Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs

Zhixi Li, MD,1,* Yifan He, BS,2,* Stuart Keel, PhD,3,* Wei Meng, BS,2 Robert T. Chang, PhD,4 Mingguang He, MD, PhD1,3

Purpose: To assess the performance of a deep learning algorithm for detecting referable glaucomatous optic neuropathy (GON) based on color fundus photographs.

Design: A deep learning system for the classification of GON was developed for automated classification of GON on color fundus photographs.

Participants: We retrospectively included 48,116 fundus photographs for the development and validation of a deep learning algorithm.

Methods: This study recruited 21 trained ophthalmologists to classify the photographs. Referable GON was defined as vertical cup-to-disc ratio of 0.7 or more and other typical changes of GON. The reference standard was made until 3 graders achieved agreement. A separate validation dataset of 8000 fully gradable fundus photographs was used to assess the performance of this algorithm.

Main Outcome Measures: The area under receiver operator characteristic curve (AUC) with sensitivity and specificity was applied to evaluate the efficacy of the deep learning algorithm detecting referable GON.

Results: In the validation dataset, this deep learning system achieved an AUC of 0.986 with sensitivity of 95.6% and specificity of 92.0%. The most common reasons for false-negative grading (n = 87) were GON with coexisting eye conditions (n = 44 [50.6%]), including pathologic or high myopia (n = 37 [42.6%]), diabetic retinopathy (n = 4 [4.6%]), and age-related macular degeneration (n = 3 [3.4%]). The leading reason for false-positive results (n = 480) was having other eye conditions (n = 458 [95.4%]), mainly including physiologic cupping (n = 267 [55.6%]). Misclassification as false-positive results amidst a normal-appearing fundus occurred in only 22 eyes (4.6%).

Conclusions: A deep learning system can detect referable GON with high sensitivity and specificity. Coexistence of high or pathologic myopia is the most common cause resulting in false-negative results. Physiologic cupping and pathologic myopia were the most common reasons for false-positive results.

Gliaoma is a leading cause of irreversible blindness worldwide.1–3 A recent global meta-analysis of 50 population-based studies reported the pooled glaucoma prevalence (age range, 40–80 years) to be 3.5%,3 corresponding to an estimated 64.3 million individuals worldwide. As a result of population growth and ageing, this figure is expected to increase to 112 million by 2040.4 Most vision loss resulting from glaucoma is avoidable through early detection and treatment strategies.4–6 Despite this, approximately 85% of cases among the Singapore Chinese, the same rate for African American population of United States, and even an overall rate of 50% among the cases in the United States are undiagnosed.7–12 High rates of undiagnosed disease can be attributed to chronic glaucoma often being asymptomatic until central visual acuity is affected in the advanced stages of disease. As glaucoma advances from the early to late stage, care costs increase by 4-fold, posing a significant financial burden worldwide.13 The assessment of the optic disc and retinal nerve fiber layer (RNFL) are the foundation of glaucoma diagnosis, although a dilated clinical fundus examination after mydriasis has been recommended for its advantage of offering a stereoscopic view of the optic disc.14 However, monoscopic optic disc photographs offer some key advantages, including convenience and affordability. Furthermore, the Glaucomatous Optic Neuropathy Evaluation project demonstrated that subjective assessments of monoscopic optic disc photographs provide an equal diagnostic accuracy for glaucoma when compared with stereoscopic photographs.15 Nevertheless, manual assessment of the optic disc is labor intensive and highly dependent on image interpretation by trained specialists. This significantly impacts the cost effectiveness of glaucoma screening,16–18 and as a result, glaucoma screening strategies are not widely implemented in the general population.19–20
Given the growing public health concern, improvements in screening methods for glaucoma are warranted. An emerging area of diagnostic imaging in ophthalmology involves the use of automated computer vision image interpretation with deep learning algorithms, which involves training the algorithm on large datasets of labelled images so that it can learn the features from the data itself rather than from predefined rules. Recent evidence suggests that deep learning algorithms can grade images with excellent diagnostic accuracy in identifying conditions such as diabetic retinopathy. If this technology could be adopted to provide accurate assessment of glaucoma, there are significant potential benefits including an increased accessibility and affordability of glaucoma screening for specific and at-risk populations, thus improving access to care and decreasing the cost of glaucoma screening especially in remote and underserved communities. The purpose of this study was to evaluate the efficacy of a newly developed deep learning algorithm for the detection of referable glaucomatous optic neuropathy (GON) from monoscopic color fundus photographs.

Methods

In the current study, 70,000 fundus photographs were downloaded by random sampling from the online dataset LabelMe (Healgoo Ltd. LabelMe dataset; 2016. http://www.labelme.org. Accessed February 16, 2016.), which contains more than 200,000 color fundus photographs collected from various clinical settings in China. Subsequently, 48,116 images with a visible optic disc were selected for the labelling of GON.

Approval from the institutional review board of Zhongshan Ophthalmic Center was obtained (identifier, 2017KYPJ049), and this project was conducted according to the tenets of the Declaration of Helsinki. Because of the retrospective nature and fully anonymized usage of images in this study, the review board indicated that the informed consent was not required.

Fundus Photograph Grading, Quality Control, and Reference Standard

Figure 1A illustrates the process of image grading. Twenty-seven licensed ophthalmologists were invited to grade the images. During the training of ophthalmologists, 4 sets each with 100 images (30 had suspect GON or worse) were used for the test. The results of graders were compared with those of 3 senior glaucoma specialists (M.H., R.C., and X.L.), and participants passed the training until they achieved an unweighted k value of 0.75 or more in any test set. Thus, 21 ophthalmologists qualified as graders to classify these images in an online grading system from September 2016 through March 2017.

Table 1 shows the details of the criteria of GON grading, which was classified into unlikely, suspect, and certain GON based on the criteria of previous population studies. The field of the retinal fundus photograph and image quality also were included in the grading. Poor location was assigned to a photograph when the optic disc was not fully visible. Poor quality was used when vessels...
Table 1. The Classification for Glaucomatous Optic Neuropathy in the Online System

<table>
<thead>
<tr>
<th>Classification</th>
<th>Presence of Clinical Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlikely</td>
<td>Does not meet any of the following criteria</td>
</tr>
<tr>
<td>Suspect</td>
<td>Any criterion of the following:</td>
</tr>
<tr>
<td></td>
<td>0.7 ≤ VCDR &lt; 0.9</td>
</tr>
<tr>
<td></td>
<td>0.05 DD &lt; rim width ≤ 0.1 DD</td>
</tr>
<tr>
<td></td>
<td>RNFL defect</td>
</tr>
<tr>
<td></td>
<td>Disc hemorrhage</td>
</tr>
<tr>
<td>Certain</td>
<td>Any criterion of the following:</td>
</tr>
<tr>
<td></td>
<td>VCDR ≥ 0.9</td>
</tr>
<tr>
<td></td>
<td>Rim width ≤ 0.05 DD or localized notches</td>
</tr>
<tr>
<td></td>
<td>RNFL defect corresponds to narrowing of rim or localized notches</td>
</tr>
<tr>
<td>Poor quality</td>
<td>Any criterion of the following:</td>
</tr>
<tr>
<td></td>
<td>Vessels within 1 DD of the optic disc margin cannot be identified</td>
</tr>
<tr>
<td></td>
<td>≥50% of the area is obscured</td>
</tr>
<tr>
<td>Poor location</td>
<td>Only part of optic disc is visible in the image</td>
</tr>
</tbody>
</table>

DD = disc diameter; RNFL = retinal nerve fiber layer; VCDR = vertical cup-to-disc ratio.

Within 1 disc diameter of the optic disc margin could not be identified or 50% or more of the area was obscured. Retinal photographs of poor quality and poor location were regarded as ungradable for referable GON and were excluded from the training and validation dataset. Referable GON was defined as suspect or certain GON.

In the process of online grading, each image was assigned randomly to a single ophthalmologist for initial grading and was assigned sequentially to other individual graders until 3 consistent grading outcomes were achieved. This specific grading outcome was considered as the conclusive grading for a given image.

Image Processing and Development of the Deep Learning Algorithm

Figure 1B displays the deep learning pipeline. The training dataset was used to train the deep learning network, and the validation dataset was used for evaluation of the algorithm. Because a large number and variation of images were collected, we performed several preprocessing steps to normalize the images for variation in our database. Image pixel values were scaled to values in a range of 0 through 1 and then downsized to a 299 × 299 matrix. Local space average color was subtracted to solve the issue of color constancy, which indicates a human observer could see the color of an object with consistency, even if the light reflected by an object changed with the illumination type. Data augmentation, a method of image transformations within a dataset to enhance image heterogeneity, but to retain prognostic features in the image itself, was performed to enhance the dataset, including random horizontal shifts of 0 to 3 pixels and random rotations of 90°, 180°, and 270°. We adopted Inception-v3 architecture, which is a convolutional neural network composed of 11 inception modules. A diagram of the deep learning architecture is shown in Figure 2. A minibatch gradient descent of size 32 was used for training, with an Adam optimizer learning rate of 0.002 for better convergence.

Characteristics of Misclassification by the Deep Learning Algorithm

To characterize misclassified photographs, an experienced ophthalmologist (Z.L.) reviewed them and categorized false-negative images according to the 3 most commonly seen features arbitrarily developed by a consensus meeting: (1) eyes with coexisting eye conditions, such as pathologic or high myopia, diabetic retinopathy, and age-related macular degeneration; (2) eyes without other eye conditions; and (3) others (only presence of RNFL defect alone or optic disc hemorrhage). Similarly, the false-positive photographs were classified into the following arbitrary features: (1) eyes with other eye conditions, such as physiologic cupping, pathologic or high myopia, nonglaucomatous optic atrophy, congenital optic disc abnormalities, and other retinal diseases; and (2) eyes with a normal fundus.

Statistical Analyses

The accuracy of each grader was calculated as a proportion of the grading results that matched the conclusive grading outcome over the total number of images graded by that individual. We used the area under the receiver operating characteristic curve (AUC) with 95% confidence intervals to evaluate the performance of this algorithm. Accuracy, sensitivity, and specificity of the system also were evaluated. All statistical analyses were performed using Stata software version 14.0 (Stata Corp., College Station, TX).

Results

A total of 48,116 fundus photographs were graded for glaucoma by program-trained ophthalmologists, with each fundus
The most common reason for false-negative classification was undetected GON with other eye conditions (n = 15) for referable GON. Recently, several reports of automated methods for identifying referable GON based on 48,116 fundus photographs was investigated. This deep learning algorithm showed a robust performance (AUC, 0.986; sensitivity, 95.6%; and specificity, 92.0%) for the detection of referable GON. Recently, several reports of automated methods for the evaluation of glaucoma have been published. \(^{30–34}\) Singh et al.\(^{30}\) developed a technique using wavelet feature extraction techniques from segmented optic discs on 44 fundus photographs and validated it in 19 images, achieving an accuracy of 94.7%. Isaac et al.\(^{32}\) described an accuracy rate of 94.1% with 100.0% sensitivity by an adaptive threshold-based image processing method on 67 fundus photographs. On a total of 2252 fundus images, Chakrabarty et al.\(^{33}\) reported an AUC of 0.792 with a sensitivity of 71.6% and a specificity of 71.7% using a feature extraction technique. Xiangyu et al.\(^{33}\) and Annan et al.\(^{34}\) developed a deep learning algorithm with combined feature extraction and reported AUCs of 0.887 and 0.838, respectively, based on a small sample of fundus images (n < 1250).

Our study had a number of unique differences when compared with the previous studies. First, most previous studies\(^{30,31,33,34}\) used technology on feature extraction, which inevitably would introduce errors in the localization and segmentation that would lead to misalignment and misclassification.\(^{34,35}\) The supervised deep learning technologies adopted in this study avoid such problems through learning predictive features directly from the global labeled images. Second, our training and validation sets were substantially larger than those of previous reports (all the above previous datasets are smaller than 2500). An insufficient sample size for the training and validation datasets compromises the accuracy to detect glaucoma on a large scale. Finally, the large dataset in this study consisted of data collected from a variety of clinical settings, and therefore is more representative of the real world.

Overfitting, whereby the algorithm performs well on the training dataset but generalizes inappropriately to unseen cases, may occur in machine learning, including deep learning. It is more common when the training sample size is small. The network in this study was built with a dropout technique in the third to last layer. Dropout is a method to drop units randomly with their connections during the training process, and it has been proven to be an effective tool to reduce overfitting.\(^{36}\) Overall, the system in this study used a large dataset, dropout, and model regularization method to help reduce the possibility of this issue.\(^{35,36}\) The evaluation of the training curve could be used to assess the overfitting possibility. The training curve (Fig 6) may suggest that this model is reasonably well fitted. If overfitting occurs, the loss of validation data increases, while training loss steadily decreases. In contrast, best-fitted curves indicate that
the validation loss has its global minimum. Although our results were insightful, certainly a further study also is expected to validate this algorithm using a larger-scale dataset.

The deep learning system strives to differentiate the images into those with or without referable GON. When looking into the reasons for false-negative classification, approximately half of all false-negative cases were a result of confounding optic disc features secondary to high myopia or pathologic myopia (42.6%). These optic discs usually are characterized by shallow cups, optic disc tilting, torsion, and peripapillary atrophy, which are also difficult for ophthalmologists or human graders to identify glaucoma even when visual field data is available.\(^{38}\) The vast majority of remaining false-negative cases (37.9%) were a result of true misclassification, where a GON image without coexisting eye disease or secondary disc change was misclassified as normal. An ideal screening program should minimize the number of false-negative results. We expect more studies to explore how this happened and to identify strategies to minimize errors.

### Table 3. The Proportion of Reasons for False-Negative Classification by the Deep Learning Algorithm

<table>
<thead>
<tr>
<th>Reason</th>
<th>No.</th>
<th>Proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>With coexisting eye conditions</td>
<td>44</td>
<td>50.6</td>
</tr>
<tr>
<td>Pathologic or high myopia</td>
<td>37</td>
<td>42.6</td>
</tr>
<tr>
<td>Diabetic retinopathy</td>
<td>4</td>
<td>4.6</td>
</tr>
<tr>
<td>AMD</td>
<td>3</td>
<td>3.4</td>
</tr>
<tr>
<td>Without other eye conditions</td>
<td>33</td>
<td>37.9</td>
</tr>
<tr>
<td>Others (RNFL defect alone or optic disc hemorrhage only present)</td>
<td>10</td>
<td>11.5</td>
</tr>
<tr>
<td>Total</td>
<td>87</td>
<td>100</td>
</tr>
</tbody>
</table>

AMD = age-related macular degeneration; RNFL = retinal nerve fiber layer.

**Figure 4.** Fundus photographs showing typical false-negative cases. A, Glaucomatous optic neuropathy in pathologic myopia. B, Glaucomatous optic neuropathy accompanied with diabetic retinopathy. C, Glaucomatous optic neuropathy without other eye conditions. D, Retinal nerve fiber layer defect alone.
In any screening program, false-positive results create unnecessary referral and burden to the healthcare system. In the present study, nearly all the false-positive cases (95.4%) were in fact a result of abnormalities of the optic disc or retina not related to GON. More than half of the eyes with false-positive results showed physiologic large cupping that also required further investigation by eye care professionals. Therefore, the increased workload resulting from false-positive classification seems to be reasonable given that most such eyes would benefit from further clinical investigation.

The key to glaucoma diagnosis is to recognize morphologic changes of optic disc and RNFL as a result of progressive retinal ganglion cell loss. However, functional abnormalities are presented first in some patients. In most cases, a single isolated observation cannot always confirm the diagnosis of glaucoma even in clinical practice. Therefore, in the future, to improve the accuracy of the deep learning algorithm further, one should ensure that the clinical labeling (classification) is as close as possible to the so-called ground truth. To improve the accuracy of glaucoma classification, more real-world clinical data, such as visual field or longitudinal changes, or even a history of glaucoma surgery, should be included in the clinical labeling of the images.

The strengths of this study include recruiting more than 20 validated ophthalmologists, multiple interpretations of each image, a large image dataset, and a high accuracy of classification. Limitations of this study also should be considered. First, although we used a large set of fundus images, these images were collected only from hospitals in China. Further research is warranted to investigate the performance of the algorithm among other ethnic groups in different settings. In addition, as is the nature of supervised deep learning, the classification of glaucoma was from the global image labeling, rather than direct definitions of local features. It is unclear what exact features were seen and learned by the network. Further work on the visualization of deep learning networks may help patients, physicians, and health care providers understand better the process of its learning.40 Overall, the current study demonstrated that deep learning can be applied to create an algorithm that is capable of identifying referable GON with high sensitivity and specificity in a large dataset. Further studies are required to explore the usefulness of this algorithm deployed in different population settings and for different ophthalmic conditions.

<table>
<thead>
<tr>
<th>Reasons</th>
<th>No.</th>
<th>Proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>With other eye conditions</td>
<td>458</td>
<td>95.4</td>
</tr>
<tr>
<td>Physiologic large cupping</td>
<td>267</td>
<td>55.6</td>
</tr>
<tr>
<td>Pathologic or high myopia</td>
<td>90</td>
<td>18.8</td>
</tr>
<tr>
<td>Nonglaucomatous optic atrophy</td>
<td>58</td>
<td>12.1</td>
</tr>
<tr>
<td>Congenital optic disc abnormality</td>
<td>3</td>
<td>0.6</td>
</tr>
<tr>
<td>Others retinal diseases</td>
<td>40</td>
<td>8.3</td>
</tr>
<tr>
<td>Normal fundus</td>
<td>22</td>
<td>4.6</td>
</tr>
<tr>
<td>Total</td>
<td>480</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4. The Proportion of Reasons for False-Positive Results by the Deep Learning Algorithm

validation increased, but that of training remained stable, appreciable decreased over the training course with accuracy increases and the loss of prediction of the algorithm and the given manual label. When the loss contributions to the data cleaning, integration, and convolutionaliated High School of South China Normal University, for his validation dataset decreased over time. The loss means the validation set. The blue and red line show that the loss of training and shows that the accuracy over the training course increased over time in the Training curve for the deep leaning algorithm. The green line Figure 6.

Acknowledgments
The authors thank Zehong Zhou, International Department, Affiliated High School of South China Normal University, for his contributions to the data cleaning, integration, and convolutional neural network construction.

References

29. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial

Footnotes and Financial Disclosures

Originally received: November 24, 2017.
Final revision: January 10, 2018.
Accepted: January 18, 2018.
Available online: n n n. Manuscript no. 2017-2689.
1 State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China.
2 Guangzhou Healgoo Interactive Medical Technology Co. Ltd., Guangzhou, China.
3 Centre for Eye Research Australia; Departments of Ophthalmology and Surgery, University of Melbourne, Melbourne, Australia.
4 Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, California.
* These authors contributed equally as first authors.
Financial Disclosure(s):
The author(s) have made the following disclosure(s): W.M.: Patent — Using deep learning models to process color fundus images
M.H.: Patent — Using deep learning models to process color fundus images
Supported in part by the Fundamental Research Funds of the State Key Laboratory in Ophthalmology, National Natural Science Foundation of China (grant no.: 81420108008); the Science and Technology Planning Project of Guangdong Province (grant no.: 2013B20400003); the University of Melbourne at Research Accelerator Program of Australia (M.H.); the CERA Foundation of Australia (M.H.); Victorian State Government of Australia (Operational Infrastructure Support to the Centre for Eye Research Australia); and Research to Prevent Blindness, Inc., New York (to Stanford University Eye Department). The sponsors or funding organizations had no role in the design or conduct of this research.

HUMAN SUBJECTS: Human subjects were included in this study. The institutional review board of Zhongshan Ophthalmic Center approved the study and determined that informed consent was not required because of retrospective nature and fully anonymized usage of images in this study. The study was performed in accordance with the tenets of the Declaration of Helsinki.

Author Contributions:
Conception and design: Li, Meng, M.H.
Analysis and interpretation: Li, Keel, Chang
Data collection: Li, Y.He, Keel, Meng
 Obtained funding: none
 Overall responsibility: Li, Y.He, Keel, Chang, M.H.

Abbreviations and Acronyms:
AMD = age related macular degeneration; AUC = area under receiver operator characteristic curve; DD = disc diameter; GON = glaucomatous optic neuropathy; RNFL = retinal nerve fiber layer; VCDR = vertical cup to disc ratio.

Correspondence:
Mingguang He, MD, PhD, State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, People’s Republic of China. E-mail: mingguang.he@unimelb.edu.au.