthy fundus' 'abnormal fundus images' and 'normal fundus'. Two of the authors (R.P. and Y.C.) completed the searches independently. In addition, they cross-checked and confirmed that all of the cases collected were referable GON fundus or non-referable GON fundus. Referable GON was defined as a ratio of vertical cup to disc diameter of 0.8 or greater, focal thinning or notching of the neuroretinal rim, optic disc hemorrhages, or localized retinal nerve fiber layer defects—features sometimes referred to as glaucoma suspects. When discrepancies arose, final labeling was achieved from 1 retinal specialist. All confirmed images (747 non-referable and 137 referable GON images) were subsequently sent to GD-CNN for evaluation.

1.4 Tele-ophthalmic Image Reading Center of Beijing Tongren Hospital

Tele-ophthalmic Image Reading Center Of Beijing Tongren Hospital，uses The Daheng Prust cloud medical system software, which is developed by Beijing Institute of Ophthalmology and Beijing Daheng Image Company，to provide remote reading services for 105 primary hospitals in 20 provinces, municipalities and autonomous regions across the country. Prior to the screening, primary hospitals ophthalmologist gave the patient verbal notification and informed consent. When reading the image remotely, the primary hospital is required to take at least one patient's 45-degree eyelid for each eye, and must also provide vision, chief complaint, and systemic medical history. If the fundus cannot be seen clearly, it is required to look outside the camera. Before going online, primary hospitals arranged for professionals to conduct uniform technical operation training for doctors and photographic staff of primary hospitals. The

reading center doctors perform reading training before taking a job. During the reading of the image， the reader will perform quality assessment of the uploaded fundus image and text medical records on the reading interface through options. Then doctors describe the uploaded image, diagnose according to the uploaded data for the interpretation of the reading, and give advice on the treatment of the patient.

2. Preprocess

The collected fundus images normally exist large redundancy that is irrelevant to GON. Thus, it is necessary to preprocess the fundus images for removing their redundancy, before making decision on glaucoma through GD-CNN. Specifically, the original fundus images are first downsized to 224×224 pixels by cropping the images that center at the optic cup and meanwhile contain part of the surrounding vessel. It is because glaucoma is highly correlated with the optic cup and the surrounding vessel. In our preprocess procedure, the optic cups are automatically detected by searching the area with highest intensity in the gray-scale map of each fundus image, since we found that the optic cup is the lightest area in the whole fundus image.

Next, we remove the mean values of all training fundus images for each of the RGB channels, such that the input to GD-CNN is around 0 for relieving the over-fitting issue. Assume that a_{ij}^{c} means the value of the $(i, j)-$th pixel at channel c. Here, $c \in \{1,2,3\}$ corresponds to the channels of red, green and blue. Then, the mean pixel value of the $c-$th channel over the training images can be calculated by

$$M_c = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{\mathrm{j}=1}^{J} a_{ij}^{\ c}, \qquad (1)$$

where $I(=224)$ and $J(=224)$ are the total numbers of pixels in each dimension of the cropped fundus image. In addition, K is the total number of training fundus images, which is 209,413 in the training set of our database. Before entering into GD-CNN, the training or test images are subtracted by the mean value at each channel,

$$\mathbf{I_c}' = \mathbf{I_c} - M_c, \qquad (2)$$

where $\mathbf{I_c}$ and $\mathbf{I_c}'$ are the RGB matrices of the fundus image before and after the subtraction of the mean value, respectively. As such, the redundancy of the fundus image can be removed for the binary classification of glaucoma in GD-CNN.

(1）Framework of DLS

The framework of the DLS algorithm for automatic glaucoma detection is shown in eFigure A. As shown in this figure, our algorithm is composed of the training and test stages. At the training stage, the extensive fundus images with manual labels of referable GON or un-referable GON are used to train GON Diagnosis CNN (GD-CNN). Note that the training fundus images can be obtained from our training dataset.

In the process of training, GD-CNN updates the trainable parameters by minimizing the cross-entropy loss between the prediction and ground truth results of GON over the training fundus images. The minimization is achieved through the back propagation (BP) algorithm with stochastic gradient descent (SGD) optimizer, calculating the derivative of the loss function to every parameters.

At the test stage, the probability of the input fundus image being GON can be decided by the output of GD-CNN, which is in terms of a sigmoid probability.

(2) Structure of GD-CNN

The GD-CNN structure is based on ResNet, which is a widely used CNN for the object classification of natural images. ResNet focuses on 1000-category classification, whereas our GD-CNN only handles the binary classification of glaucoma. Moreover, the natural images are much more diverse than the fundus images. Hence, we simplify GD-CNN to reduce the number of trainable parameters for avoiding the over-fitting issue.

Given the modules of feature extraction and sum by pixel (Figure1 B and C), the structure of GD-CNN is shown in Figure1 D. As shown in this figure, the input to GD-CNN is three 224×224 matrices, corresponding to the RGB channels of the cropped fundus image. Given the input matrices, one convolutional layer with max pooling is applied to automatically learn the feature maps. Note that the Rectified Linear Unit (ReLU) is used as the activation layer after each convolutional layer. The numbers and sizes of these convolutional layers are reported in Figure1 E. After the first convolutional operation, the four res models are designed to avoid overfitting.

With the layer of GD-CNN going deeper, the prediction accuracy may be saturated and then degrades rapidly. Hence, the residual mapping is proposed in ResNet, which is applied in GD-CNN. Subsequently, the max pooling layer and the fully connection layer are appended in GD-CNN. Finally, GD-CNN outputs the probabilities of positive and negative glaucoma represented in terms of binary labels.

(3) Loss function

For the binary classification of GON diagnosis, the cross-entropy function is applied in GD-CNN as the loss function, which measures the distance between the predicted and ground truth results of glaucoma diagnosis. When training GD-CNN, the parameters are updated to minimize the cross-entropy loss function, so that the glaucoma prediction of GD-CNN is increasingly close to the ground truth. The cross-entropy function C is defined by

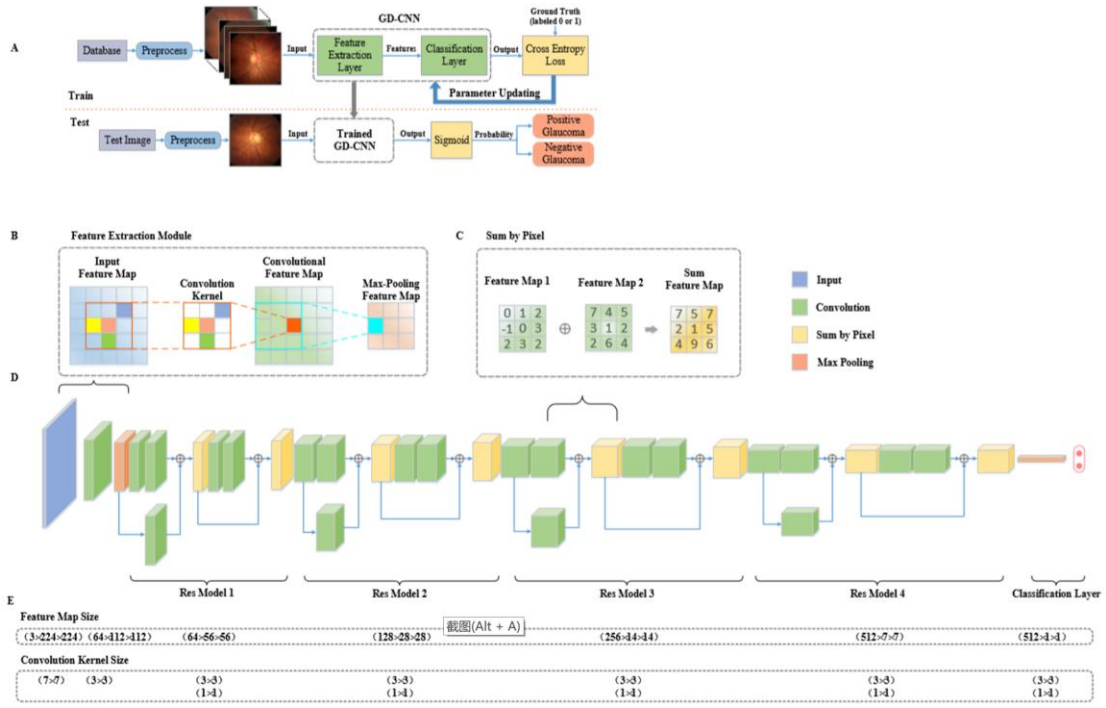
$$C = -\frac{1}{M}\sum_{m=1}^{M}[y^{(m)}\log b^{(m)} + (1-y^{(m)})\log(1-b^{(m)})], \quad (1)$$

where M is the batch size (i.e., the number of training images) in one training iteration. In (1), $y^{(m)}$ is the ground truth label of the $m-$th training fundus image, in which $y^{(m)}=1$ means that the $m-$th training image is glaucomatous and $y^{(m)}=0$ indicates the image is negative glaucoma. Moreover, $b^{(m)}$ in (1) represents the probability of GON predicted by GD-CNN for the $m-$th training image, which is modelled by the following sigmoid function:

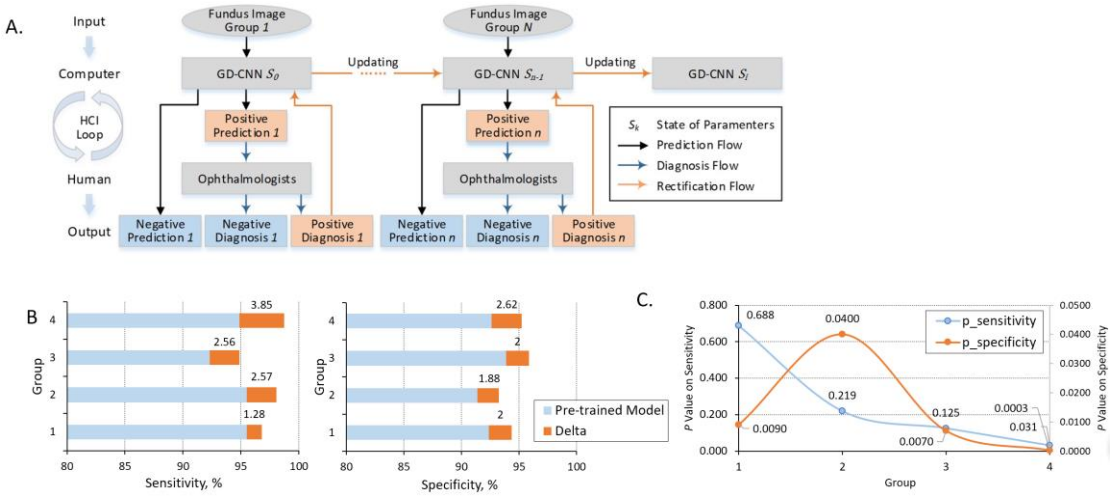
$$b^{(m)} = \frac{1}{1+e^{-z^{(m)}}}, \quad (2)$$

where $z^{(m)}$ denotes the output of GD-CNN.

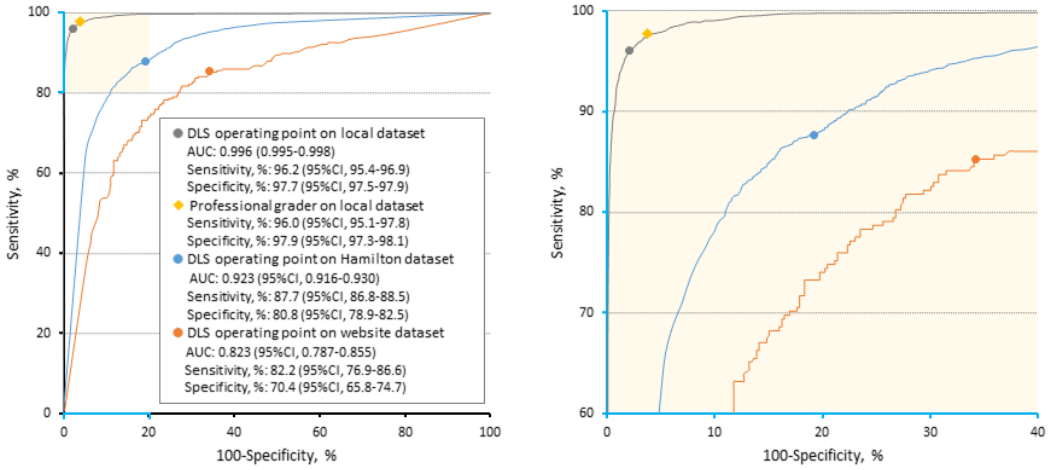
We can see from the loss function of (1) that large loss is imposed on the training images, in which the positive glaucoma or negative glaucoma images are wrongly labelled by GD-CNN. For example, the loss of (1) is large, when $b^{(m)} \to 0$ for $y^{(m)}=1$ or $b^{(m)} \to 1$ for $y^{(m)}=0$. Therefore, the parameters of GD-CNN are trained to minimize the cross-entropy loss of (1) through the back propagation (BP) algorithm. Finally, the trained GD-CNN model can be used to predict whether an input fundus image is glaucomatous according to the output value of $b^{(m)}$. When $b^{(m)} > \mathrm{th}_b$, the input fundus image is predicted as GON, and otherwise it is predicted as negative GON. Normally, we set $\mathrm{th}_b = 0.5$ for GON prediction, but the value of th_b can be adjustable for making the trade-off between sensitivity and specificity in GON diagnosis.



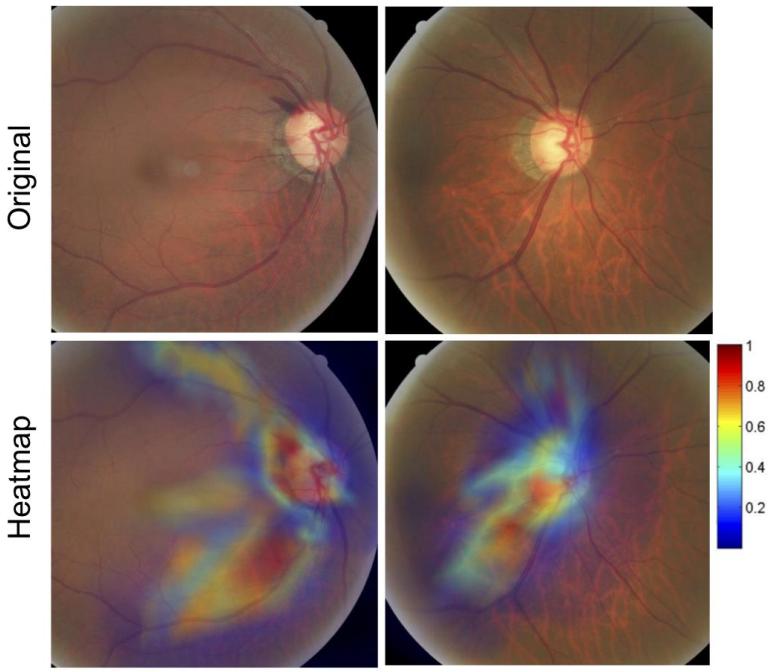
eFigure 1. Deep Learning System（DLS）A, The framework of the deep learning system (DLS) for detecting possible glaucoma. B, The basic feature extraction module in the convolutional neural network. The operation of convolution uses the kernel to multiply a block of features, sum up the multiplied features, and then slide over the input feature map. The max-pooling layer chooses the maximal value inside the region of pooling kernel as the max-pooling output, decreasing the dimension of the feature map in order to avoid overfitting. Note that the parameters of the convolution kernel are in 3 dimension which are updated through the training process, while the operation of pooling is calculated in 2 dimension and it does not exist any parameters. C, The operation of pixel-wise sum. Two 3-dimentional feature maps are summed up by pixel to create a new feature map. D, The structure of GD-CNN. It is a simplified ResNet, which contains a basic convolutional layer with max pooling, four res models, one final pooling layer and a 2-dimensional full connection layer. E. The sizes of feature maps and convolution kernel sizes in each layer.



eFigure 2. Performance of the online deep learning (ODL) system for glaucoma diagnosis alongside increasing number of samples. A, Abstraction of the ODL system. In our experiments, 5 groups of samples are sequentially obtained from the tele-ophthalmology platform. The performance of the ODL system is evaluated in terms of AUC, sensitivity and specificity, for each of sample group in a sequential order. B, Online Performance on Sensitivity and Specificity. The curves of AUC and sensitivity incrementally grow along with the increased number of diagnosed groups. C, Performance of *P* Value. This verifies the effectiveness of the ODL system, in which human (ophthalmologists) help machines improve the prediction performance and machines help human to be more efficient in diagnosing the negative samples of glaucoma



eFigure 3. Receiver Operating Characteristic Curve and Area Under the Curve of the Deep Learning System for GD-CNN in the validation Datasets. The curve is the receiver operating characteristic curve (ROC) and the orange diamond is the operating point to measure the sensitivity and specificity. AUC represents the area under the receiver operating characteristic curve.



eFigure 4. Visualization maps generated from deep features. The visualization map is created after prediction by assigning the softmax probability of the correct label to each occluded area and which can be superimposed on the input image to highlight the areas the model considered important in making its diagnosis.

eTable 1. Proportion of Definite, Probable and Unlikely Glaucomatous Optic Neuropathy in the Training and Local Validation Dataset from Chinese Glaucoma Study Alliance

				No. (%)								
	No.			Referable [a]						Non-referable		
	Total			Definite GON [b]			Probable GON [c]			Unlikely GON [d]		
Dataset	Images	Eyes	Individuals	Images	Eyes	Individuals	Images	Eyes	Individuals	Images	Eyes	Individuals
Training	241032	120522	68013	29865 (12.4%)	15872 (13.2%)	7936 (11.7%)	11046 (4.6%)	5639 (4.6%)	2820 (4.1%)	200121 (83%)	99011 (82.2%)	57257 (84.2%)
Local Validation	28569	10371	5290	4514 (15.8%)	2490 (24%)	1245 (23.5%)	571 (2%)	313 (3%)	159 (3%)	23484 (82.2%)	7568 (73%)	3886 (73.5%)

a. Referable GON is defined as Definite GON and Probable GON.
b. Definite GON is defined as, any of the following conditions: VCDR≥0.85; RNFL defects corresponds with thinning area of rim or notches.
c. Probable GON neuropathy is defined as, at least two conditions positive: 0.7VCDR < 0.85; Rim Width ≤ 0.1 DD; General Rim Thinning ≥ 60° or localized Rim Thinning < 60°（11-1o'clock or 5-7o'clock); RNFL defects ; Splinter Hemorrhages, peripapillary atrophy (Beta zone) .
d. Unlikely GON is defined as, with no sign of the features of Definite GON and/or Probable GON.

eTable 2. The Proportion of Reasons for False-Negative and False-Positive Results by the GD-CNN and Manual Grading

	GD-CNN Grading No. (%)	Manual Grading No. (%)
False-Negative Reason		
GON	6 (5.5)	6 (5.3)
With AMD	18 (16.4)	13 (11.5)
With Pathologic or high myopia	51 (46.3)	50 (44.2)
With Diabetic retinopathy	15 (13.6)	12 (10.6)
RNFL defect only	14 (12.7)	27 (23.9)
optic disc hemorrhage only	6 (5.5)	5 (4.4)
Total	110 (100)	113 (100)
False-Positive Reason		
Normal fundus	18 (3.1)	6 (1.2)
AMD	90 (15.3)	51 (9.5)
Diabetic retinopathy	64 (10.9)	37 (6.9)
Pathologic or high myopia	191 (32.3)	183 (34.0)
Physiologic large cupping	94 (16.0)	138 (25.6)
Nonglaucomatous optic atrophy	111 (18.9)	99 (18.4)
Congenital optic disc abnormalities	5 (0.9)	5 (0.9)
Congenital optic disc vessels abnormalities	11 (1.9)	7 (1.3)
Optic disc edema	4 (0.7)	12 (2.2)
Total	588 (100)	538 (100)