THE OFFICIAL INTERNATIONAL
JOURNAL OF THE JAPANESE
OPHTHALMOLOGICAL SOCIETY

CLINICAL INVESTIGATION

Evaluation of deep convolutional neural networks for glaucoma detection

Sang Phan[1] · Shin'ichi Satoh[1] · Yoshioki Yoda[2] · Kenji Kashiwagi[3] · Tetsuro Oshika[4] · The Japan Ocular Imaging Registry Research Group

Abstract

Purpose To investigate the performance of deep convolutional neural networks (DCNNs) for glaucoma discrimination using color fundus images

Study design A retrospective study

Patients and methods To investigate the discriminative ability of 3 DCNNs, we used a total of 3312 images consisting of 369 images from glaucoma-confirmed eyes, 256 images from glaucoma-suspected eyes diagnosed by a glaucoma expert, and 2687 images judged to be nonglaucomatous eyes by a glaucoma expert. We also investigated the effects of image size on the discriminative ability and heatmap analysis to determine which parts of the image contribute to the discrimination. Additionally, we used 465 poor-quality images to investigate the effect of poor image quality on the discriminative ability.

Results Three DCNNs showed areas under the curve (AUCs) of 0.9 or more. The AUC of the DCNN using glaucoma-confirmed eyes against nonglaucomatous eyes was higher than that using glaucoma-suspected eyes against nonglaucomatous eyes by approximately 0.1. The image size did not affect the discriminative ability. Heatmap analysis showed that the optic disc area was the most important area for the discrimination of glaucoma. The image quality affected the discriminative ability, and the inclusion of poor-quality images in the analysis reduced the AUC by 0.1 to 0.2.

Conclusions DCNNs may be a useful tool for detecting glaucoma or glaucoma-suspected eyes by use of fundus color images. Proper preprocessing and collection of qualified images are essential to improving the discriminative ability.

Keywords Artificial intelligence · Deep convolutional neural network · Deep learning · Ocular fundus color image · Glaucoma

Corresponding Author: Kenji Kashiwagi

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10384-019-00659-6) contains supplementary material, which is available to authorized users.

✉ Kenji Kashiwagi
  kenjik@yamanashi.ac.jp

[1] Research Center for Medical Bigdata (RCMB), National Institute of Informatics, Tokyo, Japan

[2] Yamanashi Koseiren Health Care Center, Kofu, Japan

[3] Department of Ophthalmology, Faculty of Medicine, University of Yamanashi, 1110 Shimokato, Chuo, Yamanashi, Japan

[4] Department of Ophthalmology, Faculty of Medicine, University of Tsukuba, Tsukuba, Japan

Introduction

Glaucoma is a leading cause of blindness in people aged older than 60 years and is predicted to affect approximately 80 million people by 2020 [1]. As aging is one of the major risk factors for glaucoma, the number of patients may become higher in the future as societies age, including Japanese society [2]. Glaucoma is a chronic and irreversible eye disease. The majority of glaucoma patients are often asymptomatic until the disease reaches an advanced stage, and the number of ophthalmologists is not sufficient to diagnose patients before the symptoms present. Therefore, an automatic diagnostic system would play an important role in early screening to halt or slow the progression of glaucomatous optic neuropathy (GON).

Intraocular pressure (IOP) is the most important factor in the progression of glaucoma, but a previous epidemiologic

study indicated that the IOP of the majority of Japanese glaucoma patients was within the normal range [2], indicating that IOP has low significance as a screening factor for glaucoma.

Glaucoma is a disease that causes characteristic optic nerve head damage and corresponding visual field defects, and the most effective screening for glaucoma is the detection of characteristic changes in the optic nerve structure. However, precise evaluation of GON can be performed only by trained physicians, and the number of these is insufficient for patient screening. The necessary expertise is inaccessible to a major part of the population, especially in developing countries. Therefore, an automatic diagnosis system for glaucoma is highly necessary.

Inspired by the recent success of applying deep learning for the detection of diabetic retinopathy and glaucoma in retinal fundus photos [3–6], we examined the application of deep learning, especially deep convolutional neural networks (DCNNs), for glaucoma screening among Japanese individuals. To examine their discriminative ability in detecting glaucoma and the determinant factors for its diagnosis, in the present study, we experimented with 3 popular DCNNs, VGG, ResNet, and DenseNet, using 2-dimensional color fundus photos. We also visualized the weights of the ResNet and DenseNet models that were used to make the decision on each image.

Patients and methods

This study was conducted as part of a project to construct a fundamental database of medical information, including ocular images using information and communication technology (ICT) and artificial intelligence (AI), for generalizing future ophthalmologic care undertaken and supported by the Japan Agency of Medical Research and Development (AMED).

The study was approved by the institutional review boards (IRBs) of the Japan Ophthalmologic Society, Yamanashi University, the National Institute of Informatics, and Yamanashi Koseiren Hospital. This study was also performed in accordance with the Declaration of Helsinki. The IRBs waived the need for written informed consent because of the retrospective nature of the study and the use of anonymized fundus photographs. Fundus images were anonymized with the use of identification numbers at the University of Yamanashi or at Yamanashi Koseiren Hospital, where the images were taken, and then sent to the National Institute of Informatics, where the deep convolutional neural network analyses were performed. The fundus photograph was taken at a field angle of 45 degrees using the TC50DX (Topcon) at University of Yamanashi Hospital and the CR2 (Canon) at Koseiren Hospital. Two datasets were prepared to examine

the discriminative performance for glaucoma and the factors influencing the results.

Main dataset for the discrimination of glaucoma

Two sets of fundus photo collections were prepared for this purpose. The first dataset included images from patients with a confirmed diagnosis of glaucoma who visited the Yamanashi University glaucoma outpatient clinic from January 2017 to February 2017 (referred to as YMU-gla). The second dataset included consecutive patients who received medical examinations from January 2017 to February 2017 at the Yamanashi Koseiren Hospital (referred to as KOSEI). For the diagnosis of glaucoma, in the case of the YMU group, the following conditions were satisfied: GON defined by the criteria of the Japan Glaucoma Society Taiji Glaucoma Epidemiological Survey [2] and corresponding visual field loss confirmed by a static visual field test (Humphrey static perimeter program 30-2 or 24-2; Carl Zeiss Meditec) [7]; positive evaluation by glaucoma automatic discrimination optical coherence tomography (OCT) programs including optic nerve head analysis, parapapillary retinal nerve fiber layer thickness measurement, and ganglion cell complex measurement (CIRRUS HD-OCT; Carl Zeiss Meditec); and agreement in diagnosis among more than 2 ophthalmologists. In the case of the KOSEI group, 1 glaucoma specialist (K.K.) judged between glaucoma and nonglaucoma with the same diagnostic criteria for glaucoma (KOSEI-gla) and nonglaucoma (KOSEI-normal) as those used in the YMU group. The exclusion criteria for the YMU and KOSEI groups included fundus images showing any diseases other than glaucoma; those showing any diseases causing a glaucoma-like appearance in the optic disc; and those judged by the specialists to be difficult to diagnose owing to poor quality such as defocused or blurred images, insufficient inclusion of the surroundings, or overexposure of the optic disc. We performed 3 comparisons to analyze the discriminative ability because it was impossible to eliminate cases of suspected glaucoma completely from the KOSEI-gla images that may have influenced the discriminative ability. Comparison 1 compared the YMU-gla images with the KOSEI-normal images, comparison 2 compared the KOSEI-gla images with the KOSEI-normal images, and comparison 3 compared the YMU-gla images combined with the KOSEI-gla images and the KOSEI-normal images.

Supplemental dataset

In clinical settings, ophthalmologists must commonly diagnose images of poor quality, which is a challenge when evaluating fundus images, even among experts. To investigate the effect of poor-quality images on the results, we prepared a dataset

containing images of poor quality from the YMU and KOSEI groups.

Poor-quality images of glaucoma eyes were defined as those that satisfied the following conditions: the glaucoma specialist (K.K.) experienced difficulty diagnosing glaucoma using the fundus image owing to a defocused or blurred image; insufficient inclusion of the surroundings; or overexposure of the optic disc head and some additional information supporting the existence of glaucoma including a history of glaucoma, previous clear fundus images, OCT images, and a visual field defect corresponding to GON. We also used poor-quality images of glaucomatous and nonglaucomatous eyes from the KOSEI groups that met the following conditions: poor-quality images with similar characteristics to those described for the poor-quality glaucoma images, but the glaucoma specialist (K.K.) could conclude whether the eye was glaucomatous or not.

Data preprocessing

We first cropped the background of the input images; the resulting images were then resized to squares of 256-x-256 and 512-x-512 pixels. During training, we applied several data augmentation methods, such as random flipping, random rotation, and random color jittering. The input image size to the network was 224/448 for the 256/512 input images.

Evaluation criteria

We used the area under the receiver operating characteristic curve (AUC) to evaluate the performance of the glaucoma diagnosis. This curve shows the trade-off between sensitivity and specificity, which are defined as:

$$Sensitivity = \frac{TP}{TP + FN}, Specificity = \frac{TN}{TN + FP},$$

where TP, TN, FP, and FN are the number of true-positives, true-negatives, false-positives, and false-negatives, respectively.

Frameworks

We formulated glaucoma detection as a binary classification problem (Fig. 1) and used convolutional neural networks for the learning of the image representation. The whole network was trained end-to-end by minimizing the binary cross-entropy loss. In this work, we used VGG19 [8], ResNet152 [9], and DenseNet201 [10]. All of these networks consist of multiple convolutional layers followed by pooling layers and activation layers. At the end of these networks are 1 or 2 fully connected layers. ResNet and DenseNet further use residual connections or dense connections in their architectures to transmit input as well as gradients between distant layers.

Impact of image resolution on discriminative performance

Because image resolution may influence the results, we compared the results with respect to the input image size. First, if the input image size was larger than the standard size, we inserted an adaptive average pooling layer before the first fully connected (FC) layer. This preprocessing guaranteed a fixed-size input into the FC layer regardless of the input image size. Then, we compared the discriminative ability between the 256-x-256 and 512-x-512 images using VGG19.

Heatmap visualization

The definition of the parts of the image that contribute the most to the decision of the network is important and can be accomplished using the class activation map method proposed by Zhou and colleagues [11]. This method can be

Fig. 1 Performance comparison among 3 DCNNs
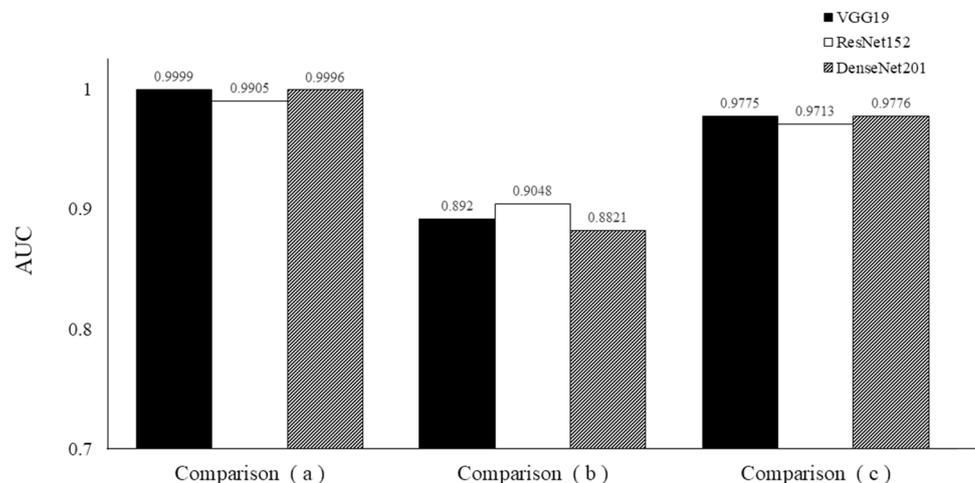
applied directly with the output of the ResNet and DenseNet models because they, unlike the VGG network, use global average pooling (GAP) before the FC layer.

Interference caused by poor-quality images in evaluation accuracy

The influence of poor-quality images contained in a dataset on discriminative ability is important to investigate. We combined the main and supplemental datasets to analyze the discriminative ability using the same 3 models.

Results

Main dataset

Of all the collected images, 18.1% of the images were excluded from the analysis because they did not satisfy the enrollment criteria. The final dataset contained 3312 images in total, of which 369 were glaucoma cases collected at YMU (referred to as YMU-gla); 256, glaucoma cases collected at KOSEI (referred to as KOSEI-gla); and 2687, nonglaucoma images obtained at the KOSEI hospital (referred to as KOSEI-normal). To investigate differences in the results between the datasets from the different institutes, we divided the whole collection into subgroups including training (50%), validating (25%), and testing splits (25%), as indicated in Table 1.

Supplemental data set

We also collected 465 poor-quality images, comprising 363 KOSEI-normal images, 64 YMU-gla images, and 38 KOSEI-gla images, for the supplemental dataset. Some example images are shown in the supplemental figure.

Comparison of DCNNs

Figure 1 shows the performances of the VGG, ResNet, and DenseNet in the 3 comparisons. Both the VGG19 and

Table 1 Conditions of 3 datasets and data splits

Conditions	Training	Validating	Testing
(a) YMU-gla. vs KOSEI normal	1528 (14.5)	764 (12.2)	764 (13.7)
(b) KOSEI-gla. vs KOSEI normal	1471 (9.4)	736 (9.7)	736 (9.5)
(c) YMU-gla. + KOSEI gla. vs KOSEI normal	1656 (22.8)	828 (24.1)	828 (23.2)

Numbers in parentheses indicate the percentage of glaucoma or glaucomatous cases

DenseNet201 showed similarly high AUCs of more than 0.97, and the ResNet152 had an AUC of approximately 0.9. The comparison between the YMU-gla and KOSEI-normal images showed a slightly better AUC than that between the KOSEI-gla and KOSEI-normal images.

Effect of image size

Table 2 shows the comparison of the results with respect to input image size. While image size was an important factor for the detection of diabetic retinopathy in our previous study, we found that it did not affect glaucoma detection, at least in our dataset.

Heatmap visualization

Figure 2 shows 2 output heatmaps, one for predicting a positive image (top) and the other for predicting a negative image (bottom). The heatmaps show that the optic disc area tends to be a greater area of focus in the predictions. This observation is interesting because the optic disc is the primary area examined for glaucoma diagnosis.

However, owing to their low resolution (interpolation from the 7-x-7 grid), we did not consider the heatmaps to be particularly helpful in understanding why the network made the decision.

Therefore, we focused on the optic disc area to obtain higher resolution heatmaps. We trained the classifiers again on the optic disc images. First, we segmented the optic disc area using a previously described method [12]. We were able to detect the optic discs for 3143 of 3312 images (94.8%) and reevaluated the discriminative ability. The models that were trained on the optic disc images performed similarly to the models that were trained on the full image, therefore confirming our observation that the network relies mainly on the optic disc area to make the prediction.

Effects using poor-quality images on discrimination

Thereafter, we used dataset (c) as the training data to train the classifier. Two datasets, the YMU and KOSEI, were prepared to be used as test data to examine the effect of poor-quality images (Table 3). A set of 465 images

Table 2 Performance of VGG model with different sizes of the input images

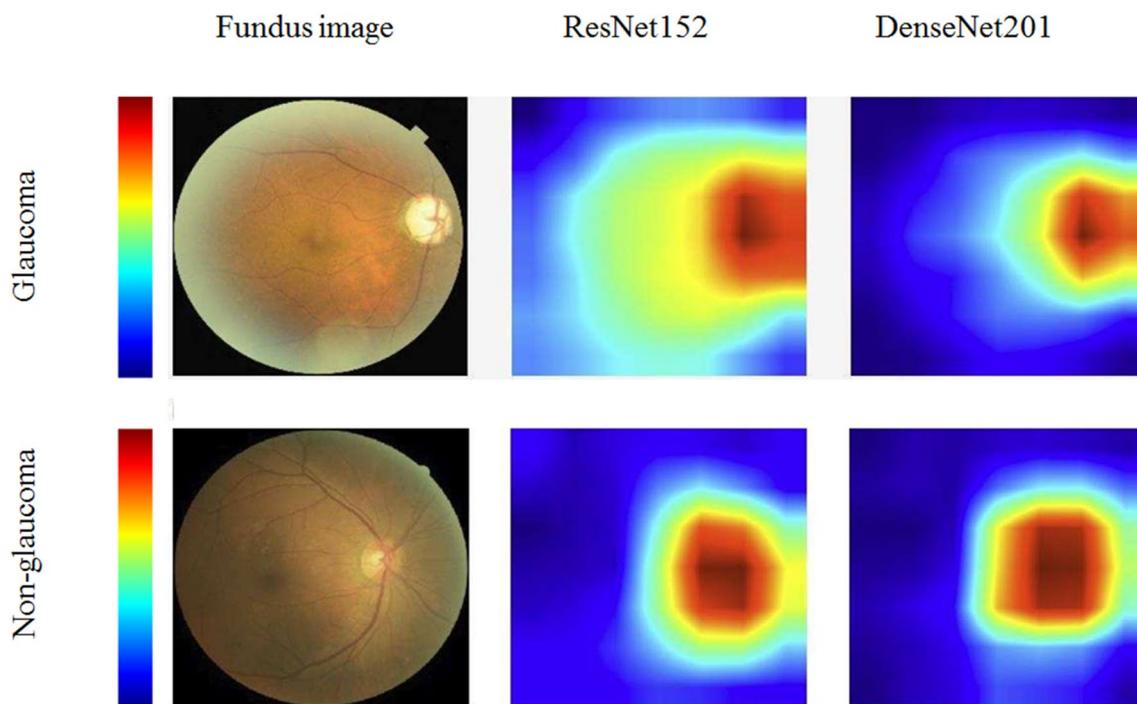| Image size (pixel)	YMU-gla vs KOSEI-normal	KOSEI-gla vs KOSEI-normal	YMU-gla+KOSEI-gla vs KOSEI-normal
256 x 256	0.9999	0.8920	0.9775
512 x 512	0.9998	0.8934	0.9776

Fig. 2 Representative heatmap visualization

Table 3 Conditions of 2 datasets including poor-quality images and data splits

Conditions	Training	Validating	Testing	Testing of poor images
1	1656 (22.8)	828 (24.1)	828 (23.2)	465 (21.9)
2	2484 (23)	828 (24.1)	0	465 (21.9)

Numbers in parentheses indicate the percentage of glaucoma or glaucomatous cases

was included as poor-quality images. We prepared 2 settings for this study. Setting 1 consisted of 1656 images for training, 828 images for validation, and 828 clear images and 465 poor images for testing. On the other hand, setting 2 consisted of 2484 images for training, 828 images for validation, and 465 poor images for testing. A significant drop in performance, as indicated by a significant decrease in the AUC, was observed when testing the same model on the dataset that included the poor-quality images (Fig. 3).

Figure 4 shows the plots for the AUC scores associated with the inclusion of a varied number of training images. We used the VGG19 network and ran each experiment 10 times and calculated the mean along with the standard error. We observed that increasing the number of training images definitely improved the AUC.

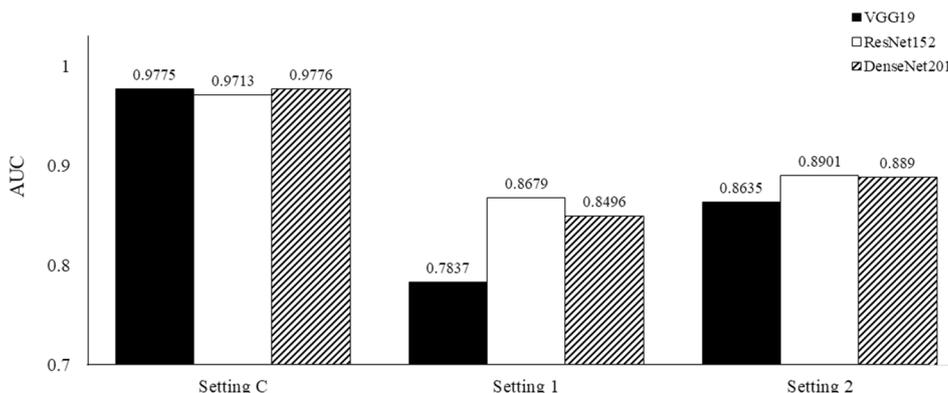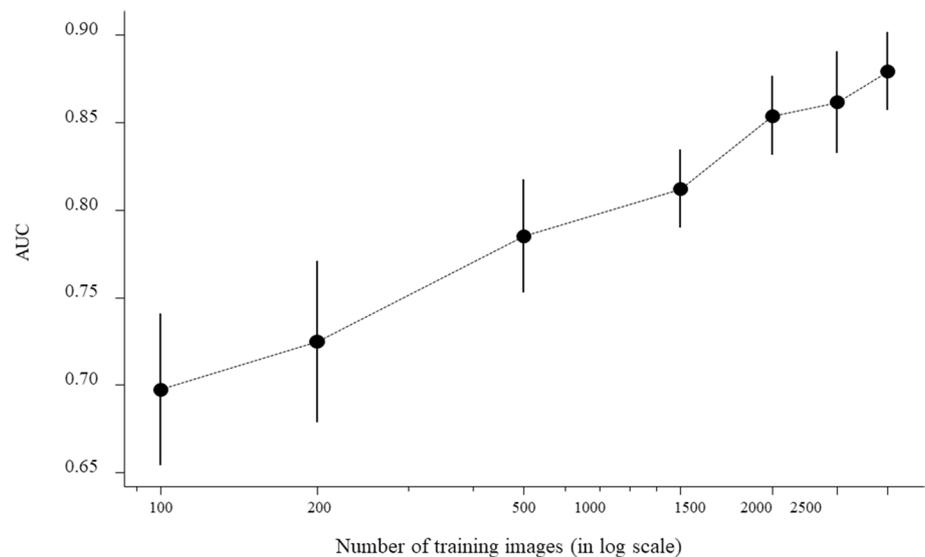**Fig. 3** Influence of poor-quality images on performance of 3 DCNNS

Fig. 4 Effects of the number of training images on discriminative performance for glaucoma

Discussion

The present study investigated the performance of 3 conventional DCNNs for glaucoma discrimination using color fundus images and the factors related to this discriminative ability. We found that the 3 DCNNs examined showed similar but not equivalent discriminative abilities, with AUCs of approximately 0.9. Furthermore, the optic disc area was shown to be the most important region for glaucoma discrimination, but no significant difference was found in the discriminative ability when the resolution was increased by focusing on the optic disc area. Finally, the image quality affected the discriminative ability.

We observed a dataset bias in our results; the AUC of the glaucoma patients from the YMU dataset was much better than that of the glaucoma patients from the KOSEI dataset. Similarly, a previous paper showed that the AUC values for glaucoma discrimination were not the same among multiple datasets [4]. Some possible explanations are available. First, the GON among the glaucoma-confirmed cases may be more advanced than that among the glaucoma-suspected cases. Although we consecutively selected glaucoma-confirmed cases that satisfied the entry and exclusion criteria in this study, differences in the severity of GON between the glaucoma-confirmed cases and the glaucoma-suspected cases cannot be ruled out. Second, other diseases that appear similar to GON may have contaminated our dataset because we did not perform a visual field test or OCT examination to confirm the diagnosis for these eyes. Third, the diagnosis of the glaucoma-suspected cases was judged by 1 glaucoma specialist, which may have resulted in diagnostic bias. Fourth, the quality of the photos, including the resolution and instrument, were not the same between YMU

and KOSEI, which may have influenced the discriminative ability. To minimize this possibility, we prepared a dataset that included images from the 2 datasets. Indeed, the discriminative abilities were very similar. Because the number of patients in this study was not large, further study with a much larger dataset is necessary to address these possibilities.

The present study showed that the quality of the image affects the discriminative ability. Therefore, collecting high-quality images is important for better discrimination. However, in actual clinical practice, some fundus images are of poor quality, exhibiting poor characteristics such as defocused, blurred, and insufficient inclusion of the surroundings and overexposure of the optic disc. Indeed, Li and colleagues reported that GON with coexisting pathologic or high myopia is the most common cause of false-negative results [6]. In their study, approximately 17% of all enrolled fundus images were excluded from the analysis owing to poor quality or poor image location, which is a similar percentage to that observed in the present study. Because the heatmap analysis revealed that the optic head is the most important region in glaucoma discrimination, we may obtain a similar discriminative ability by focusing on the optic nerve head region to that obtained by targeting the whole fundus region. Data cleaning or other image preprocessing methods for better discriminative ability should also be considered.

Generally, availability of more images results in a higher discriminative ability. The 3 DCNNs used in the present study showed comparatively good AUC results with those observed in previous studies [4, 6, 13], although the number of images enrolled was smaller than those in the previous studies. Glaucoma may be a suitable disease for AI analysis, and the use of AI may be supportive for glaucoma discrimination during medical examination.

The present study has some limitations. The number of images studied was small, and the glaucoma discriminative ability exhibited an AUC of approximately 0.9. For clinical use, a higher AUC is necessary. In the present study, the discriminative ability was shown to increase with an increase in the number of included images. Therefore, the efficacy of the DCNNs in discriminating glaucoma should be investigated using a greater number of images. The present study examined confirmed cases of glaucoma and suspected cases judged by glaucoma specialists, but some cases exhibited highly deformed optic discs that are even difficult to diagnose by glaucoma specialists. Future studies must discriminate glaucoma more precisely. For example, some previous studies have used OCT or other instruments for this purpose [14]. The development of an AI system with a high discriminative ability must use various images. Although there was an imbalance in the 3 datasets between glaucomatous and healthy individuals, the evaluation by AUC had the effect of avoiding the influence of imbalance. It is well known that glaucoma severity may influence discriminative ability. Since we did not investigate the effect of glaucoma severity in this study, it is necessary to perform this in further studies.

Because glaucoma is an irreversible and progressive disease, the prognosis must be confirmed to provide proper treatment choices. Research on prognosis prediction using visual field test results is underway [15, 16], but the development of a system that can make a prognosis judgment is necessary. Furthermore, various eye diseases other than glaucoma or pathologic conditions suggesting the existence of systemic diseases can be detected in the fundus, and the development of systems to distinguish these diseases should be pursued.

We have empirically shown that the optic disc is the most crucial part of the image for glaucoma diagnosis. However, there are several structural features that indicate the presence of glaucoma, including nerve fiber layer defects and optic disc hemorrhage, which may improve the discriminative ability for glaucoma. Unfortunately, excellent annotations of glaucoma-specific evidence in AI analysis have not been fully undertaken. Our future work employing AI systems for early detection and proper treatment of glaucoma should further our understanding of this field.

References

1. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. Br J Ophthalmol. 2006;90:262–7.
2. Iwase A, Suzuki Y, Araie M, Yamamoto T, Abe H, Shirato S, et al. The prevalence of primary open-angle glaucoma in Japanese: the Tajimi Study. Ophthalmology. 2004;111:1641–8.
3. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
4. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, 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;318:2211–23.
5. Cerentini A, Welfer D, Cordeiro d'Ornellas M, Pereira Haygert CJ, Dotto GN. Automatic identification of glaucoma using deep learning methods. Stud Health Technol Inform. 2017;245:318–21.
6. 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;125:1199–206.
7. Anderson DR, Patella VM. Automated static perimetry. 2nd ed. St. Louis, MO: Mosby; 1999.
8. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Paper presented at: International Conference on Learning Representations; May 7–9, 2015; San Diego, CA.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings from the IEEE Conference on Computer Vision and Pattern Recognition; June 27–30, 2016; Las Vegas, NV.
10. Huang G, Liu Z, van der Maaten L, Weinberger K. Densely connected convolutional networks. In: Proceedings from the IEEE Conference on Computer Vision and Pattern Recognition; July 21–26, 2017; Honolulu, HI.
11. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings from the IEEE Conference on Computer Vision and Pattern Recognition; June 27–30, 2016; Las Vegas, NV.
12. Sevastopolsky A. Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. Pattern Recogn Image Anal. 2017;27:618–24.

13. Keel S, Lee PY, Scheetz J, Li Z, Kotowicz MA, MacIsaac RJ, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. Sci Rep. 2018;8:4330.

14. Devalla SK, Chin KS, Mari JM, Tun TA, Strouthidis NG, Aung T, et al. A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head. Invest Ophthalmol Vis Sci. 2018;59:63–74.

15. Murata H, Zangwill LM, Fujino Y, Matsuura M, Miki A, Hirasawa K, et al. Validating variational Bayes linear regression method with multi-central datasets. Invest Ophthalmol Vis Sci. 2018;59:1897–904.

16. Yousefi S, Kiwaki T, Zheng Y, Sugiura H, Asaoka R, Murata H, et al. Detection of longitudinal visual field progression in glaucoma using machine learning. Am J Ophthalmol. 2018;193:71–9.