

## RESEARCH ARTICLE

# Deep learning assisted detection of glaucomatous optic neuropathy and potential designs for a generalizable model

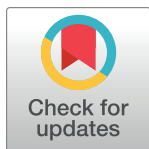
Yu-Chieh Ko<sup>1,2,3</sup>, Shih-Yu Wey<sup>4</sup>, Wei-Ta Chen<sup>5</sup>, Yu-Fan Chang<sup>1,3</sup>, Mei-Ju Chen<sup>1,2</sup>, Shih-Hwa Chiou<sup>3,6</sup>, Catherine Jui-Ling Liu<sup>1,2,†\*</sup>, Chen-Yi Lee<sup>4,†\*</sup>

**1** Department of Ophthalmology, Taipei Veterans General Hospital, Taipei, Taiwan, **2** Faculty of Medicine, National Yang-Ming University School of Medicine, Taipei, Taiwan, **3** Institute of Clinical Medicine, National Yang-Ming University School of Medicine, Taipei, Taiwan, **4** Department of Electronics Engineering, National Chiao-Tung University, HsinChu, Taiwan, **5** Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan, **6** Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

☉ These authors contributed equally to this work.

‡ These authors also contributed equally to this work.

\* [jliliu@vghtpe.gov.tw](mailto:jliliu@vghtpe.gov.tw) (CJL); [cylee@si2lab.org](mailto:cylee@si2lab.org) (CYL)



## OPEN ACCESS

**Citation:** Ko Y-C, Wey S-Y, Chen W-T, Chang Y-F, Chen M-J, Chiou S-H, et al. (2020) Deep learning assisted detection of glaucomatous optic neuropathy and potential designs for a generalizable model. PLoS ONE 15(5): e0233079. <https://doi.org/10.1371/journal.pone.0233079>

**Editor:** Yuchen Qiu, University of Oklahoma, UNITED STATES

**Received:** October 8, 2019

**Accepted:** April 28, 2020

**Published:** May 14, 2020

**Copyright:** © 2020 Ko et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Data cannot be shared publicly because of the regulation of the Ministry of Science and Technology of Taiwan. The images have been uploaded in the National Center for High-performance Computing, the LIONS Data Framework and is available for researchers who meet the criteria for access to confidential data. The detailed information can be found at <http://omim.nchc.org.tw/LIONSdata/search.php>. The readers can contact Ya-yuan Cheng at [yycheng907@gmail.com](mailto:yycheng907@gmail.com) for further information.

## Abstract

### Purpose

To evaluate ways to improve the generalizability of a deep learning algorithm for identifying glaucomatous optic neuropathy (GON) using a limited number of fundus photographs, as well as the key features being used for classification.

### Methods

A total of 944 fundus images from Taipei Veterans General Hospital (TVGH) were retrospectively collected. Clinical and demographic characteristics, including structural and functional measurements of the images with GON, were recorded. Transfer learning based on VGGNet was used to construct a convolutional neural network (CNN) to identify GON. To avoid missing cases with advanced GON, an ensemble model was adopted in which a support vector machine classifier would make final classification based on cup-to-disc ratio if the CNN classifier had low-confidence score. The CNN classifier was first established using TVGH dataset, and then fine-tuned by combining the training images of TVGH and Drishti-GS datasets. Class activation map (CAM) was used to identify key features used for CNN classification. Performance of each classifier was determined through area under receiver operating characteristic curve (AUC) and compared with the ensemble model by diagnostic accuracy.

### Results

In 187 TVGH test images, the accuracy, sensitivity, and specificity of the CNN classifier were 95.0%, 95.7%, and 94.2%, respectively, and the AUC was 0.992 compared to the 92.8% accuracy rate of the ensemble model. For the Drishti-GS test images, the accuracy of the CNN, the fine-tuned CNN and ensemble model was 33.3%, 80.3%, and 80.3%,

**Funding:** CJL and YCK received funds from the Ministry of Science and Technology, TAIWAN (MOST 106-2634-F-075-001 and MOST 107-2634-F-075-001 to CJL; MOST 1062314-B-075-037-; MOST 105-2314-B-075-047 to YCK ); YCK received funds from Taipei Veterans General Hospital (V107C-114). The funding source had no role in the design of this study and will not have any role during its execution, analyses, interpretation of the data, or decision to submit results.

**Competing interests:** The authors have declared that no competing interests exist.

respectively. The CNN classifier did not misclassify images with moderate to severe diseases. Class-discriminative regions revealed by CAM co-localized with known characteristics of GON.

## Conclusions

The ensemble model or a fine-tuned CNN classifier may be potential designs to build a generalizable deep learning model for glaucoma detection when large image databases are not available.

## Introduction

Glaucoma is one of the leading causes of irreversible blindness [1]. Early diagnosis and intervention are key to prevent glaucoma blindness [2,3]; however, it is difficult to achieve early detection because of the asymptomatic nature of most types of glaucoma. Even in developed countries, the diagnostic rate of glaucoma is below 50% [4,5]. Population-based screening can identify glaucoma at an asymptomatic stage but is not feasible currently due to low cost-effectiveness [6]. Nevertheless, the combination of automated glaucoma detection with teleglaucoma may enable large-scale glaucoma screening targeted at high-risk populations [7].

Optic disc photography remains the fundamental approach in glaucoma screening [8,9]. Disc photography reviewed by experienced specialists can achieve similar diagnostic performance as that of advanced imaging technology such as optical coherence tomography (OCT) [10], and combining other modalities with fundus images did not necessarily increase the diagnostic power in teleglaucoma [7]. However, evaluation of the optic disc is subjective, and likely to encounter interobserver discordance even among glaucoma specialists [8,11]. Qualified interpretation of fundus images at limited cost through automated glaucoma detection using artificial intelligence may make fundus photography-based glaucoma screening feasible.

In the pre-deep learning era, automated glaucoma detection using fundus images was based primarily on the cup-to-disc ratio (CDR), and in some cases, defects of the retinal nerve fiber layer (RNFL) or the presence of parapapillary atrophy [12]; however, the CDR-based approach may miss important diagnostic information such as disc hemorrhage or have inadequate accuracy if the disc size is not adjusted [8,13]. The development of deep learning techniques in recent years, especially the use of convolutional neural network (CNN) and its variants for computer vision, has allowed improved medical image analysis through expanded capability to extract either low-level, coarse visual features, or high-level refined features [14]. CNN-based algorithm achieved a high diagnostic rate in differentiating the glaucomatous fundus from the healthy fundus [15–17], but its generalizability remains unclear because most studies verified the performance of the algorithm only in the images from the same dataset [15,16]. Furthermore, with the complicated diagnostic process of glaucoma involving both structural and functional assessment [18], it is difficult to build large fundus image datasets for the analysis of glaucoma. Nevertheless, the number of training images will affect the diagnostic performance and generalizability of a deep learning model [19,20]. Therefore, it is intriguing to know whether it is possible to build a CNN classifier for glaucoma detection using limited fundus images and to improve its applicability across different datasets by adding universal diagnostic characteristics, such as enlarged CDR to the model.

This study aimed to evaluate the performance of a CNN algorithm constructed with limited training images to discern glaucomatous change in images from different resources, benefits

of including CDR estimation as an ensemble model, and the key features extracted for the classification in the CNN algorithm.

## Materials and methods

### Database

This study included fundus photographs in JPEG format of 465 non-glaucomatous eyes and 479 eyes with primary open angle glaucoma (POAG), which were collected retrospectively from the image database of the Department of Ophthalmology, Taipei Veterans General Hospital (TVGH). The fundus images were obtained from subjects with POAG at regular follow-up, and subjects who visited our department for ocular screening or disorders not affecting the optic disc and retina. This study followed the tenets of the Declaration of Helsinki, and was approved by the Institutional Review Board of TVGH. The written informed consent was waived because of the retrospective nature of the study and the analysis was performed using anonymized clinical data and images.

Eyes with POAG were diagnosed in the presence of glaucomatous disc changes with corresponding RNFL loss and reproducible visual field (VF) defects, and a normal open angle on gonioscopy. In all subjects with POAG, VF analysis using the 24–2 Swedish interactive threshold algorithm standard of the Humphrey Field Analyzer 750i (version 4.2, Zeiss-Humphrey Instruments, Dublin, CA, USA) and RNFL scanning using Cirrus HD-OCT (Model 4000; Carl Zeiss Meditec, Inc., Dublin, CA, USA) were performed. The normal eyes had an intraocular pressure (IOP) of < 21 mmHg, and a disc with normal appearance without RNFL defect. The fundus images were acquired using three fundus cameras, Canon CR-DGi NM fundus camera, Canon CX-1 hybrid mydriatic/non-mydriatic digital retinal camera, and Canon CR-2 PLUS AF non-mydriatic retinal camera (Canon USA, Inc., Lake Success, NY, USA) at 45° centered at the macula. All images included had a clear view of the disc. The images were reviewed by two glaucoma specialists (YK and CL) to confirm the diagnosis and define the vertical CDR. Standard anonymization was performed for all images before further analysis.

The fundus images were 3888 x 2552 pixels in size, with resolution of 120 dpi and bit depth of 24. The images were divided into a training set and test set for model training and evaluation based on stratified sampling with 763 images and 181 images at the training and test sets, respectively. The clinical information of the images is listed in Table 1. The CNN model established based on the training images from TVGH dataset was specified as TVGH-CNN model. In addition, images from the Drishti-GS Dataset [21], (<http://cvit.iit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php>), an open-access datasets containing training and test datasets of 50 and 51 images, respectively were used to evaluate the diagnostic ability of our model.

**Table 1. Clinical and demographic characteristics of the image dataset from Taipei Veterans General Hospital.**

	POAG	Control
Number of images	479	465
Age (y)	58.78±13.8	61.20±19.34
Gender (Female, %)	168 (35.07%)	226 (48.60%)
Cup-to-disc ratio	0.80±0.12	0.42±0.20
Mean Deviation (dB)	-6.89±6.01	
Average RNFL thickness (μm)	72.73±11.34	

Values are presented as mean ± SD or percentage (%). POAG, primary open angle glaucoma; RNFL, retinal nerve fiber layer.

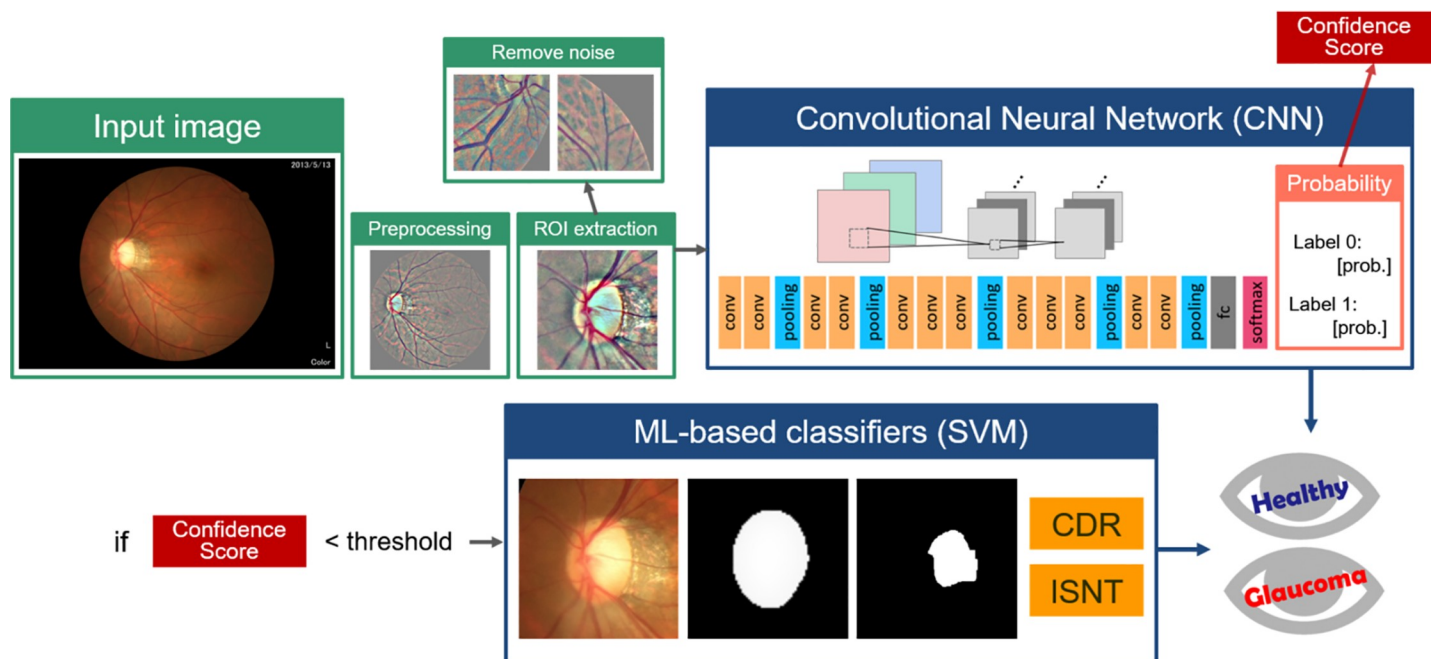
<https://doi.org/10.1371/journal.pone.0233079.t001>

## Proposed detection system

The workflow of our detection system is illustrated in Fig 1. The fundus images were preprocessed and region of interest (ROI) of area of  $256 \times 256$  pixels centered at the disc were extracted. Images with incorrectly extracted ROI were deleted automatically. Subsequently, a CNN model was used for classification. In the last layer of this CNN model, each image was classified as normal or glaucoma with a probability value, which was treated as the confidence score of the CNN classifier. If the confidence score was lower than a self-defined threshold, then the image was subjected to the support vector machine (SVM) classifier for final classification based on customized features, including enlarged vertical CDR, and thinning of the superior and inferior neuroretinal rim. The ensemble model coupling a SVM classifier based on these features was used to avoid missing subjects with moderate to advanced diseases, who will inevitable have enlarged CDR.

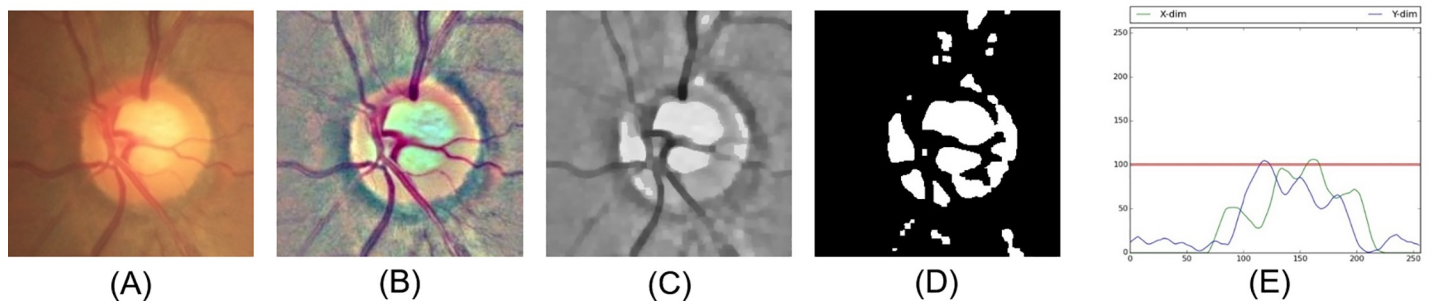
## Image preprocessing and ROI extraction

Fundus images were first preprocessed to compensate for the levels of brightness and clarity of each image. Unsharp masking by Gaussian blur method was used to enhance the optic disc structure and blood vessels. ROI detection centered at the disc was implemented to remove redundant features using the concept of guided backpropagation [22]. Following ROI extraction, images with the optic disc located at the central area of the image were considered as qualified images. Invalid cases with incorrect ROI were automatically excluded through distribution of the histogram on the red and green channel. Images with the disc centered at the ROI would show a peak at the center and were included in subsequent analysis (Fig 2). Finally, data augmentation by adding deformations and noise to the dataset was performed by rotation, shifting, sheering, and zooming using the Keras library (<https://keras.io/>).



**Fig 1. Overall Framework.** Original color images were preprocessed to contrast the vessels and then submitted for ROI extraction. Image with incorrect ROI selection, referred as noise, were excluded from the CNN model. Images with low confidence score in CNN classification were forwarded to an ML-based classifier based on the key characteristics of CDR value and thinning of the superior and inferior neuroretinal rim according to the ISNT rule. CDR, cup-to-disc ratio; CNN, convolutional neural network; ML, machine-learning; ROI, region of interest; SVM, support vector machine.

<https://doi.org/10.1371/journal.pone.0233079.g001>

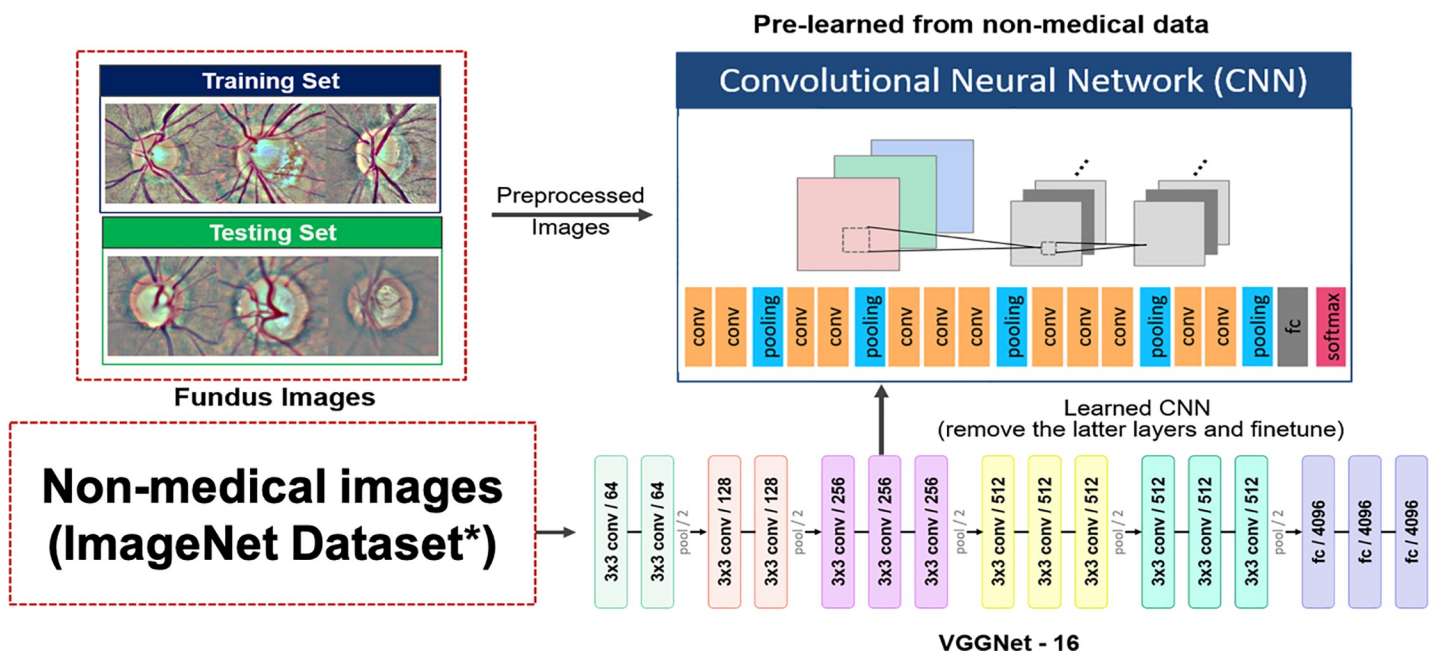


**Fig 2. Process to determine noisy cases with incorrect ROI selection.** (A) Color fundus image; (B) fundus image after preprocessing; (C) further preprocessing on the red and green channel; (D) binary display of the image; (E) histogram of the projection on x-axis and y-axis, which indicates correct ROI selection with both peaks of the axes located at the center. ROI, region of interest.

<https://doi.org/10.1371/journal.pone.0233079.g002>

### Detection based on CNN

In our CNN model, preprocessed and qualified images were submitted for processing through a series of convolutional, nonlinear functions (ReLU), pooling, fully connected (FC) layers, and finally softmax loss layer, which incorporated the concept of cross entropy and softmax layer to provide a probability value for the final classification of each image. Moreover, Stochastic Gradient Descent was performed for optimization to minimize the loss of function. Instead of training a model from scratch, transfer learning was performed using VGGNet (invented by Visual Geometry Group from University of Oxford) [23], a pre-trained CNN model through the following steps: removal of top layers such as the FC layers; retraining the CNN model initialized from the trained weights of VGGNet with a series of convolutional layers of 3 x 3 filter size (Fig 3) [24,25]. Regularization methods, such as data augmentation and dropout layer, were used to improve the condition of convergence and prevent overfitting.



**Fig 3. Transfer learning and VGGNet architecture.** Under transfer learning strategy, the CNN model was initialized with the trained weight of VGGNet with a 3 x 3-sized filter. The final model contained convolutional layers, pooling layers, and the fc layer. VGGNet, invented by Visual Geometry Group from University of Oxford; CNN, convolutional neural network; conv, convolutional; fc, fully connected. \*Russakovsky O, et al. ImageNet large scale visual recognition challenge. *IJCV*. 115, 211–252 (2015).

<https://doi.org/10.1371/journal.pone.0233079.g003>



### Cross-validation of the CNN model

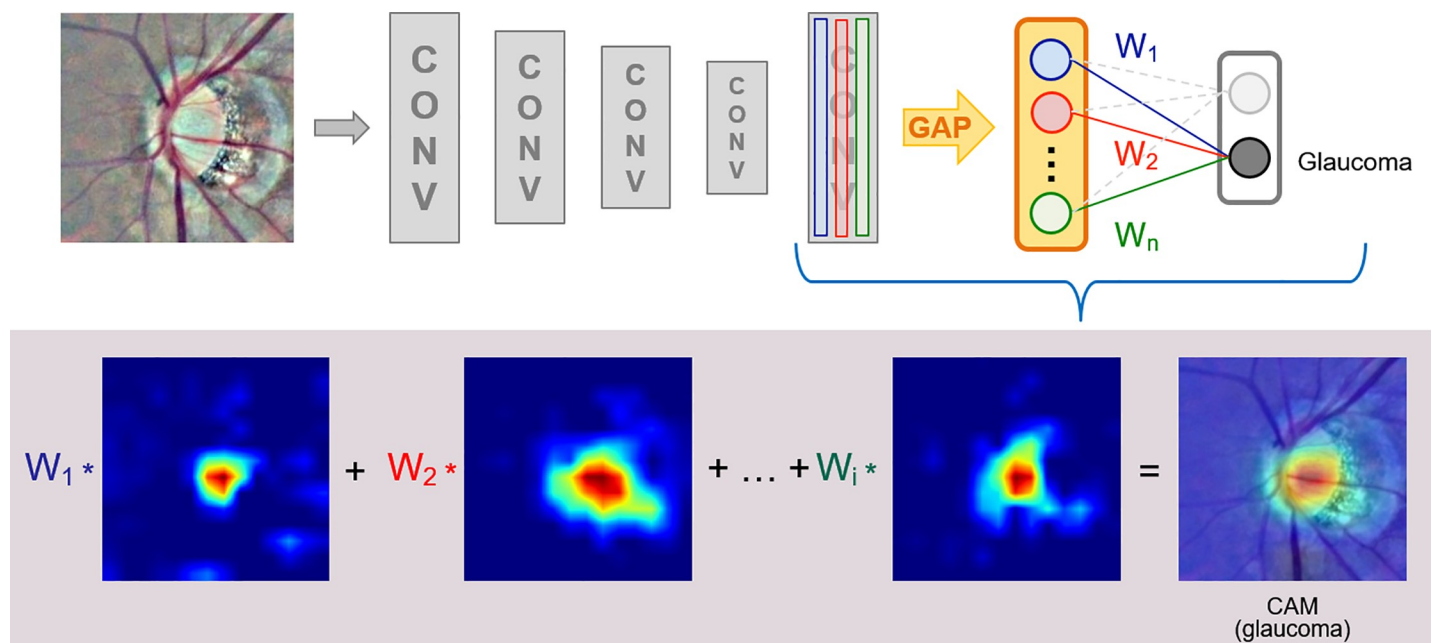
We performed 5-fold cross validation to evaluate the performance and to avoid overfitting of the CNN model. Within each fold, the combined images were partitioned into independent training and validation sets using an 80–20 percentage split. The performance of each fold was evaluated not only on the validation sets, but also the test images. The parameters of the fold with best performance were selected as the final model, unless there were significant variation on the performance of each fold.

### Class activation map (CAM) for visualization

CAM was applied to identify features recognized by the proposed CNN model to differentiate between the glaucomatous and normal discs [26]. Through global average pooling (GAP) performed on the convolutional feature maps and an FC layer that produced the desired output, the important regions of each image were identified by projecting the weights of the output layer back on to the convolutional feature maps (Fig 4) [27].

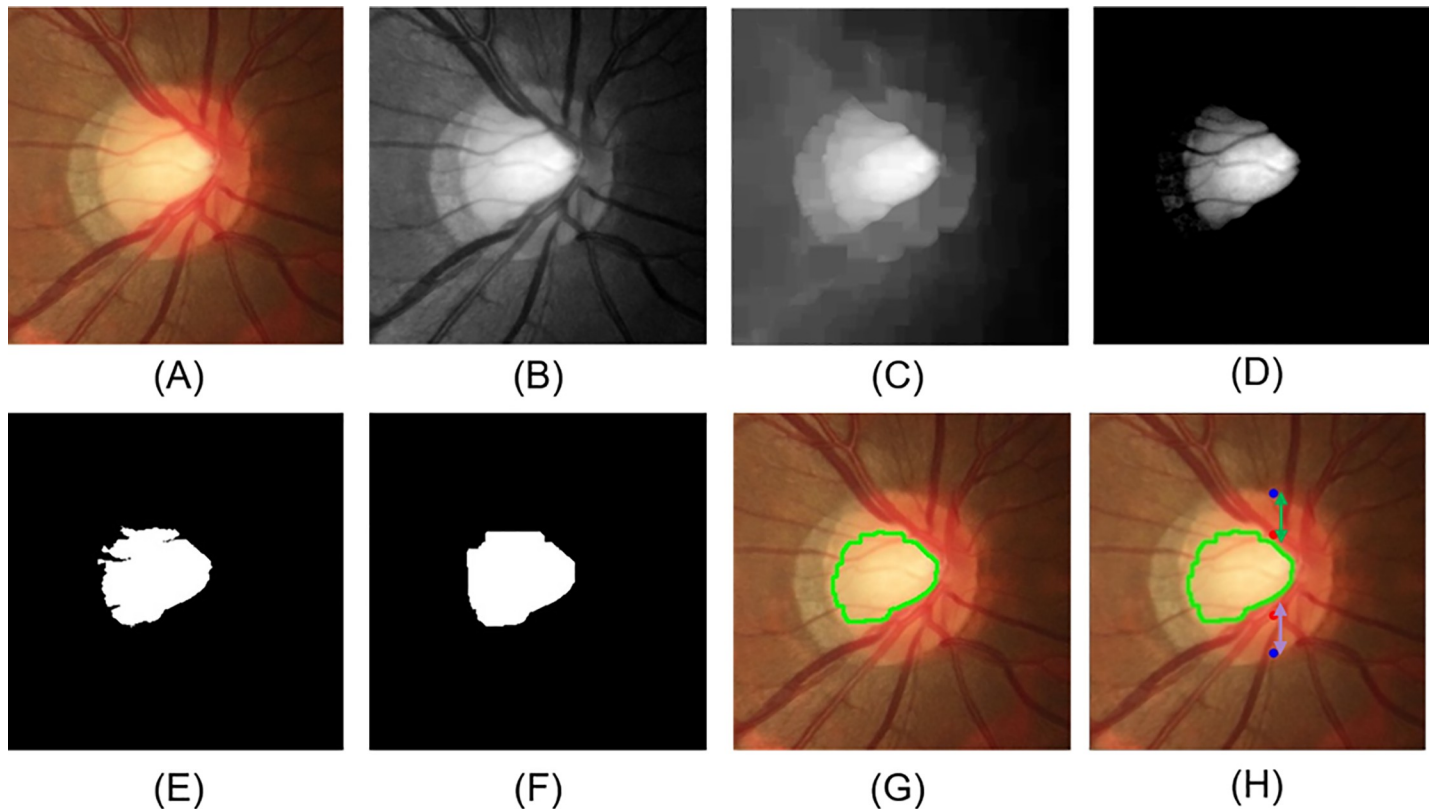
### SVM classifier based on enlarged CDR

The characteristic changes of the glaucomatous disc, including an increased value of CDR, and thinning of the superior or inferior neuroretinal rim were selected as the features for SVM classifier. Preprocessed images were first processed with the green and red channel. The disc margin was determined by the red channel, and the cup area by green channel. The vertical disc diameter was decided according to the superior and inferior junctional points of the disc boundary along the vertical meridian. K-means clustering was used to determine the cup area after removing the redundant information by subtracting the mean and standard deviation



**Fig 4. Class activation map attached to the final layer of the proposed CNN model.** CAM revealed the important areas for the final classification and was generated using GAP with the multiplication of each channel by the given weight from the fully connected (fc) layer ( $W_1, W_2, \dots, W_n$ ). CAM, class activation map; CNN, convolutional neural network; CONV, convolutional; GAP, global average pooling.

<https://doi.org/10.1371/journal.pone.0233079.g004>



**Fig 5. Process to determine the disc margin, cup area, vertical cup-to-disc ratio, and superior and inferior neuroretinal rim-to-disc ratio.** (A) Color fundus image after ROI extraction; (B) green and red channel processing; (C) removal of the vessels; (D) processing by subtracting the mean and standard deviation; (E) K-means clustering; (F) morphological operations to remove noise; (G) the cup boundary (green line); (H) vertical disc diameter (between blue points), cup diameter (between red points), superior rim (green arrow), and inferior rim (purple arrow). ROI, region of interest.

<https://doi.org/10.1371/journal.pone.0233079.g005>

from the green channels, and the K value was set to 2 as dividing the preprocessing image into two regions (Fig 5) [28]. The CDR was calculated as the vertical cup length divided by the vertical disc diameter; the rim-to-disc ratio was calculated as the remaining part at each side including the superior and inferior neuroretinal rim width divided by the vertical disc diameter. The three features were input into the SVM classifier with RBF kernel for classification. Prediction was implemented by using a grid search and the SVM model provided by scikit-learn [29].

### Statistical analysis

The performance of our model was evaluated by the indices of sensitivity, specificity, and accuracy defined as the sum of correct classification of both the healthy and glaucomatous images divided by the total number of test images. The receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) were plotted using the python scikit-learn library according to the definition, wherein, the true positive rate is the Y-axis, and the false positive rate is the X-axis. Nonparametric tests were used to compare the clinical characteristics between images with correct and incorrect prediction because most measurements of the test images were not normally distributed. Mann-Whitney U test was used for continuous variables, and chi-square test for categorical variables. The analyses were conducted using SPSS (version 18.0.0, SPSS Inc., Chicago, IL, USA).

## Result

### Diagnostic performance of the CNN models

In a total of 187 test images from the Taipei Veterans General Hospital (TVGH) dataset, the diagnostic accuracy of the TVGH-CNN model was 95.0%, sensitivity was 95.7%, and specificity was 94.2% in differentiating the glaucomatous eyes from the normal eyes; the AUC of the TVGH-CNN model for identifying glaucomatous eyes was 0.992 (Table 2 and Fig 6). However, the accuracy of the TVGH-CNN model used to classify the 51 test images of the Drishti-GS dataset was comparatively lower at 33.3%, which indicated poor generalizability of the TVGH-CNN model. Two approaches were used to overcome this problem: the first was to fine-tune the TVGH-CNN model using the training images of the Drishti-GS dataset; the second was to develop an ensemble model which used CDR for classification when the confidence score of the CNN classifier was low.

Fifty training images of the Drishti-GS dataset were added to the TVGH training dataset to establish a fine-tuned CNN model (T&D-CNN model), which improved the classification accuracy of the CNN model applied to the Drishti-GS test images. The diagnostic accuracy of the T&D-CNN model for Drishti-GS test images was improved to 80.3%, with a sensitivity of 73.6%, specificity of 100%, and an AUC of 0.937 (Table 2 and Fig 7). To determine optimum incorporation of the limited training images of Drishti-GS dataset into the TVGH dataset, we analyzed the effects of preprocessing, augmentation, and ROI extraction of the training images on the performance of the T&D-CNN model. The result indicated that ROI selection increased the performance of the model. For dataset with limited images such as Drishti-GS, preprocessing and augmentation may increase the model performance (Table 3).

### Diagnostic performance of the ensemble model

An ensemble model for classification was also established with the aim to avoid missing subjects with moderate to advanced disease, which used a SVM classifier based on CDR for final classification when the probability value of the CNN classifier was lower than the threshold of 0.85. The probability value refers to the binary classification confidence of the CNN classifier for the predicted classification. The threshold value was selected by calculating the sensitivity and specificity of the ensemble model with different threshold values starting from 0.5 with an increment of 0.05 till 1.0. The ensemble model had highest sensitivity and specificity on test images including TVGH and Drishti-GS datasets at a threshold value of 0.85. The SVM classifier, which was established based on the TVGH dataset, had a diagnostic accuracy of 76.7% and 72.5% for the TVGH and Drishti-GS test images, respectively. The ensemble model had

**Table 2. Diagnostic capability of the CNN model, SVM model, ensemble model and fine-tuned CNN model in identifying glaucomatous disc.**

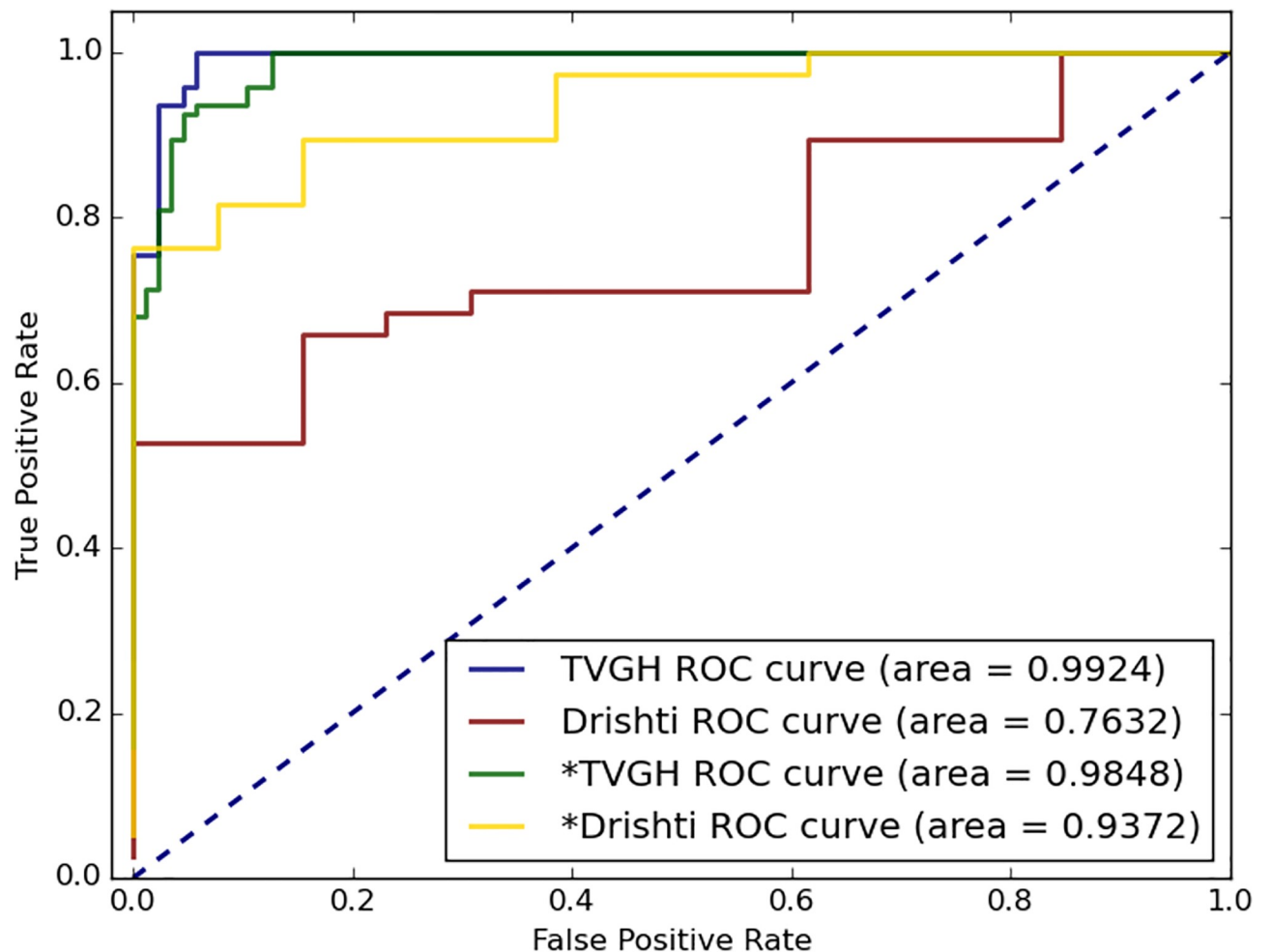
Models	Dataset of test images							
	TVGH				Drishti-GS			
	Accuracy	Specificity	Sensitivity	AUC (95% CI)	Accuracy	Specificity	Sensitivity	AUC (95% CI)
TVGH-CNN model <sup>a</sup>	95.0%	94.2%	95.7%	0.992 (0.984–1.0)	33.3%	100%	10.0%	0.763 (0.632–0.894)
SVM model <sup>a</sup>	76.7%	78.1%	75.5%	0.806 (0.740–0.872)	72.5%	76.9%	71.0%	0.783 (0.617–0.950)
Ensemble model	92.8%	95.4%	90.4%		80.3%	92.3%	76.3%	
Fine-tuned T&D-CNN model <sup>b</sup>	93.9%	95.4%	92.6%	0.984 (0.972–0.998)	80.3%	100%	73.6%	0.937 (0.874–1.0)

<sup>a</sup>The model was constructed based on the TVGH training dataset.

<sup>b</sup>The model was constructed based on the mixed training images of the TVGH and Drishti-GS datasets (T&D-CNN model). CNN, convolutional neural network; SVM, support vector machine; TVGH, Taipei Veterans General Hospital; AUC, area under the curve; CI, confidence interval.

<https://doi.org/10.1371/journal.pone.0233079.t002>





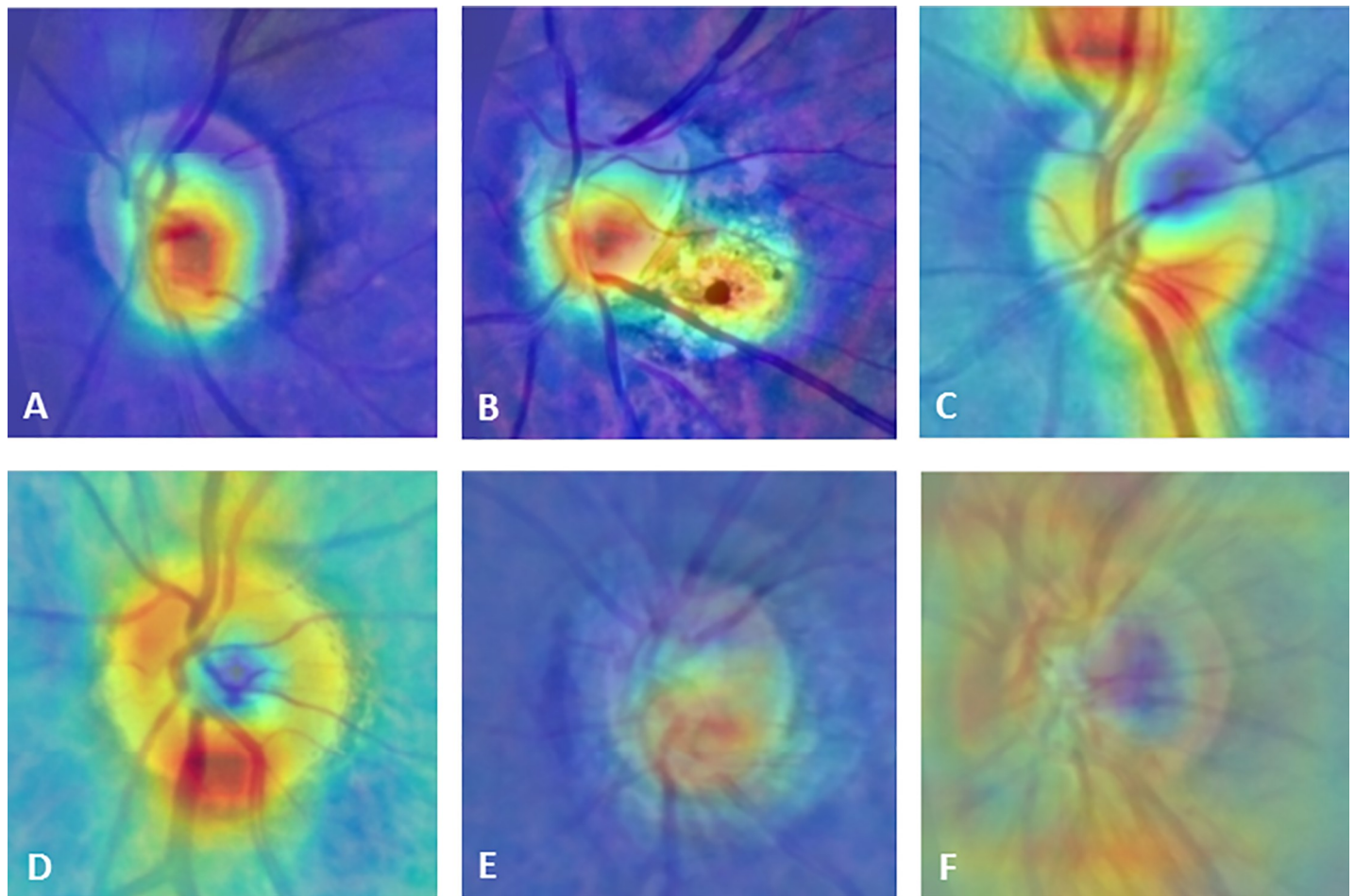
**Fig 6. ROC curve of the CNN models to differentiate between the normal disc and glaucomatous disc.** The TVGH ROC curve indicated the result through the TVGH-CNN model established based on the TVGH training images to predict the TVGH test images. The Drishti ROC curve indicated the result through the CNN model established using the Drishti-GS training images to predict the Drishti-GS test images. \*TVGH and \*Drishti used the combined TVGH and Drishti training dataset to predict the TVGH and Drishti test images, respectively. ROC, receiver operating characteristic; CNN: convolutional neural network.

<https://doi.org/10.1371/journal.pone.0233079.g006>

better diagnostic accuracy for the Drishti-GS test images than that of the TVGH-CNN model, but similar diagnostic accuracy as that of the fine-tuned T&D-CNN model; moreover, there was no increase in the overall diagnostic accuracy for the TVGH test images than that of the TVGH-CNN model (Table 2).

### Factors affecting the performance of the CNN classifier

To understand the potential factors leading to misclassification through our model, we compared the clinical characteristics for the images with correct and incorrect prediction (Table 4). For glaucoma, the images being correctly classified as glaucoma had larger CDR and worse visual field defect than those being incorrectly classified. All images with GON that were incorrectly classified as normal had a mean deviation of better than -6 dB. The images of healthy eyes that were incorrectly classified as glaucoma had larger CDR values than those that were correctly classified.



**Fig 7. Representative CAMs and superimposed images showing different patterns between the glaucomatous and healthy eyes.** The hot spots (red color) were localized at the cup and the region of parapapillary atrophy in the glaucomatous eyes (A, B), and at the neuroretinal rim or along the nerve fibers bundles in the healthy eyes (C, D). The superimposed images with CAM centered at the disc, showed the different patterns of CAM between the glaucomatous (E) and healthy eyes (F). CAM, class activation maps.

<https://doi.org/10.1371/journal.pone.0233079.g007>

**Table 3. Impact of image processing on the performance of the proposed CNN model in identifying glaucomatous discs using a mixed training dataset including training images of Drishti-GS and TVGH datasets.**

Methods included in image processing	Dataset of test images					
	TVGH			Drishti-GS		
	Specificity	Sensitivity	AUC (95% CI)	Specificity	Sensitivity	AUC (95% CI)
All <sup>a</sup>	95.4%	92.6%	0.984 (0.972–0.998)	100%	73.6%	0.937 (0.874–1.0)
All except ROI selection	93.4%	89.4%	0.978 (0.960–0.996)	100%	55.2%	0.887 (0.789–0.984)
All except preprocessing	97.7%	94.6%	0.993 (0.985–1.0)	84.6%	84.2%	0.897 (0.794–0.999)
All except augmentation	95.4%	96.8%	0.990 (0.997–1.0)	92.3%	68.4%	0.925 (0.849–1.0)

<sup>a</sup>ROI selection, preprocessing, and augmentation were performed. The augmentation was performed by rotation, shifting, sheering, and zooming with a final image number of 30000 using the 50 training images from the Drishti-GS dataset. CNN: convolutional neural network; TVGH, Taipei Veterans General Hospital; AUC, area under the curve; CI, confidence interval; ROI, region of interest.

<https://doi.org/10.1371/journal.pone.0233079.t003>

Table 4. Clinical characteristics for the test images and relationship with the prediction of the deep learning algorithm.

	POAG			Control		
	CNN prediction			CNN prediction		
	Correct N = 87	Incorrect N = 7	P- value <sup>a</sup>	Correct N = 83	Incorrect N = 4	P- value <sup>a</sup>
Age (years)	58.85±15.71	62.00±2.52	0.389	62.26±19.55	51.75±23.26	0.297
Sex (Female, %)	33 (37.9%)	1 (14.28%)	0.210	49 (59.03%)	4 (100%)	0.101
CDR	0.80±0.11	0.67±0.21	0.037	0.41±0.14	0.57±0.10	0.018
Visual field						
Mean deviation (dB)	-6.49±5.47	-2.16±2.27	0.015			
PSD (dB)	6.28±3.72	2.97±1.51	0.011			
Average RNFL thickness (μm)	72.79±11.06	80.71±10.92	0.750			

Seven images with wrong automated ROI selection were excluded in the analysis.

Values are presented as mean ± SD or percentage (%).

<sup>a</sup> Mann-Whitney U test was used for continuous variables, and Chi-square test for categorical variables. POAG, primary open angle glaucoma; CNN, convolutional neural network; D, diopter; CDR, cup-to-disc ratio; PSD, pattern standard deviation; RNFL, retinal nerve fiber layer.

<https://doi.org/10.1371/journal.pone.0233079.t004>

## Visualization of predictive features

CAM was used to reveal features recognized by our CNN model for classification. Areas critical for the classification were marked as red, while areas with low impact as blue. The CAM was different between the glaucomatous and healthy eyes. The hot spots were located most commonly at the cup area and region of parapapillary atrophy in the images of the glaucomatous eyes, but at the neuroretinal rim or along the retinal nerve bundles in the images of the healthy eyes. The superimposed images revealed the difference between the two groups (Fig 7).

## Discussion

This study found that CNN-based automated glaucoma detection using color fundus images acquired by a desktop fundus camera achieved high accuracy on test images whose quality is similar to that of the training dataset. However, the generalizability of the classification model depended on the diversity of the training data. Addition of a limited number of images from the target test population to the training dataset significantly increased the accuracy of the CNN detection model. With this modification, CNN achieved similar or better diagnostic performance than the ensemble model that considered CDR. CAM revealed that our CNN model captured characteristic areas of the optic disc as that through human perception to identify glaucomatous changes.

The key to achieving accurate classification of retinal images using deep learning is the number and quality of the training images [19,20]. Unlike the diagnosis and grading of diabetic retinopathy which uses only fundus photographs, the diagnosis of glaucoma involves various structural and functional examinations [18]. It is very difficult to collect tens of thousands of glaucomatous fundus images with definite diagnosis and complete profiles of functional and structural measurements. This issue was settled in some studies aiming for building CNN classifier using large number of fundus photographs by identifying referable GON but not definite GON based on findings of the fundus images and agreement between doctors [16,30]. The major limitation of this approach is that one cannot evaluate the diagnostic performance of any proposed system for images with different level of disease severity. Christopher et al reported that the performance of deep learning algorithm differed according to the severity of GON [15]. In line with their findings, glaucomatous eyes at an earlier stage were more likely to be misclassified as normal in our model. Thus, diagnostic performance of deep learning

algorithms as currently available diagnostic tool for glaucoma may vary according to the disease severity and is more likely to miss early GON [15].

The generalizability of our deep learning model trained with limited fundus images was inadequate. Lack of diversity of our database may explain the poor generalizability of our model; however, it is unclear if the generalizability of a deep learning model would be improved by using a larger number of images of various origins because verification of the CNN model on other database has not been thoroughly evaluated [15,16]. Phene et al reported that a deep learning algorithm trained with 58033 fundus images still had variable diagnostic performance in detecting GON on two validation datasets with an AUC of 0.940 and 0.858 [17]. Liu et al used 241032 images from 68013 patients to build a CNN classifier for automated glaucoma detection [30]. However, even with the largest image dataset to date, the CNN classifier still had suboptimal performance on images from subjects of different ethnicities and of variable qualities, with an AUC dropping from 0.996 with internal validation set to 0.923 and 0.823, respectively. Gulshan et al. found that the performance of a deep learning model in detecting diabetic retinopathy reached a plateau at around 60000 images [19]. Therefore, to improve the generalizability of a CNN classifier for glaucoma detection, approaches other than expanding the number of images of the training dataset should also be considered.

We proposed two approaches to improve the generalizability of our model: The first was to fine-tune the CNN model with limited number of images from the target test dataset which could dramatically improve the diagnostic performance of this mixed model for images of different origin. This approach is not novel but practical, with an aim to build clinic-based glaucoma screening algorithm using limited images from each clinic when neither a universal deep learning algorithm for glaucoma detection nor the access to large-scale glaucoma image datasets is available. Similar finding was noted in Diaz-Pinto et al's study [31]. They used 5 public datasets to evaluate the generalizability of their CNN model and found that the diagnostic performance of the model varied according to whether the training images of the test dataset were included for model training or not with an AUC varied from 0.9605 to 0.7678, respectively. The second was to use an ensemble model that considered the CDR when the confident score of the CNN classifier was low. The original goal of the ensemble model was to avoid missing cases with advanced GON. However, the results revealed that the fine-tuned CNN model achieved similar diagnostic performance as that of the ensemble model. Therefore, when a large image dataset of images with definite diagnosis of GON is not available, a CNN model could be first constructed using hundreds of images with high quality and then fine-tuned using some images from the target test population to improve the diagnostic power of the model in the target test population.

The CAM revealed that the class-discriminative regions of the CNN model co-localized with the known characteristics of GON. The hot spots of the CAM were located at the cup area in most images that were correctly classified as glaucoma, but at the neuroretinal rim and nerve bundles in the images of healthy eyes. This finding agreed with the report of Christopher et al that the inferior and superior regions of the neuroretinal rim are the most important parts for distinguishing between the eyes with GON and healthy eyes through deep learning algorithm [15]. These findings clarified the classification of glaucoma based on CNN model and may enable further modification of the CNN model to achieve higher accuracy.

Our study has some limitations. The major limitation is that the training dataset was collected at one medical center with limited number of images, which may lead to questionable generalizability of our model. As expected, our model showed poor performance on test images of other origin, which could be overcome by using the fine-tuned model. Further evaluation to verify the generalizability of deep learning algorithm for glaucoma detection is needed since the fundus images taken for glaucoma diagnosis may vary in layout, clarity, and

visibility of the RNFL, especially for population-based glaucoma screening. The second limitation of our study is that ROI centered at the ONH was used to construct the model to reduce the image size and complexity, which may result in loss of information and complicated pre-processing at clinical application. However, ROI extraction increased both the sensitivity and specificity of the model as compared with those of the whole images (Table 3). Third, we did not verify the performance of our approaches using different deep learning algorithms in addition to VGGNet because the performance and generalizability of the TVGH-CNN model did not change significantly with other CNN algorithms, such as AlexNet, GoogLeNet, or Xception (data not shown). In line with our findings, Christopher et al's study revealed that the difference of VGGNet, Inception, and ResNet50 in detecting GON is marginal, with an AUC of 0.89, 0.91, and 0.91 respectively [15]. Finally, cases with anomalous or extreme disc appearance and retinopathy except for drusen were excluded from the dataset; hence, our model may not be applicable to such cases. Meanwhile, we did not measure disc size, not count disc hemorrhage or peripapillary atrophy in this study, therefore the impact of disc size, disc hemorrhage and peripapillary atrophy on the performance of CNN classifier is unknown.

## Conclusions

The proposed deep learning model trained with limited fundus images achieved high accuracy in identifying GON on images of similar quality, and did not misclassify any cases with moderate to advanced disease. The model could be successfully fine-tuned to increase its diagnostic accuracy for images of different origins, which might be a feasible approach to build CNN classifier for clinic-based glaucoma screening. The CAM revealed that the model captured similar characteristics in identifying GON as those through human perception. Based on these findings, deep learning assisted detection of GON may be a viable option to assist glaucoma screening, but the generalizability of a deep learning model must be verified first before being applied to the targeted test population.

## Author Contributions

**Conceptualization:** Yu-Chieh Ko, Catherine Jui-Ling Liu, Chen-Yi Lee.

**Data curation:** Yu-Chieh Ko, Shih-Yu Wey, Yu-Fan Chang, Mei-Ju Chen.

**Formal analysis:** Yu-Chieh Ko, Shih-Yu Wey, Chen-Yi Lee.

**Funding acquisition:** Yu-Chieh Ko.

**Investigation:** Yu-Chieh Ko, Catherine Jui-Ling Liu, Chen-Yi Lee.

**Methodology:** Yu-Chieh Ko, Shih-Yu Wey, Wei-Ta Chen, Chen-Yi Lee.

**Project administration:** Shih-Hwa Chiou, Catherine Jui-Ling Liu.

**Resources:** Yu-Chieh Ko, Wei-Ta Chen, Yu-Fan Chang, Mei-Ju Chen, Shih-Hwa Chiou, Catherine Jui-Ling Liu.

**Software:** Shih-Yu Wey, Chen-Yi Lee.

**Supervision:** Shih-Hwa Chiou, Catherine Jui-Ling Liu, Chen-Yi Lee.

**Validation:** Yu-Chieh Ko, Shih-Yu Wey.

**Writing – original draft:** Yu-Chieh Ko, Shih-Yu Wey, Wei-Ta Chen, Catherine Jui-Ling Liu, Chen-Yi Lee.



**Writing – review & editing:** Yu-Chieh Ko, Shih-Hwa Chiou, Catherine Jui-Ling Liu, Chen-Yi Lee.

## References

1. Kingman S. Glaucoma is second leading cause of blindness globally. *Bull World Health Organ.* 2004; 82(11):887–8. <https://doi.org/S0042-96862004001100019> PMID: 15640929
2. Ernest PJ, Schouten JS, Beckers HJ, Hendrikse F, Prins MH, Webers CA. An evidence-based review of prognostic factors for glaucomatous visual field progression. *Ophthalmology.* 2013; 120(3):512–19. <https://doi.org/10.1016/j.ophtha.2012.09.005> PMID: 23211636
3. Chen PP. Blindness in patients with treated open-angle glaucoma. *Ophthalmology.* 2003; 110(4):726–733. [https://doi.org/10.1016/S0161-6420\(02\)01974-7](https://doi.org/10.1016/S0161-6420(02)01974-7) PMID: 12689894
4. Dielemans I, Vingerling JR, Wolfs RC, Hofman A, Grobbee DE, de Jong PT. The prevalence of primary open-angle glaucoma in a population-based study in The Netherlands. The Rotterdam Study. *Ophthalmology.* 1994; 101(11):1851–5. [https://doi.org/10.1016/s0161-6420\(94\)31090-6](https://doi.org/10.1016/s0161-6420(94)31090-6) PMID: 7800368
5. Mitchell P, Smith W, Attebo K, Healey PR. Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. *Ophthalmology* 1996; 103(10):1661–9. [https://doi.org/10.1016/s0161-6420\(96\)30449-1](https://doi.org/10.1016/s0161-6420(96)30449-1) PMID: 8874440
6. Burr JM, Mowatt G, Hernández R, Siddiqui MA, Cook J, Lourenco T, et al. The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation. *Health Technol Assess.* 2007; 11(41):iii-iv, ix-x, 1–190. <https://doi.org/10.3310/hta11410> PMID: 17927922
7. Thomas SM, Jeyaraman MM, Hodge WG, Hutnik C, Costella J, Malvankar-Mehta MS. The effectiveness of teleglaucoma versus in-patient examination for glaucoma screening: a systematic review and meta-analysis. *PLoS One.* 2014; 9(12):e113779. <https://doi.org/10.1371/journal.pone.0113779> PMID: 25479593
8. Hertzog LH, Albrecht KG, LaBree L, Lee PP. Glaucoma care and conformance with preferred practice patterns. Examination of the private, community-based ophthalmologist. *Ophthalmology.* 1996; 103(7):1009–13. [https://doi.org/10.1016/s0161-6420\(96\)30573-3](https://doi.org/10.1016/s0161-6420(96)30573-3) PMID: 8684788
9. Myers JS, Fudemberg SJ, Lee D. Evolution of optic nerve photography for glaucoma screening: a review. *Clin Exp Ophthalmol.* 2018; 46(2):169–176. <https://doi.org/10.1111/ceo.13138> PMID: 29280542
10. Badalà F, Nouri-Mahdavi K, Raoof DA, Leeprechanon N, Law SK, Caprioli J. Optic disk and nerve fiber layer imaging to detect glaucoma. *Am J Ophthalmol.* 2007; 144(5):724–32. <https://doi.org/10.1016/j.ajo.2007.07.010> PMID: 17868631
11. O'Neill EC, Gurria LU, Pandav SS, Kong YX, Brennan JF, Xie J, et al. Glaucomatous optic neuropathy evaluation project: factors associated with underestimation of glaucoma likelihood. *JAMA Ophthalmol.* 2014; 132(5):560–6. <https://doi.org/10.1001/jamaophthalmol.2014.96> PMID: 24699817
12. 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):581–96. <https://doi.org/10.1016/j.compmedimag.2013.09.005> PMID: 24139134
13. Garway-Heath DF, Ruben ST, Viswanathan A, Hitchings RA. Vertical cup/disc ratio in relation to optic disc size: its value in the assessment of the glaucoma suspect. *Br J Ophthalmol.* 1998; 82(10):1118–24. <https://doi.org/10.1136/bjo.82.10.1118> PMID: 9924296
14. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. *arXiv:1512.07108v6.* 2017 [cited 2017 Oct 19]. Available from: <https://arxiv.org/abs/1512.07108>.
15. Christopher M, Belghith A, Bowd C, Proudfoot JA, Goldbaum MH, Weinreb RN, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci Rep.* 2018; 8(1):16685. <https://doi.org/10.1038/s41598-018-35044-9> PMID: 30420630
16. 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. <https://doi.org/10.1016/j.ophtha.2018.01.023> PMID: 29506863
17. Phene S, Carter Dunn R, Hammel N, Liu Y, Krause J, Kitade N, et al. Deep learning to assess glaucoma risk and associated features in fundus images. *arXiv:1812.08911v1.* 2018 [cited 2018 Dec 21]. Available from: <https://arxiv.org/abs/1812.08911>.
18. Jonas JB, Aung T, Bourne RR, Bron AM, Ritch R, Panda-Jonas S. Glaucoma. *Lancet.* 2017; 390(10108):2183–2193. [https://doi.org/10.1016/S0140-6736\(17\)31469-1](https://doi.org/10.1016/S0140-6736(17)31469-1) PMID: 28577860

19. 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(22):2402–2410. <https://doi.org/10.1001/jama.2016.17216> PMID: 27898976
20. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018; 125(8):1264–1272. <https://doi.org/10.1016/j.ophtha.2018.01.034> PMID: 29548646
21. Sivaswamy J, Krishnadas S, Chakravarty A, Joshi GD, Ujjwal, Syed TA. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomed Imaging Data Pap*. 2015, 2(1):1004.
22. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. arXiv:1412.6806. 2014 [cited 2014 Dec 21]. Available from: <https://arxiv.org/abs/1412.6806>.
23. Simonyan K, Zisserman A. Very Deep Convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014 [cited 2014 Sep 4]. Available from: <https://arxiv.org/abs/1409.1556>.
24. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? arXiv:1411.1792v1. 2014 [cited 2014 Nov 6]. Available from: <https://arxiv.org/abs/1411.1792v1>
25. Orlando JI, Prokofyeva E, Fresno MD, Blaschko MB. Convolutional neural network transfer for automated glaucoma identification. Proceedings of the 12th International Symposium on Medical Information Processing and Analysis. 2016 Dec 5–7; Tandil Argentina; SPIE;2016. Vol. 10160; pp. 1–10. <https://doi.org/10.1117/12.2255740>.
26. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. arXiv:1512.04150. 2015 [cited 2015 Dec 14]. Available from: <https://arxiv.org/abs/1512.04150>.
27. Lin M, Chen Q, Yan S. Network in network. arXiv:1312.4400. 2013 [cited 2013 Dec 16]. Available from: <https://arxiv.org/abs/1312.4400>.
28. Ayub J, Ahmad J, Muhammad J, Aziz L, Ayub S, Akram U, et al. Glaucoma detection through optic disc and cup segmentation using K-mean clustering. Proceedings of 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube).2016 April 11–12; Quetta Pakistan; IEEE; 2016. <https://doi.org/10.1109/ICECUBE.2016.7495212>.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *JMLR*. 2011; 12:825–2830. <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>
30. Liu H, Li L, Wormstone IM, Qiao C, Zhang C, Liu P, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. 2019; 137(12):1353–1360. <https://doi.org/10.1001/jamaophthalmol.2019.3501>
31. Diaz-Pinto A, Morales S, Naranjo V, Kohler T, Mossi JM, Navea A. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed Eng Online*. 2019; 18(1):29. <https://doi.org/10.1186/s12938-019-0649-y> PMID: 30894178