Establishing a Generalized Deep Learning System for Detection of Glaucomatous Optic Neuropathy using Fundus Photographs

Hanruo Liu[1], MD, PhD; Liu Li[2], BEng; I. Michael Wormstone[3], PhD; Chunyan Qiao[1], MD, PhD; Chun Zhang[4], MD, PhD; Ping Liu[5], MD, PhD; Shuning Li[1], MD, PhD; Huaizhou Wang[1], MD, PhD; Dapeng Mou[1], MD, PhD; Ruiqi Pang[1], MD; Diya Yang[1], MD, PhD; Lai Jiang[2], BEng; Yihan Chen[1], MD; Man Hu[6], MD, PhD; Yongli Xu[7], PhD; Hong Kang[8], PhD; Xin Ji[9], BEng; Robert Chang[10], MD, PhD; Clement Tham[11], MD, PhD; Carol Cheung[11], PhD; Daniel Shu Wei Ting[12], MD, PhD; Tien Yin Wong[12], MD, PhD; Zulin Wang[2], PhD; Robert N. Weinreb[13], MD, PhD; Mai Xu[2*], PhD; Ningli Wang[1*], MD, PhD

1. Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing Ophthalmology & Visual Science Key Lab, Beijing, China.
2. School of Electronic and Information Engineering, Beihang University, Beijing 100191, China.
3. School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
4. Department of Ophthalmology, Peking University Third Hospital, Beijing, China
5. Ophthalmology Hospital, First Hospital of Harbin Medical University, Harbin 150001, Heilongjiang, China
6. National Key Discipline of Pediatrics, Ministry of Education, Department of Ophthalmology, Beijing Children's Hospital, Capital Medical University, Beijing, China
7. Department of Mathematics, Beijing University of Chemical Technology, Beijing, China
8. Nankai University, the College of Computer Science, Tianjin, China
9. Beijing Shanggong Medical Technology co., Ltd, Beijing, China

10. Department of Ophthalmology, Byers Eye Institute at Stanford University, Palo Alto, CA, US
11. Department of Ophthalmology & Visual Sciences, Faculty of Medicine, The Chinese University of Hong Kong, 147K Argyle Street, Kowloon, Hong Kong, China
12. Singapore National Eye Center, Singapore Eye Research Institute, Duke-National University of Singapore Graduate School of Medicine, National University of Singapore and National University, Singapore 168751, Singapore
13. Shiley Eye Institute, University of California San Diego, 9500 Gilman Drive, MC 0946, La Jolla, CA 92093, US

*These authors contributed equally to this work.

Correspondence should be addressed to Ningli Wang (email: wningli@vip.163.com), or Mai Xu. (maixu@buaa.edu.cn).

Running head: Deep Learning System for Detection of Glaucomatous Optic Neuropathy

Number of Tables: 3

Number of Figures: 2

Word count (total): Abstract: 338 Text: 2996 words (excluding abstract, references and tables), Key Points 100 words

Key Points

Question:
How does a deep learning system (DLS) compare with professional human graders in detecting glaucomatous optic neuropathy (GON)?

Findings:
The DLS showed a sensitivity of 96.2% and specificity of 97.7% for detecting GON in a local validation dataset；93.6-96.1% sensitivity and 95.6-97.1% specificity in three clinical-based datasets；91.0% sensitivity and 92.6% specificity in a real-world distribution dataset; 87.7% sensitivity and 80.8% specificity in a multi-ethnic dataset; 82.2% sensitivity and 70.4% specificity in a website-based dataset.

Meaning:
This assessment of fundus images suggests DLS can provide a tool with high sensitivity, specificity and might expedite screening for GON.

Abstract:

IMPORTANCE A deep learning system (DLS) that could automatically detect glaucomatous optic neuropathy (GON) with high sensitivity and specificity might expedite screening for GON.

OBJECTIVE To establish a DLS for detection of GON using retinal fundus images and convoluted neural networks (GD-CNN) that has the ability to be generalized across populations.

DESIGN, SETTING, AND PARTICIPANTS A DLS for the classification of GON was developed for automated classification of GON using retinal fundus images. To build and validate GD-CNN, a total of 355 339 fundus images were included. Of those, 241 032 images and 114 307 images were selected as the training and validation dataset, respectively. The generalization of the DLS was tested in several validation datasets, which allowed assessment of the DLS in a clinical setting without exclusions, testing against variable image quality based on fundus photographs obtained from websites, evaluation in a population-based study that reflects a natural distribution of glaucoma patients within the cohort and an additive dataset that has a diverse ethnic distribution. An online learning system was established to transfer the trained and validated DLS to generalize the results with fundus images from new sources. To better understand the DLS decision making process, a prediction visualization test was performed that identified regions of the fundus images utilized by the DLS for diagnosis.

EXPOSURES Use of a deep learning system.

MAIN OUTCOMES AND MEASURES Area under the receiver operating characteristics curve (AUC), sensitivity and specificity for DLS with reference to professional graders.

RESULTS The AUC of the GD-CNN model in primary local validation datasets was 0.996 (95% CI, 0.995-0.998), with sensitivity of 0.962, and specificity of 0.977. The most common reasons for both false-negative and false-positive grading by GD-CNN (46.3% and 32.3%) and manual grading (44.2% and 34.0%) was pathologic or high myopia.

CONCLUSIONS AND RELEVANCE Application of GD-CNN to fundus images from different settings and varying image quality demonstrated a high sensitivity, specificity and generalization for detecting GON. These findings suggest automated DLS might enhance current screening programs in a cost-effective and time-efficient manner.

Glaucoma is the leading cause of irreversible blindness.[1] It is predicted to affect 80 million people worldwide by 2020 and 111.8 million by 2040.[2] Glaucoma is a chronic neurodegenerative disease of the eye. [3] The majority of glaucoma patients are unaware of their condition until late in the course of their disease, when central visual acuity is affected.[4] Screening and early detection of glaucoma, along with timely referral and treatment, is a generally accepted strategy for preventing vision loss.[5] Digital fundus image evaluation has emerged as a modality for large-scale glaucoma screening due to convenience and relative affordability.[6,7] Nevertheless, this process of manual image assessment is labor intensive and time-consuming.[8] In addition, glaucoma diagnosis from fundus images is subjective, and efficiency is likely linked to the experience and skill of the observer.

Artificial intelligence has been successfully applied in image-based medical diagnoses, such as skin cancer, breast cancer, brain tumors and diabetic retinopathy. [9-13] The deep learning system (DLS) approach also has recently been adopted to provide high sensitivity and specificity (>90%) for detecting glaucomatous optic neuropathy (GON) from high-quality retinal fundus images.[14] However, the use of DLS for medical diagnosis has inferior performance when applied to data obtained from different sources.[12,14] This is an important consideration, as ideally a DLS would need to be generally utilized in different settings in which the images will be of varying quality, ethnicity and population sources if maximum reach and clinical benefit is to be achieved.[15-17]

In this study, we established a large-scale database of fundus images for glaucoma diagnosis ('FIGD' database) and developed from fundus images Glaucoma Diagnosis with Convoluted Neural Networks (GD-CNN), as an advanced DLS approach for automatically detecting GON that has the ability to be generalized across populations.

Methods

Training datasets

The study was conducted according to the tenets of the Declaration of Helsinki and it was approved by the institutional review board (IRB) of Beijing Tongren Hospital (identifier, TRECKY2018-034). As the study was a retrospective review and analysis of fully anonymized colour retinal fundus images, the medical ethics committee exempted the need for the patients' informed consent.

To establish an automatic diagnosis system for GON, a total of 274 413 fundus images were obtained from the Chinese Glaucoma Study Alliance (CGSA, Appendix 1.1 available online at www.aaojournal.org) between 2009 and 2017 (Table 1). The CGSA uses a tele-ophthalmology platform and a cloud-based online dataset (http://www.funduspace.com Accessed May 2017), which has established its own electronic data capture system to achieve effective data quality control. For each patient, two fundus images of each eye were recorded. For this study, each image in the training dataset was subjected to a tiered grading system consisting of multiple layers of trained graders of increasing expertise. Each image imported into the database started with a label matching the most recent diagnosis of the patient. The first tier of graders consisted of five trained medical students and non-medical undergraduates. They conducted initial quality control according to the following rules: 1) the image did not contain severe resolution reductions or significant artifacts; 2) the image field included the entire optic nerve head and macula; 3) the illumination was acceptable i.e. not too dark or too light; 4) the image was focused sufficiently for grading the optic nerve head and retinal nerve fiber layer (RNFL). The second tier of graders consisted of twenty-two Chinese board-certified ophthalmologists or postgraduate ophthalmology trainees (>2 years'

experience) who had passed a pre-training test. In the process of grading, each image was assigned randomly to two ophthalmologists for grading. Each grader independently graded and recorded each image according to the criteria of GON (Table 2). The third tier of graders consisted of two senior independent glaucoma specialists (>10 years of experience with glaucoma diagnosis); they were consulted to adjudicate disagreement in tier 2 grading (Appendix 1.2， available online at www.aaojournal.org). Following this process images were classified as unlikely, probable, and definite GON. Referable GON was defined as probable or definite GON.

GD-CNN Model

The training images with assigned labels were utilized to establish a state-of-the-art DLS, GD-CNN, based on the Residual Net (ResNet) platform.[18] (eFigures 1 & 2, Appendix 2.0 - www.aaojournal.org). In the current study, we restricted the analysis to the binary classification problem of glaucoma in fundus images. The basic operation of ResNet is to apply convolution repeatedly, which is computationally quite expensive for high-resolution images. Therefore, we pre-process images by down-sampling them to 224×224 pixel resolution. In addition, these images were centered on the optic cup and contained part of the surrounding vessels, as glaucoma is highly correlated with alteration in these regions.[19] To achieve this, the optic cups were automatically detected by recognition of the area with the highest intensity on the grayscale map of each fundus image; this was found to consistently be associated with the optic cup. Next, we calculate the mean values of red, green and blue (RGB) channels, respectively, among all the fundus images in the training dataset. Then, for each sample, we remove the three mean values on RGB channels, such that the input to GD-CNN is around 0 for relieving the over-fitting issue.[20] As such, the redundancy of the fundus image can be removed for the binary classification of glaucoma in GD-CNN. Since the GON

diagnosis was formulated as a binary classification problem, predicting whether GON was positive or negative, a cross-entropy function was applied in GD-CNN as the loss function. For each parameter assessed, GD-CNN was trained to minimize the cross-entropy loss over the large-scale training samples of positive and negative GON. The minimization was achieved through the back-propagation algorithm with the stochastic gradient descent optimizer. Once training of GD-CNN was established, the system was applied to validation sets.

Validation datasets

Details of all validation datasets are described in Table 1 and eTable 1. The initial local validation dataset did not overlap with the image data used in training. Images previously not seen by the network were presented to GD-CNN for assessment and automated diagnosis. The images were also independently assessed by three experienced professional graders (>2 years' experience) in detecting referable GON.

Online deep learning (ODL) system

The central challenge of applying DLSs in medicine is the ability to guarantee generalizability in prediction. Generalization refers to the ability of DLSs to successfully perform when assessing previously unseen samples from different data sources. An ODL system was developed to improve the generalization ability of the GD-CNN model, making automatic GON diagnosis practical. In the ODL system, the GD-CNN model is used to sequentially predict GON with a Human-Computer Interaction (HCI) loop (eFigure 2 A). The HCI loop consisted of three iterative steps: (1) The computer used GD-CNN to initially diagnose glaucoma of fundus images with a high sensitivity rate; (2) the ophthalmologists manually confirmed the positive samples predicted by the computer; (3) the confirmed

samples fine-tuned the GD-CNN model, which was used for initial GON diagnosis of the subsequent fundus images (i.e., go to step 1).

Visualization of Prediction

Following Zeiler and Fergus,[21] we visualized the contributions of different regions to GD-CNN prediction of GON on fundus images. The visualization is represented by heat maps, which highlight strong prognostic regions of the fundus images. The experiment of occlusion testing was conducted to obtain the visualization results. First, original fundus image was resized into a 360x360 RGB image. Then, a 60x60 gray block was used to slice through the fundus image (with a stride of 10 pixels), alongside both horizontal and vertical axes. Consequently, the fundus image generates 961 (=31x31) visualization testing images, each of which has a 60x60 gray block at different position, respectively. Second, the visualization testing images were predicted using the GD-CNN model. For each visualization test image, the prediction probability output refers to the value of the visualization heat map at the corresponding position. Hence, the visualization heat map was 31x31. Finally, the heat map was mapped to the original fundus image to visualize the importance of each region in GON prediction.

The deep features refer to the output of the final max pooling layer, which is in 512 dimensions. In order to visualize the distribution of the deep features from different categories, the dimensionality of deep features was reduced by t-distributed stochastic neighbor embedding visualization (t-SNE) from 512 to 3. Note that t-SNE is a state-of-the-art nonlinear dimensionality reduction method. The deep features from glaucoma and negative glaucoma are clustered into two groups once the training loss converges. The groups of two

clusters can be clearly separated, verifying the effectiveness of the deep features learned in GD-CNN.

Statistical analysis

The performance of our algorithm was evaluated in terms of area under the curve (AUC) of receiver operating characteristic (ROC) curves. 95% confidence intervals for AUC were calculated non-parametrically through logit-transformation-based confidence intervals, which was found to have good coverage accuracy over unbiased samples. In addition to AUC, sensitivity and specificity of each operating point in ROC curves were also measured with 2-sided 95% confidence intervals. These confidence intervals were calculated as Clopper-Pearson intervals, which are “exact” intervals based on cumulative probabilities.

Furthermore, to determine if the ODL system has an effect on diagnosing glaucoma, McNemar tests were conducted between the original GD-CNN model and the fine-tuned GD-CNN models. Specifically, two 2x2 contingency tables were applied to count the diagnosis changes after ODL, for positive and negative samples, respectively. Then a Chi-squared based P value was calculated along with the sensitivity/specificity over each validation dataset.

All statistical analyses were computed using the Stats Models (version 0.6.1) python package and Matlab AUC (version 1.1) package.

Results

Training, validation and evaluation of the GD-CNN model

From a total of 274 413 fundus images initially obtained from CGSA, 269 601 images passed initial image quality review and were graded for GON by the second-tier graders of Chinese board-certified ophthalmologists. The median quantity of images per ophthalmologist graded was 14 756 (range, 8 762-55 389) and ten ophthalmologists graded more than 15 000 images. 13 254 images of disagreement in tier 2 grading were adjudicate by senior glaucoma specialists. 241 032 images (definite GON 29 865 (12.4%), probable GON 11 046 (4.6%), unlikely GON 200 121 (83%) from 68 013 patients were selected, using random sampling, to train the GD-CNN model. Validation and evaluation of the GD-CNN model was assessed using the remaining 28 569 images from CGSA. Distribution of the three diagnostic categories was 15.8% definite GON, 2% probable GON and 82.2% unlikely GON (eTable 1). In local validation dataset, the AUC of the GD-CNN model was 0.996 (95%CI, 0.995-0.998), and sensitivity and specificity in detecting referable GON was comparable with that of trained professional graders (96.2%vs 96.0%; P = 0.76; 97.7% vs 97.9%; P = 0.81 respectively) (eFigure 3). To evaluate the ability of the GD-CNN to work across different populations, three clinical based studies were performed to reflect the routine functioning of an ophthalmic center. When images from these cohorts from different hospitals were diagnosed through GD-CNN and compared to clinical evaluation, performance remained high (Table 3), such that the AUC for referable GON ranged from 0.995 to 0.987 , with both sensitivity and specificity of greater than 90% (range: 93.6-96.1% and 95.6-97.1% respectively). Further evaluation was undertaken using the Handan Eye Study dataset to provide a real-world distribution of glaucoma patients. In this case AUC was 0.964 with a sensitivity of 91.0% and specificity 92.6% (Table 3). To test GD-CNN across a range of ethnic backgrounds, a multi-ethnic dataset (73.0% White, 19.3% Black/African American, 5.4% Asian, 0.3% Middle Eastern) from the Hamilton Glaucoma Center was utilized, with AUC of 0.923, sensitivity of 87.7% and specificity 80.8%. GD-CNN showed an AUC of

0.823 with 82.2% sensitivity and 70.4% specificity in a varied range of image quality dataset from worldwide web (Table 3).

Understanding the basis for incorrect diagnosis

Among the local validation datasets, an additional analysis was conducted to further evaluate GD-CNN's performance, to better establish the basis for false positive and negative diagnosis (eTable 2). The most common reason for undetected GON from fundus images was pathological or high myopia for both GD-CNN and manual grading (n = 51 [46.3%] and n = 50 [44.2%] respectively). Interestingly, the most likely cause for a false-positive classification by DLS or manual grading was also pathological or high myopia (n = 191 [32.3%] and n =183 [34%] respectively). Physiologically large cupping was also a common cause of false positives with manual diagnosis (n = 138 [25.6%]), and to a lesser degree with GD-CNN (n = 94 [16.0%]).

Implementation of the ODL system

The ODL system was implemented in the tele-ophthalmic image reading platform of Beijing Tongren Hospital (Appendix 1.4), which collected a group of fundus images every week (around 600 images). It was found that the ODL system both sensitivity and specificity improve with each group of samples collected sequentially over a five-week period (eFigure 2).

Visualization of prediction

To visualize the learning procedure and represent the areas contributing most to the DLS, we created a heatmap which superimposed a convolutional visualization layer at the end of our network; performed on 1000 images (Figure 1 and eFigure 4). The regions of interest

identified to have made the greatest contribution to the neural network's diagnosis were also shared with 91.8% of ophthalmologists (Figure 2A). All areas containing optic nerve head variance and neuroretinal rim loss were located correctly on all the images used for testing, while RNFL defects and peripapillary atrophy (PPA) on occasions did not present a clear point of interest with an accuracy of 90.0% and 87.0% respectively. Figure 2B represents a t-distributed stochastic neighbor embedding visualization of this data set by our automated method, clearly showing 2 clusters of fundus images and indicating the ability of our model to separate normal from those with glaucoma.

Discussion

In this study, we focused on automating the diagnosis of glaucoma from fundus images by establishing a DLS (GD-CNN) with an ability to work across numerous populations. Previous studies have reported automated methods for the evaluation of glaucoma with most employing technology on feature extraction[22-26]. Recently, the DLS approach also has been adopted to provide high sensitivity and specificity for detecting GON from high-quality retinal fundus images.[14,27,28] The ambition of deep learning is to create a "fully-automated" screening model, which can automatically learn the features for glaucoma diagnoses without any human effort, avoiding misalignment or/and misclassification caused by introduced errors in the localization and segmentation. Compared with previous work, the GD-CNN model differs from conventional learning-based algorithms in a number of aspects.

The GD-CNN model was trained using a larger dataset than previous studies [13,14,27-32]. It is reasonable to assume that access to a greater pool of training images is likely to increase the accuracy of the DLS to detect glaucoma. A major issue with deep learning algorithms is their general applicability to systems and settings beyond the site of development. To address this

issue, additional data sets were employed. Datasets resulting from ophthalmic settings are likely to provide a higher incidence of glaucoma patients than is present in the general population. Therefore, to provide a realistic disease-screening test for GD-CNN, a population dataset obtained from the Handan Eye Study was employed, which provided a real-world ratio of individuals with and without diagnosed glaucoma[33,34]. Ethnicity can also present different anatomical/clinical features and incidence of glaucoma[35]. A number of the cohorts derived from Chinese centers have limited ethnic diversity. Therefore, to test GD-CNN across a range of ethnic backgrounds a multi-ethnic dataset, which includes White, African American, Asian, and Middle Eastern, from the Hamilton Glaucoma Center was utilized. Despite the different challenges imposed by these different data sets, GD-CNN consistently performed with high sensitivity and specificity. Another major factor that can impact on the generalization of DLSs is the image quality provided on which the DLS is making decisions and diagnosis. To address this important concern, GD-CNN was externally evaluated using a multi-quality image dataset of retinal fundus photographs established from website sources. Examination of 884 images available on the worldwide web using GD-CNN as expected proved a greater challenge, but analysis showed acceptable performance with AUC of 0.823 with 82.2% sensitivity and 70.4% specificity.

The current study addressed the issue of false positive and negative diagnosis by the DLS and manual grading. The main reason for both false-negative and false-positive diagnosis by GD-CNN and manual grading was high or pathologic myopia, which are characterized by peripapillary atrophy (beta-zone), shallow cups, tilting and/or torsion optic disc. More studies assessing textural based properties are planned to allow more accurate classification by the algorithm which can distinguish among the optic disc region, central β-zone and peripheral α-zone of peripapillary atrophy and other retinal areas.

To further evaluate the ability of the GD-CNN model across multiple populations, an ODL system was proposed in which the GD-CNN model iteratively updated with an HCI loop. Consequently, in the ODL system, the generalization ability of GD-CNN can be improved through human-computer interaction, such that each can educate and inform the other. An ODL system using a pre-trained GD-CNN model to reinforce training on limited local images would likely generate a more accurate model requiring less time for local dataset classifications. In principle, the ODL system we have described here could potentially be employed on a wide range of medical images across multiple disciplines. Further benefit may come from the use of AI with digital images like a combination of structural and functional testing, and even multiple other orthogonal datasets, for example, cardiovascular data and genomic data, to further enhance the value of data utilization for the health care system.

Conclusions

The GD-CNN model, which was driven by a large-scale database of fundus images, has high sensitivity and specificity for detecting glaucoma. The experimental results show the potential of automated DLSs in enhancing current screening programs in a cost-effective and time efficient manner. The generalization of this approach might be facilitated by training the GD-CNN model on large-scale data and implementing GD-CNN in an ODL system, which may be further refined through a human computer interface.

1. Tham YC, Li X, Wong TY, et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121(11):2081-2090. doi: 10.1016/j.ophtha.2014.05.013

2. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol.* 2006;90(3):262-267. doi:10.1136/bjo.2005.081224
3. Hood DC, Raza AS, de Moraes CG, et al. Glaucomatous damage of the macula. *Prog Retin Eye Res.* 2013; 32:1-21. doi: 10.1016/j.preteyeres.2012.08.003
4. Tatham AJ, Weinreb RN, Medeiros FA. Strategies for improving early detection of glaucoma: the combined structure-function index. *Clin Ophthalmol.* 2014; 8:611-621. doi: 10.2147/OPTH.S44586
5. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *JAMA.* 2014; 311(18):1901-1911. doi: 10.1001/jama.2014.3192
6. Pizzi LT, Waisbourd M, Hark L, et al. Costs of a community-based glaucoma detection programme: analysis of the Philadelphia Glaucoma Detection and Treatment Project. *Br J Ophthalmol.* 2018; 102(2):225-232. doi: 10.1136/bjophthalmol-2016-310078
7. Zhao D, Guallar E, Gajwani P, et al. Optimizing glaucoma screening in high-risk population: design and 1-year findings of the screening to prevent (SToP) glaucoma study. *Am J Ophthalmol.* 2017; 180:18-28. doi: 10.1016/j.ajo.2017.05.017
8. Fleming C, Whitlock EP, Beil T, et al. Screening for primary open-angle glaucoma in the primary care setting: an update for the US preventive services task force. *Ann Fam Med.* 2005; 3(2):167-170. doi: 10.1370/afm.293
9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017; 542(7639):115-118. doi: 10.1038/nature21056
10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *JAMA.* 2017; 318(22):2199-2210. doi: 10.1001/jama.2017.14585

11. Korfiatis P, Kline TL, Coufalova L, et al. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas. *Med Phys.* 2016; 43(6):2835-2844. doi: 10.1118/1.4948668
12. 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(22):2402-2410. doi: 10.1001/jama.2016.17216
13. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017; 318(22):2211-2223. doi: 10.1001/jama.2017.18152
14. Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology.* 2018; 125(8):1199-1206. doi: 10.1016/j.ophtha.2018.01.023
15. Wong TY, Bressler NM. Artificial Intelligence with Deep Learning Technology Looks into Diabetic Retinopathy Screening. JAMA. 2016; 316(22):2366-2367. doi: 10.1001/jama.2016.17563
16. Castelvecchi D. Can we open the black box of AI? *Nature.* 2016; 538(7623):20-23. doi: 10.1038/538020a
17. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA.* 2018; 319(1):19-20. doi: 10.1001/jama.2017.19198
18. He K, Zhang X, Ren S. et al. Deep residual learning for image recognition. Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*. 2016:770-778. doi: 10.1109/CVPR.2016.90
19. Haleem MS, Han L, van Hemert J. et al. Automatic extraction of retinal features from

colour retinal images for glaucoma diagnosis: A review. *Computerized Medical Imaging and Graphics*. 2013; 37(7-8):581-596. doi: 10.1016/j.compmedimag.2013.09.005

20. Szegedy, C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*. 2015:1-9. doi:10.1109/CVPR.2015.7298594
21. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:818-833. doi: 10.1007/978-3-319-10590-1_53
22. Singh A, Dutta MK, Partha SM, et al. Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. *Comput Methods Programs Biomed.* 2016; 124:108-120. doi: 10.1016/j.cmpb.2015.10.010
23. Issac A, Partha SM, Dutta MK. An adaptive threshold-based image processing technique for improved glaucoma detection and classification. *Comput Methods Programs Biomed.* 2015; 122(2):229-244. doi: 10.1016/j.cmpb.2015.08.002
24. Chakrabarty L, Joshi GD, Chakravarty A, et al. Automated Detection of Glaucoma from Topographic Features of the Optic Nerve Head in Color Fundus Photographs. *J. Glaucoma.* 2016; 25(7):590-597. doi: 10.1097/IJG.0000000000000354
25. Chen X, Xu Y, Wong DWK, et al. Glaucoma detection based on deep convolutional neural network. *Conf Proc IEEE Eng Med Biol Soc.* 2015:715-718. doi: 10.1109/EMBC.2015.7318462
26. Annan L, Jun C, Damon WKW, et al. Integrating holistic and local deep features for glaucoma classification. *Conf Proc IEEE Eng Med Biol Soc.* 2016:1328-1331. doi: 10.1109/EMBC.2016.7590952
27. Christopher M, Belghith A, Bowd C, et al. Performance of Deep Learning

438 Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy
439 in Fundus Photographs. *Sci Rep*. 2018;8(1):16685. doi: 10.1038/s41598-018-35044-9
440 28. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning
441 algorithm to screen for glaucoma from fundus photography. *Sci Rep.* 2018;8(1):14665.
442 doi: 10.1038/s41598-018-33013-w
443 29. Meier J, Bock R, Michelson G, et al. Effects of Preprocessing Eye Fundus Images on
444 Appearance Based Glaucoma Classification. Proceedings of the International
445 Conference on Computer Analysis of Images and Patterns. 2007:165-172. doi:
446 10.1007/978-3-540-74272-2_21
447 30. Bock R, Meier J, Michelson G, et al. Classifying Glaucoma with Image-Based
448 Features from Fundus Photographs. Proceedings of the 29th DAGM Symposium,
449 Heidelberg, Germany, September 12-14, 2007. doi:10.1007/978-3-540-74936-3_36
450 31. Bock R, Meier J, Nyúl L, et al. Glaucoma risk index: automated glaucoma detection
451 from color fundus images. *Med Image Anal.* 2010; 14(3): 471-481. doi:
452 10.1016/j.media.2009.12.006
453 32. Keerthi SS, Shevade SK, Bhattacharyya C, et al. Improvements to Platt's SMO
454 Algorithm for SVM Classifier Design. *Neural Computation*. 2014; 13(3):637-649.
455 doi:10.1162/089976601300014493
456 33. Wang NL, Hao J, Zhen Y, *et al.* A Population-based Investigation of Circadian
457 Rhythm of Intraocular Pressure in Habitual Position Among Healthy Subjects: The
458 Handan Eye Study. *J. Glaucoma*. 2016; 25(7):584-589. doi:
459 10.1097/IJG.0000000000000351
460 34. Zhang Y, Li SZ, Li L, et al. The Handan Eye Study: comparison of screening methods
461 for primary angle closure suspects in a rural Chinese population. *Ophthalmic*
462 *Epidemiol*. 2014; 21(4):268-275. doi: 10.3109/09286586.2014.929707

35. Cho HK, Kee C. Population-based glaucoma prevalence studies in Asians. *Surv Ophthalmol*. 2014; 59(4):434-447. doi: 10.1016/j.survophthal.2013.09.003

Legends:

Figure 1. Visualization of deep features of the GD-CNN deep learning system. Visualization maps generated from deep features, which can be superimposed on the input image to highlight the areas the model considered important in making its diagnosis.

Figure 2. Training loss and visualization of deep features at different training iterations. (A) Training loss with accuracy with training iterations. (B) Feature clustering with the progress of training. The dimensionality of deep features was nonlinearly reduced by t-distributed stochastic neighbor embedding (t-SNE) method for visualization.

Acknowledgements

Funding/Support: The research has received funding from National Natural Science Fund Projects of China (81271005), Beijing municipal administration of hospitals Qingmiao projects (QMS20180210), The priming scientific research foundation for the junior researcher in Beijing Tongren Hospital (2016-YJJ-ZZL-021), Beijing Tongren Hospital Top Talent Training Program. Medical Synergy Science and Technology Innovation Research (Z181100001918035)

Role of the Funder/Sponsor: The funding organizations had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflict of Interest Disclosures: All authors have completed and submitted the authorship forms.

Author Contributions: Drs H. Liu, M. Xu and N. Wang had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Drs M. Xu and N. Wang contributed equally.

Concept and design: H. Liu, M. Xu, N. Wang

Acquisition, analysis, or interpretation of data: H. Liu, L. Li, C. Qiao, C. Zhang, P. Liu, S. Li, H. Wang, D. Mou, R. Pang, D. Yang, L. Jiang, Y. Chen, M. Hu, Y. Xu, H. Kang, X. Ji, C. Tham, C. Cheung

Critical revision of the manuscript for important intellectual content: H. Liu, I. M. Wormstone, D. Yang, R. Chang, D. S. Ting, Z. Wang, T. Y. Wong, M. Xu, R. N. Weinreb

Statistical analysis: H. Liu, D. Yang, L. Li, L. Jiang

Obtained funding: H. Liu, M. Xu, N. Wang

Administrative, technical, or material support: H. Liu, C. Qiao, C. Zhang, P. Liu, S. Li, H. Wang, Z. Wang, M. Xu, N. Wang

Supervision: H. Liu, Z Wang, M. Xu, N. Wang

504 Table 1. Summary of Source Datasets

Source Datasets	Images No.	Eyes[a] No.	Individuals No.	Age, Mean[b] (SD), y	Female[b] No./Total (%)	Cohort	Ethnicity/Race	Camera	Assessor
CGSA	274 413	138 210	69 105	54.1 (14.5)	20 167 (55.8 %)	Clinic-based	Han Chinese (78.3%)	Topcon, Canon, Carl Zeiss	Professional grader team
Beijing Tongren Hospital	20 466	10 308	5 154	52.8 (16.7)	1 068 (49.7 %)	Clinic-based	Han Chinese (81.7%)	Topcon, Canon	2 Ophthalmologists; arbitration by 1 glaucoma specialist
Peking University Third Hospital	12 718	64 60	3 230	57.2 (10.9)	327 (43.1 %)	Clinic-based	Han Chinese (79.5%)	Topcon	2 Ophthalmologists; arbitration by 1 glaucoma specialist
Harbin Medical University First Hospital	9 305	4 732	2 366	59.9 (11.2)	771 (57.3 %)	Clinic-based	Han Chinese (82.9%)	Topcon	2 Professional senior graders; arbitration by 1 glaucoma specialist
Handan Eye Study	29 676	13 404	6 702	55.2 (10.9)	2 589 (42.2 %)	Population-based	Han Chinese (80.1%)	Topcon, Canon	3 Glaucoma specialists
Hamilton Glaucoma Center	7 877	3 938	1 969	58.2 (19.2)	1041 (52.9 %)	Clinic-based	White (73.0%), Black/African American (19.3%), Asian (5.4%), Middle Eastern (0.3%)	Topcon, Canon	3 Glaucoma specialists
Website	884	884	884	N/A	N/A	Website-based	N/A	N/A	2 Professional senior graders; arbitration by 1 glaucoma specialist

[a] For each patient, 2 fundus images were taken and recorded of each eye. [b] Individual data including age sex and ethnicity/race were available for CGSA (52.3%), Beijing Tongren Hospital (41.7%), Peking University Third Hospital (23.5%), Harbin Medical University First Hospital (56.9%), Handan Eye Study (99.6%), Hamilton Glaucoma Center (100%), Website (N/A).

505
506

507 Table 2. The Classification for Glaucomatous Optic Neuropathy

Classification	**Clinical Features**
Unlikely glaucomatous optic neuropathy	With no sign of the following
Probable glaucomatous optic neuropathy	At least two conditions positive: 0.7 ≤ VCDR < 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)
Definite glaucomatous optic neuropathy	Any of the following conditions: VCDR ≥ 0.85; RNFL defects corresponds with thinning area of rim or notches.
VCDR: vertical cup-to-disc ratio. DD: disc diameter RNFL: retinal nerve fiber layer	

508
509

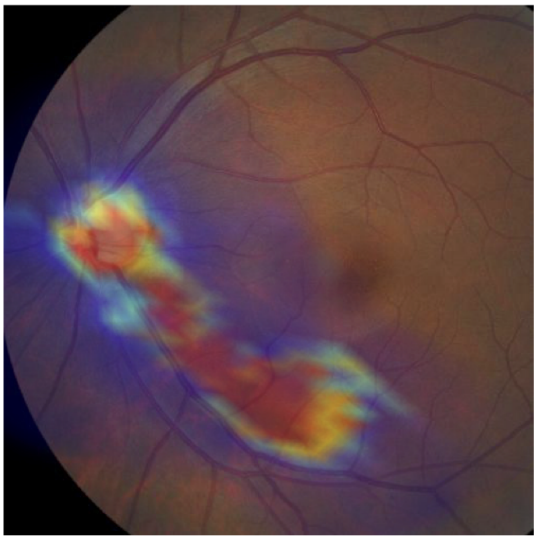
510 Table 3. The Performance of the GD-CNN in Validation Datasets

		% (95% CI)		Confusion Result No. (%)				
Datasets (No. of Images)	AUC (95% CI)	Sensitivity	Specificity	True-Positive	False-Positive	False-Negative	True-Negative	Total Concordant Images
Local Validation								
Chinese Glaucoma Study Alliance (N =28 569)	0.996 (0.995-0.998)	96.2 (95.4 – 96.9)	97.7 (97.5-97.9)	2 786 (9.8)	588 (2.1)	110 (0.4)	25 085 (87.8)	27 871 (97.6)
Clinical Validation								
Beijing Tongren Hospital (N =20 466)	0.995 (0.996-0.996)	96.1 (95.2-96.9)	97.1 (96.8-97.3)	2 226 (10.9)	534 (2.6)	90 (0.4)	17 616 (86.1)	19 842 (97.0)
Peking University Third Hospital (N = 12 718)	0.994 (0.991-0.996)	96.0 (93.9-97.2)	96.1 (95.8-96.5)	593 (4.7)	468 (3.7)	26 (0.2)	11 631 (91.5)	12 224 (96.1)
Harbin Medical University First Hospital (N = 9 305)	0.987 (0.982-0.991)	93.6 (90.9-95.6)	95.6 (95.1-96.0)	435(4.7)	392 (4.2)	30 (0.3)	8 448 (90.8)	8 883 (95.5)
Population Screening Validation								
Handan Eye Study (N = 29 676)	0.964 (0.952-0.972)	91.0 (88.4-93.1)	92.6 (92.2-92.8)	543 (1.8)	2 175 (7.3)	54 (0.2)	26 904 (90.7)	27 447 (92.5)
Multi-ethnic Validation								
Hamilton Glaucoma Center (N=7 877)	0.923 (0.916-0.930)	87.7 (86.8-88.5)	80.8 (78.9-82.5)	5224 (66.3)	369 (4.7)	733 (9.3)	1551 (19.7)	6 775 (86.0)
Multi-quality Validation								
Website (N = 884)	0.823 (0.787-0.855)	82.2 (76.9-86.6)	70.4 (65.8-74.7)	212 (31.0)	126 (18.4)	46 (6.7)	300 (43.9)	512 (74.9)

511
512

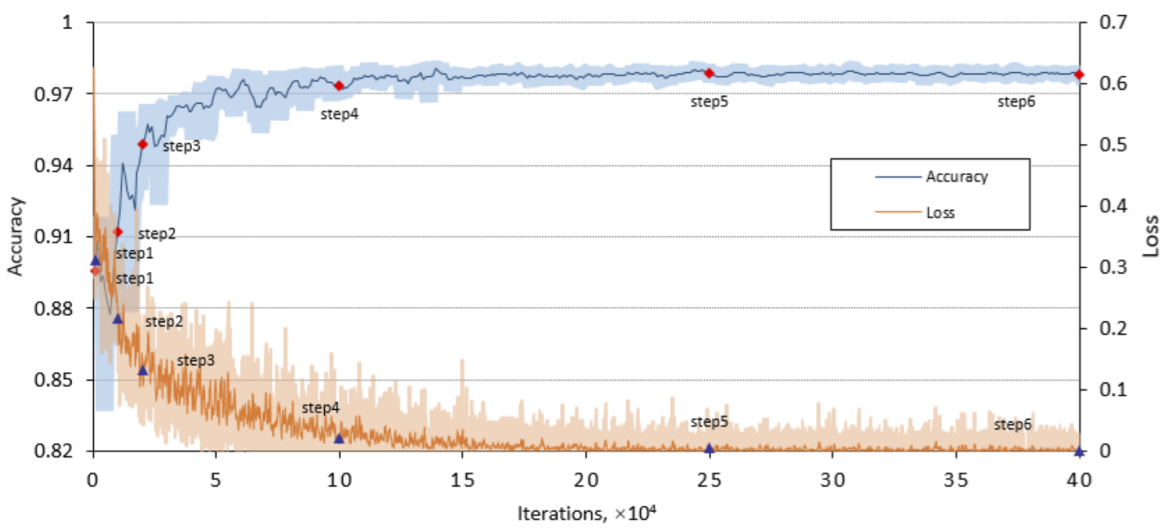
513

Original

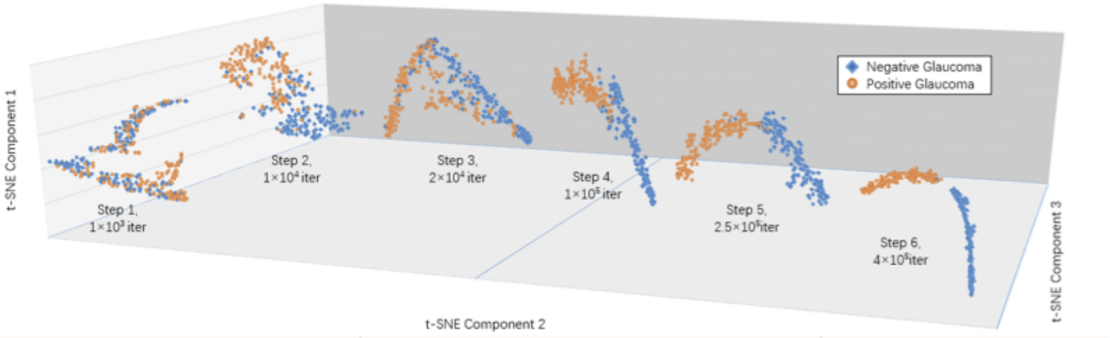


Heatmap

A


Accuracy
Loss
Iterations, ×10⁴
step1
step2
step3
step4
step5
step6

B


Negative Glaucoma
Positive Glaucoma
t-SNE Component 1
t-SNE Component 2
t-SNE Component 3
Step 1, 1×10³ iter
Step 2, 1×10⁴ iter
Step 3, 2×10⁴ iter
Step 4, 1×10⁵ iter
Step 5, 2.5×10⁵iter
Step 6, 4×10⁵iter