JAMA Ophthalmology | Original Investigation

Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs

Hanruo Liu, MD, PhD; Liu Li, BEng; I. Michael Wormstone, PhD; Chunyan Qiao, MD, PhD; Chun Zhang, MD, PhD; Ping Liu, MD, PhD; Shuning Li, MD, PhD; Huaizhou Wang, MD, PhD; Dapeng Mou, MD, PhD; Ruiqi Pang, MD; Diya Yang, MD, PhD; Linda M. Zangwill, PhD; Sasan Moghimi, MD; Huiyuan Hou, MD, PhD; Christopher Bowd, PhD; Lai Jiang, BEng; Yihan Chen, MD; Man Hu, MD, PhD; Yongli Xu, PhD; Hong Kang, PhD; Xin Ji, BEng; Robert Chang, MD, PhD; Clement Tham, MD, PhD; Carol Cheung, PhD; Daniel Shu Wei Ting, MD, PhD; Tien Yin Wong, MD, PhD; Zulin Wang, PhD; Robert N. Weinreb, MD, PhD; Mai Xu, PhD; Ningli Wang, MD, PhD

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

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

DESIGN, SETTING, AND PARTICIPANTS In this cross-sectional study, a DLS for the classification of GON was developed for automated classification of GON using retinal fundus images obtained from the Chinese Glaucoma Study Alliance, the Handan Eye Study, and online databases. The researchers selected 241 032 images were selected as the training data set. The images were entered into the databases on June 9, 2009, obtained on July 11, 2018, and analyses were performed on December 15, 2018. The generalization of the DLS was tested in several validation data sets, 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 patients with glaucoma within the cohort and an additive data set 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 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. A total of 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. The AUC of the GD-CNN model in primary local validation data sets was 0.996 (95% CI, 0.995-0.998), with sensitivity of 96.2% and specificity of 97.7%. The most common reason for both false-negative and false-positive grading by GD-CNN (51 of 119 [46.3%] and 191 of 588 [32.3%]) and manual grading (50 of 113 [44.2%] and 183 of 538 [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 generalizability for detecting GON. These findings suggest that automated DLS could enhance current screening programs in a cost-effective and time-efficient manner.

Invited Commentary page 1361

Supplemental content and Journal Club Slides

CME Quiz at jamanetwork.com/learning and CME Questions page 1472

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Authors: Ningli Wang, MD, PhD, Beijing Tongren Hospital, Capital Medical University; Beijing Institute of Ophthalmology, No.1 Dongjiaominxiang Street, Dongcheng District, Beijing 100730, China (wningli@vip.163.com); Mai Xu, PhD, School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (maixu@buaa.edu.cn).

JAMA Ophthalmol. 2019;137(12):1353-1360. doi:10.1001/jamaophthalmol.2019.3501
Published online September 12, 2019. Corrected on December 12, 2019.

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 patients with glaucoma 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 owing to its convenience and relative affordability.[6,7] Nevertheless, the process of manual image assessment is labor-intensive and time-consuming.[7] 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.[8-12] 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.[13] However, the use of DLS for medical diagnosis has inferior performance when applied to data obtained from different sources.[12,13] This is an important consideration, because if maximum reach and clinical benefit are to be achieved, ideally a DLS would be used in different settings with images of varying quality, patient ethnicity, and population sources.[14-16]

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

Key Points

Question How does a deep learning system compare with professional human graders in detecting glaucomatous optic neuropathy?

Findings In this cross-sectional study, the deep learning system showed a sensitivity and specificity of greater than 90% for detecting glaucomatous optic neuropathy in a local validation data set, in 3 clinical-based data sets, and in a real-world distribution data set. The deep learning system showed lower sensitivity when tested in multiethnic and website-based data sets.

Meaning This assessment of fundus images suggests that deep learning systems can provide a tool with high sensitivity and specificity that might expedite screening for glaucomatous optic neuropathy.

Methods

Training Data Sets

This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline. The study was conducted according to the tenets of the Declaration of Helsinki and it was approved by the institutional review board of Beijing Tongren Hospital. Because the study was a retrospective review and analysis of fully anonymized color retinal fundus images, the medical ethics committee declared it exempt from 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; eAppendix in the Supplement) between 2009 and 2017 (**Table 1**). The images were entered into the databases on June 9, 2009, obtained on July 11, 2018, and analyses were performed on December 15, 2018. The CGSA uses a tele-ophthalmology platform and a cloud-based online data set (http://www.funduspace.com), which has established its own electronic data capture system to achieve effective data quality control. For each patient, 2 fundus images of each eye were recorded. For this study, each image in the training data set 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 label was masked to evaluators. The first tier of graders consisted of 5 trained medical students and nonmedical undergraduates. They conducted initial quality control according to the following criteria: (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 (ie, not too dark or too light); (4) the image was focused sufficiently for grading the optic nerve head and retinal nerve fiber layer. The second tier of graders consisted of 22 Chinese board-certified ophthalmologists or postgraduate ophthalmology trainees, with more than 2 years' experience, who had passed a pretraining test. In the process of grading, each image was assigned randomly to 2 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 2 senior independent glaucoma specialists with more than 10 years of experience with glaucoma diagnosis (H.W. and S.L.); they were consulted to adjudicate disagreement in tier 2 grading (eAppendix in the Supplement). After this process, images were classified as unlikely, probable, and definite GON. Referable GON was defined as probable or definite GON.

Glaucoma Diagnosis With Convoluted Neural Networks Model

The training images with assigned labels were used to establish a state-of-the-art DLS, GD-CNN, based on the Residual Net (ResNet) platform[17] (eFigures 1 and 2 in the Supplement). 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 or high-resolution images, because processing them requires more computational resources, such as memory, and time. Therefore, we preprocessed images by downsampling them to 224 × 224 pixel resolution. In addition, these images were centered on the optic

Table 1. Summary of Source Data Sets

Source Data Sets	No.			Age, Mean (SD), y[b]	Female No./Total (%)[b]	Cohort	Ethnicity/Race, (%)	Camera	Assessor
	Images	Eyes[a]	Individuals						
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	5154	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	6460	3230	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	9305	4732	2366	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	6702	55.2 (10.9)	2 589 (42.2)	Population-based	Han Chinese (80.1)	Topcon, Canon	3 Glaucoma specialists
Hamilton Glaucoma Center	7877	3938	1969	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	NA	NA	Website-based	NA	NA	2 Professional senior graders; arbitration by 1 glaucoma specialist

Abbreviations: CGSA, Chinese Glaucoma Study Alliance; NA, not applicable.

[a] For each patient, 2 fundus images of each eye were taken and recorded.

[b] Individual data, including age, sex, and race/ethnicity, were available for CGSA (36 142 of 69 105 individuals [52.3%]), Beijing Tongren Hospital (2150 of 5154 [41.7%]), Peking University Third Hospital (759 of 3230 [23.5%]), Harbin Medical University First Hospital (1346 of 2366 [56.9%]) (%), Handan Eye Study (6675 of 6702 [99.6%]), and the Hamilton Glaucoma Center (100%). This information was NA for the website.

Table 2. Classification for Glaucomatous Optic Neuropathy

Classification	Clinical Features
Unlikely glaucomatous optic neuropathy	No sign of the conditions below
Probable glaucomatous optic neuropathy	At least 2 conditions positive: 0.7 ≤ VCDR < 0.85; rim width ≤0.1 DD; general rim thinning ≥60° or localized rim thinning <60° (11 to 1 o'clock or 5 to 7 o'clock); RNFL defects; splinter hemorrhages; and peripapillary atrophy (β zone)
Definite glaucomatous optic neuropathy	Any of the following conditions: VCDR ≥ 0.85; RNFL defects correspond with thinning area of rim or notches.

Abbreviations: DD, disc diameter; RNFL, retinal nerve fiber layer; VCDR, vertical cup-disc ratio.

cup and contained part of the surrounding vessels, because glaucoma is highly associated with alteration in these regions.[18] To achieve this, the optic cups were automatically detected by recognition of the area with the highest intensity on the gray-scale map of each fundus image; this was found to consistently be associated with the optic cup. Next, we calculated the mean values of red, green, and blue channels, respectively, among all the fundus images in the training data set. Then, for each sample, we remove the 3 mean values on red, green, and blue channels, such that the input to GD-CNN was approximately 0 for relieving the overfitting.[19] As such, the redundancy of the fundus image could be removed for the binary classification of glaucoma in GD-CNN. Because the GON diagnosis was formulated as a binary classification problem, estimating 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 Data Sets

Details of all validation datasets are described in Table 1 and eTable 1 in the Supplement. The initial local validation data set 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 3 experienced professional graders (D.M., R.P., Y.C.) with more than 2 years' experience in detecting referable GON.

Online Deep Learning 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 grade 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 loop (eFigure 2A in the Supplement). The human-computer interaction loop consisted of 3 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 (ie, return to step 1).

Visualization of Predictive Imaging Features

Following Zeiler and Fergus,[20] we visualized the contributions of different regions to GD-CNN prediction of GON on fundus images. The visualization is represented by heatmaps, which highlight strong prognostic regions of the fundus images. The experiment of occlusion testing was conducted to obtain the visualization results. First, the original fundus image was resized into a 360 × 360 red, green, and blue image. Then, a 60 × 60 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 (31 × 31) visualization testing images, each of which has a 60 × 60 gray block at a different position. 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 heatmap at the corresponding position. Hence, the visualization heat map was 31 × 31. Finally, the heatmap 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. 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. *t*-Distributed stochastic neighbor embedding visualization is a state-of-the-art nonlinear dimensionality reduction method. The deep features from glaucoma and no finding of glaucoma are clustered into 2 groups once the training loss converges. The groups of 2 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 receiver operating characteristic curve (AUC). The 95% CIs for AUC were calculated nonparametrically through logit-transformation–based CIs, which was found to have good coverage accuracy over unbiased samples. In addition to AUC, the sensitivity and specificity of each operating point in ROC curves were also measured with 2-sided 95% CIs. These CIs were calculated as Clopper-Pearson intervals, which are exact intervals based on cumulative probabilities.

Furthermore, to determine whether 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 2 × 2 contingency tables were applied to count the diagnosis changes after ODL, for positive and negative samples, respectively. Then a χ^2-based *P* value was calculated along with the sensitivity/specificity over each validation data set. Statistical significance was set at 2-sided $P < .05$.

All statistical analyses were computed using the Stats Models Python package, version 0.6.1 (http://www.statsmodels.org) and Matlab AUC, version 1.1 (MathWorks).

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, 8762-55 389) and 10 ophthalmologists graded more than 15 000 images. Senior glaucoma specialists adjudicated 13 254 images in which there was disagreement in tier 2 grading. We selected 241 032 images (definite GON, 29 865 [12.4%]; probable GON, 11 046 [4.6%]; and unlikely GON 200, 121 [83%]) from 68 013 patients, 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 3 diagnostic categories was 15.8% definite GON, 2% probable GON, and 82.2% unlikely GON (eTable 1 in the Supplement). In the local validation data set, the AUC of the GD-CNN model was 0.996 (95% CI, 0.995-0.998), and sensitivity and specificity in detecting referable GON were comparable with that of trained professional graders (sensitivity, 96.2% vs 96.0%; $P = .76$; specificity, 97.7% vs 97.9%; $P = .81$) (eFigure 3 in the Supplement). To evaluate the ability of the GD-CNN to work across different populations, 3 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 with clinical evaluation, performance remained high (**Table 3**); the AUC for referable GON ranged from 0.995 to 0.987, with both sensitivity and specificity greater than 90% (range: sensitivity, 93.6% to 96.1%; specificity, 95.6% to 97.1%). Further evaluation was undertaken using the Handan Eye Study data set to provide a real-world distribution of individuals with glaucoma. In this case, the AUC was 0.964 with a sensitivity of 91.0% and specificity of 92.6% (Table 3). To test GD-CNN across a range of ethnic backgrounds, a multiethnic data set (73.0% white, 19.3% African American, 5.4% Asian, 0.3% Middle Eastern) from the Hamilton Glaucoma Center was used, with an AUC of 0.923, sensitivity of 87.7%, and specificity 80.8%. Glaucoma Diagnosis With Convoluted Neural Networks showed an AUC of 0.823 with 82.2% sensitivity and 70.4% specificity in a data set composed of images of a varied range of quality obtained online (Table 3).

Understanding the Basis for Incorrect Diagnosis

Among the local validation data sets, an additional analysis was conducted to further evaluate GD-CNN's performance to better establish the basis for false-positive and false-negative diagnoses (eTable 2 in the Supplement). The most common reason for undetected GON from fundus images was pathologic or high myopia for both GD-CNN (51 of 110 [46.3%]) and manual grading (50 of 113 [44.2%]). The most likely cause for a false-

Table 3. Performance of the GD-CNN in Validation Data Sets

Data Sets (No. of Images)	AUC (95% CI)	% (95% CI)		Confusion Result No. (%)				
		Sensitivity	Specificity	True-Positive	False-Positive	False-Negative	True-Negative	Total Concordant Images
Local Validation								
Chinese Glaucoma Study Alliance (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 (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 (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 (9305)	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 (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)
Multiethnic Validation								
Hamilton Glaucoma Center (7877)	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)
Multiquality Validation								
Website (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)

Abbreviations: AUC, area under the receiver operating characteristic curve; GD-CNN, Glaucoma Diagnosis with Convoluted Neural Networks.

positive classification by DLS or manual grading was also pathologic or high myopia (DLS: 191 of 588 [32.3%]; manual: 183 of 538 [34.0%]). Physiologically large cupping was also a common cause of false-positive results with manual diagnosis (138 of 538 [25.6%]), and to a lesser degree with GD-CNN (94 of 588 [16.0%]).

Implementation of the Online Deep Learning System

The ODL system was implemented in the tele-ophthalmic image reading platform of Beijing Tongren Hospital (eAppendix in the Supplement), which collected a group of fundus images every week (approximately 600 images). It was found that both sensitivity and specificity of the ODL system improve with each group of samples collected sequentially across a 5-week period (eFigure 2 in the Supplement). Specifically, the improvement in sensitivity was 1.3%, 2.6%, 2.6%, and 3.9%, respectively, and the improvement of specificity was 2.0%, 2.4%, 2.1%, and 2.6%.

Visualization of Prediction

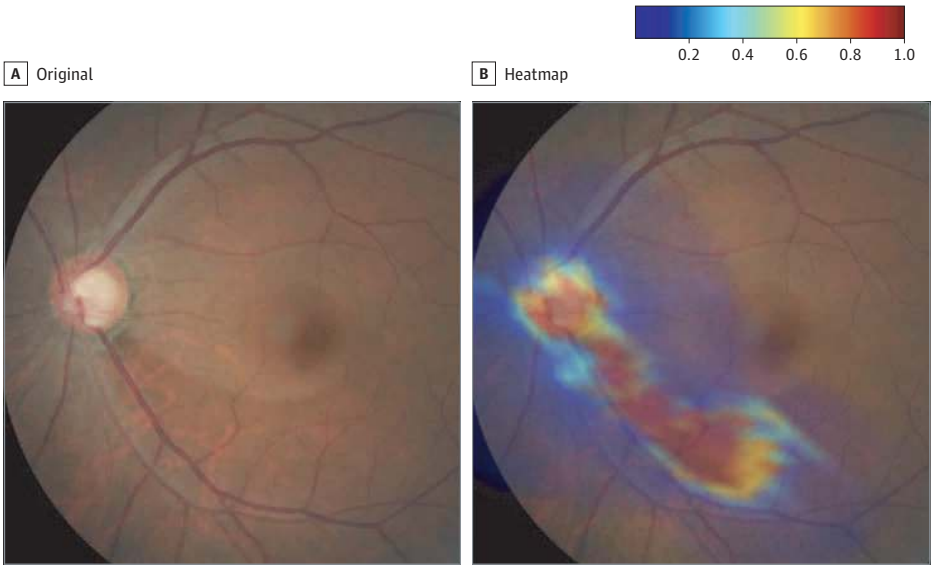
To visualize the learning procedure and represent the areas contributing most to the DLS, we created a heatmap that superimposed a convolutional visualization layer at the end of our network, performed on 1000 images (**Figure 1**; eFigure 4 in the Supplement). The regions of interest identified to have made the greatest contribution to the neural network's diagnosis were also shared with 91.8% of ophthalmologists (**Figure 2**A). All areas containing optic nerve head variance and neuroretinal rim loss were located correctly on all the images used for testing, whereas retinal nerve fiber layer defects and peripapillary atrophy 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 using technology on feature extraction.[21-25] Recently, the DLS approach also has been adopted to provide high sensitivity and specificity for detecting GON from high-quality retinal fundus images.[2,26,27] 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 and misclassification caused by introduced errors in localization and segmentation. Compared with previous work, the GD-CNN model differs from conventional learning-based algorithms in a number of aspects.

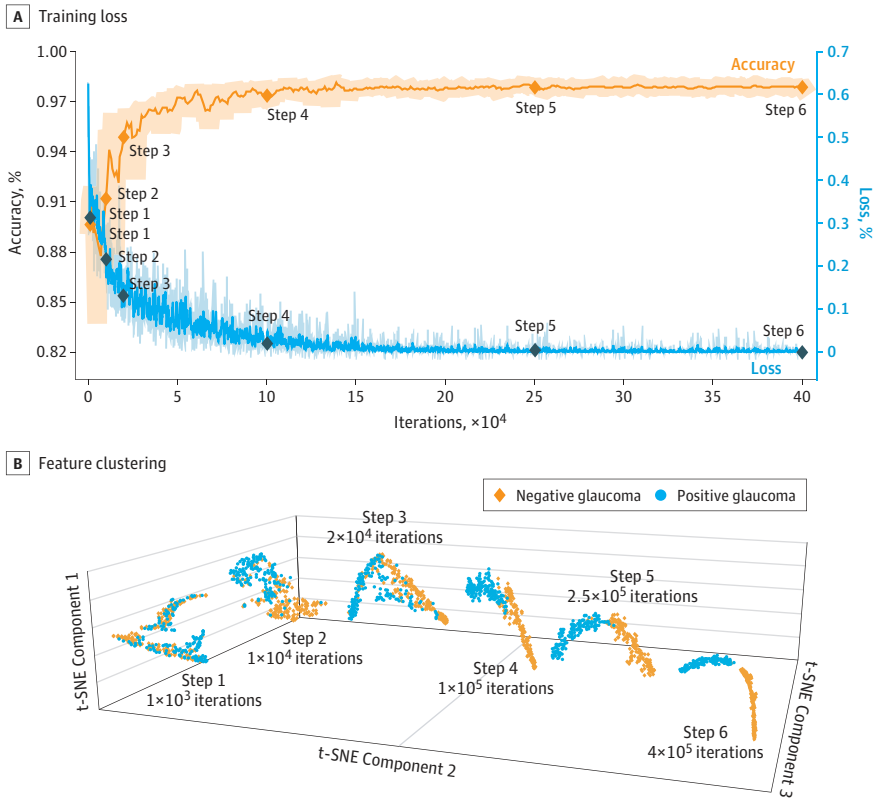
The GD-CNN model was trained using a larger data set than used in previous studies.[12,13,26-31] It is reasonable to assume that access to a greater pool of training images is likely to increase the accuracy of the DLS in detecting glaucoma. A major challenge with deep learning algorithms is their general applicability to systems and settings beyond the site of development. To address this challenge, additional data sets were used. Data sets resulting from ophthalmic settings are likely to produce a higher incidence of glaucoma than is present in the general population. Therefore, to provide a realistic disease-screening test for GD-CNN, a population data set obtained from the Handan Eye Study was used, which provided a real-world ratio of individuals with and without diagnosed glaucoma.[32,33] Ethnicity is also associ-

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 of the model important for 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 the *t*-distributed stochastic neighbor embedding (*t*-SNE) method for visualization.

ated with different anatomical and clinical features and a different incidence of glaucoma.[34] 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 multiethnic data set from the Hamilton Glaucoma Center, which includes white, African American, Asian, and Middle Eastern individuals, was used. Despite the different challenges imposed by these different data sets, GD-CNN consistently performed with high sensitivity and specificity. Another major factor in the generalization of DLSs is the quality of images on which the DLS is making decisions and diagnosis. To address this important concern, GD-CNN was externally evalu-

ated using a multiquality image data set of retinal fundus photographs established from website sources. Examination of 884 images available on the World Wide 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 false-negative diagnosis by the DLS and manual grading. The main reason for both false-negative and false-positive diagnoses by GD-CNN and manual grading was high or pathologic myopia, which are characterized by peripapillary atrophy (β-zone), shallow cups, and tilting, torsion, or both of the optic disc. More studies assessing textural properties are planned to allow more accurate classification by the algorithm to allow it to 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 a human-computer interaction loop.

Limitations

This study has some limitations. 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 pretrained GD-CNN model to reinforce training on limited local images would likely generate a more accurate model requiring less time for local data set classifications. In principle, the ODL system we have described here could potentially be used on a wide range of medical images across multiple disciplines. Further benefit may come from the use of artificial intelligence with digital images in a combination of structural and functional testing, and even multiple other orthogonal data sets, for example, cardiovascular data and genomic data, to further enhance the value of data use 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.

ARTICLE INFORMATION

Accepted for Publication: July 14, 2019.

Published Online: September 12, 2019. doi:10.1001/jamaophthalmol.2019.3501

Correction: This article was corrected on December 1, 2019, to fix an error in the byline.

Author Affiliations: Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing, China (H. Liu, Qiao, S. Li, H. Wang, Mou, Pang, Yang, Chen, N. Wang); Beijing Ophthalmology and Visual Science Key Lab, Beijing, China (H. Liu, Qiao, S. Li, H. Wang, Mou, Pang, Yang, Chen, N. Wang); School of Electronic and Information Engineering, Beihang University, Beijing, China (L. Li, Jiang, Z. Wang, M. Xu); School of Biological Sciences, University of East Anglia, Norwich, United Kingdom (Wormstone); Department of Ophthalmology, Peking University Third Hospital, Beijing, China (Zhang); Ophthalmology Hospital, First Hospital of Harbin Medical University, Harbin, Heilongjiang, China (P. Liu); Department of Ophthalmology, Beijing Children's Hospital, Capital Medical University, Beijing, China (Hu); Department of Mathematics, Beijing University of Chemical Technology, Beijing, China (Y. Xu); College of Computer Science, Nankai University, Tianjin, China (Kang); Beijing Shanggong Medical Technology Co., Ltd, Beijing, China (Ji); Department of Ophthalmology, Byers Eye Institute at Stanford University, Palo Alto, California (Chang); Department of Ophthalmology and Visual Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Kowloon, Hong Kong, China (Tham, Cheung); Singapore Eye Research Institute, Singapore National Eye Center, Singapore (Ting, Wong); Shiley Eye Institute, University of California, San Diego, La Jolla, California (Zangwill, Moghimi, Hou, Bowd, Weinreb).

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 to this work.
Concept and design: H. Liu, Qiao, Zhang, H. Wang, Yang, Moghimi, Lai, Hu, Y. Xu, Kang, Ji, Tham, Ting, Wong, Z. Wang, M. Xu, N. Wang.
Acquisition, analysis, or interpretation of data: H. Liu, L. Liu, Wormstone, Qiao, P. Liu, Li, H. Wang, Mou, Pang, Yang, Zangwill, Hou, Bowd, Chen, Chang, Tham, Cheung, Wong, Weinreb, M. Xu.
Drafting of the manuscript: H. Liu, L. Liu, Qiao, Zhang, Li, Pang, Chen, Tham, M. Xu.
Critical revision of the manuscript for important intellectual content: H. Liu, Wormstone, Qiao, P. Liu, H. Wang, Mou, Yang, Zangwill, Moghimi, Hou, Bowd, Lai, Hu, Y. Xu, Kang, Ji, Chang, Tham, Cheung, Ting, Wong, Z. Wang, Weinreb, M. Xu, N. Wang.
Statistical analysis: H. Liu, L. Liu, Pang, Lai, Chen, Tham, Wong, M. Xu.
Obtained funding: H. Liu, Mou, Pang, Zangwill, Chen, Weinreb.
Administrative, technical, or material support: L. Liu, Qiao, Zhang, P. Liu, Li, H. Wang, Yang, Zangwill, Moghimi, Hou, Bowd, Chen, Hu, Y. Xu, Kang, Ji, Tham, Cheung, Z. Wang, Weinreb, M. Xu.
Supervision: P. Liu, Chen, Tham, Ting, Wong, Z. Wang, M. Xu, N. Wang.

Conflict of Interest Disclosures: Dr Zangwill reports grants from the National Eye Institute during the conduct of the study and research and equipment support from Heidelberg Engineering, Optovue, Carl Zeiss Meditec, and Topcon. Dr Ting reported having a patent pending for a deep learning system for retinal diseases, not related to this work. Dr Wong reported receiving personal fees from Allergan, personal fees from Bayer, personal fees from Boehringer Ingelheim, personal fees from Genentech, personal fees from Merck, personal fees from Novartis, personal fees from Oxurion, and personal fees from Roche outside the submitted work and he is a shareholder in Plano and EyRIS. No other disclosures were reported.

Funding/Support: The research has received funding from the 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 (Dr H. Liu; 2016-YJJ-ZZL-021), Beijing Tongren Hospital Top Talent Training Program, and 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.

REFERENCES

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. 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, Liebmann JM, Ritch R. 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. Zhao D, Guallar E, Gajwani P, et al; SToP Glaucoma Study Group. 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

7. Fleming C, Whitlock EP, Beil T, Smit B, Harris RP. 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

8. 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

9. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al; the CAMELYON16 Consortium. 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

10. 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

11. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

12. 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

13. Li Z, He Y, Keel S, Meng W, Chang RT, He M. 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

14. 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

15. Castelvecchi D. Can we open the black box of AI? *Nature*. 2016;538(7623):20-23. doi:10.1038/538020a

16. 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

17. 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*. Boston, MA: IEEE; 2016:770-778. doi:10.1109/CVPR.2016.90

18. Haleem MS, Han L, van Hemert J, Li B. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: a review. *Comput Med Imaging Graph*. 2013;37(7-8):581-596. doi:10.1016/j.compmedimag.2013.09.005

19. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA: IEEE; 2015:1-9. doi:10.1109/CVPR.2015.7298594

20. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. *Computer Vision—ECCV 2014*. Cham, Switzerland: Springer; 2014:818-833. doi:10.1007/978-3-319-10590-1_53

21. Singh A, Dutta MK, ParthaSarathi M, Uher V, Burget R. 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

22. Issac A, Partha Sarathi M, 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

23. Chakrabarty L, Joshi GD, Chakravarty A, Raman GV, Krishnadas SR, Sivaswamy J. 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

24. Xiangyu Chen, Yanwu Xu, Jiang Liu, Damon Wing Kee Wong, Tien Yin Wong. Glaucoma detection based on deep convolutional neural network. *Conf Proc IEEE Eng Med Biol Soc*. 2015;2015:715-718. doi:10.1109/EMBC.2015.7318462

25. Annan Li, Jun Cheng, Jiang Liu, Damon Wing Kee Wong. Integrating holistic and local deep features for glaucoma classification. *Conf Proc IEEE Eng Med Biol Soc*. 2016;2016:1328-1331. doi:10.1109/EMBC.2016.7590952

26. Christopher M, Belghith A, Bowd C, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci Rep*. 2018;8(1):16685. doi:10.1038/s41598-018-35044-9

27. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. 2018;8(1):14665. doi:10.1038/s41598-018-33013-w

28. Meier J, Bock R, Michelson G, et al. Effects of preprocessing eye fundus images on appearance based glaucoma classification. *Proceedings of the International Conference on Computer Analysis of Images and Patterns*. Berlin, Germany: Springer; 2007:165-172. doi:10.1007/978-3-540-74272-2_21

29. Bock R, Meier J, Michelson G, et al. Classifying glaucoma with image-based features from fundus photographs. In: Hamprecht FA, Schnörr C, Jähne B, eds. *Pattern Recognition. DAGM 2007*. Heidelberg, Germany: Springer; 2007. doi:10.1007/978-3-540-74936-3_36

30. Bock R, Meier J, Nyúl LG, Hornegger J, Michelson G. Glaucoma risk index: automated glaucoma detection from color fundus images. *Med Image Anal*. 2010;14(3):471-481. doi:10.1016/j.media.2009.12.006

31. Keerthi SS, Shevade SK, Bhattacharyya C, et al. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput*. 2014;13(3):637-649. doi:10.1162/089976601300014493

32. Wang NL, Hao J, Zhen Y, et al. A population-based investigation of circadian rhythm of intraocular pressure in habitual position among healthy subjects: the Handan Eye Study. *J Glaucoma*. 2016;25(7):584-589. doi:10.1097/IJG.0000000000000351

33. Zhang Y, Li SZ, Li L, Thomas R, Wang NL. The Handan Eye Study: comparison of screening methods for primary angle closure suspects in a rural Chinese population. *Ophthalmic Epidemiol*. 2014;21(4):268-275. doi:10.3109/09286586.2014.929707

34. 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