A retrospective comparison of deep learning to manual annotations for optic disc and optic cup segmentation in fundus photos

Huazhu Fu²*, Fei Li¹*, Yanwu Xu³+, Jingan Liao⁴, Jian Xiong¹,
Jianbing Shen², Jiang Liu⁵,⁶, Xiulan Zhang¹+ for iChallenge-GON study group⁹

Affiliation:
1 State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University,
Guangzhou, Guangdong, China
2 Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates
3 Baidu Inc., Beijing, China
4 School of Computer Science and Engineering, South China University of Technology, Guangzhou,
Guangdong, China
5 Department of Computer Science and Engineering, Southern University of Science and Technology, Guangzhou, Guangdong, China
6 Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, Ningbo, Zhejiang, China

* These authors share first authorship on and contributed equally to this work.
+ These authors share corresponding authorship.
# iChallenge-GON study group includes:
Dr. Chunman Yang: The 2nd Affiliated Hospital of Guizhou Medical University, Kaili, Guizhou, China.
Dr. Fengbin Lin: Zhongshan Ophthalmic Center, Guangzhou, Guangdong, China.
Dr. Huang Luo: Guangzhou Hospital of TCM, Guangzhou, Guangdong, China.
Dr. Hao Li: Zhongshan Ophthalmic Center, Guangzhou, Guangdong, China.
Dr. Huixin Che: Aier Eye Hospital, Jinzhou, Liaoning, China.
Dr. Nuhui Li: Zhongshan Ophthalmic Center, Guangzhou, Guangdong, China.
Dr. Yazhi Fan: The 2nd Affiliated Hospital of Xi’an Jiaotong University, Shaanxi, China.

Corresponding Authors: Yanwu Xu, PhD, Baidu Inc., Beijing 100085, China. E-mail: ywxu@ieee.org; and Xiulan Zhang, MD, PhD, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510275, China. E-mail: zhangxl2@mail.sysu.edu.cn.

Short title: Deep learning for disc and cup segmentation in fundus

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abbreviations and Acronyms:

OD = optic disc; OC = optic cup; CDR = cup-to-disc ratio; MAE = mean absolute error; CI = confidence interval; AUC = area under the receiver operating characteristic curve; ONH = optic nerve head; RNFL = retinal nerve fiber layer; IOP = intraocular pressure; VF = visual fields; SD = standard deviation; OCT = optical coherence tomography; AS-OCT = anterior segment OCT.

Financial Support: The sponsor or funding organization had no role in the design or conduct of this research.

Conflict of Interest: no conflicting relationship exists for any author.
ABSTRACT

**Purpose:** Optic disc (OD) and cup (OC) segmentation are fundamental for fundus image analysis. Manual annotation is time consuming, expensive, and highly subjective, while an automated system is invaluable to the medical community. The aim of this study is to develop a deep learning system to segment OD and OC in fundus photos, and evaluate how the algorithm compares against manual annotations.

**Methods:** A total of 1200 fundus photos with 120 glaucoma cases were collected. The OD and OC annotations were labeled by seven licensed ophthalmologists, and glaucoma diagnoses were based on comprehensive evaluations of the subject medical records. A deep learning system for OD and OC segmentation was developed. The performances of segmentation and glaucoma discriminating based on the cup-to-disc ratio (CDR) of automated model were compared against the manual annotations.

**Results:** The algorithm achieved an OD dice of 0.938 (95% confidence interval (CI), 0.934-0.941), OC dice of 0.801 (95% CI, 0.793-0.809), and CDR mean absolute error (MAE) of 0.077 (95% CI, 0.073-0.082). For glaucoma discriminating based on CDR calculations, the algorithm obtained an area under receiver operator characteristic curve (AUC) of 0.948 (95% CI, 0.920-0.973), with a sensitivity of 0.850 (95% CI, 0.794-0.923) and specificity of 0.853 (95% CI, 0.798-0.918).

**Conclusions:** We demonstrated the potential of the deep learning system to assist ophthalmologists in analyzing OD and OC segmentation and discriminating glaucoma from non-glaucoma subjects based on CDR calculations.

**Translational Relevance:** We investigate the segmentation of OD and OC by deep learning system compared against the manual annotations.
INTRODUCTION

Glaucoma is the leading cause of irreversible blindness around the world\textsuperscript{1}. In clinical practice, glaucoma is diagnosed by evaluating the thickness of the retinal nerve fiber layer (RNFL), and the morphology of the optic nerve head (ONH)\textsuperscript{2,3}. Some other features are considered when making a diagnosis of glaucoma\textsuperscript{1,4}, including visual field, intraocular pressure, family history, corneal thickness, history of disc hemorrhages, etc. In fundus examinations, glaucoma is usually characterized by a larger cup-to-disc ratio (CDR), focal notching of the neuroretinal rim, etc.\textsuperscript{5,6}. An enlarged CDR may also indicate the existence of other ocular ailments, such as neuro-ophthalmic diseases. Previous studies have shown that a larger vertical CDR is closely associated with the progression of glaucoma\textsuperscript{7–9}. However, calculations for the CDR often vary among ophthalmologists and are relatively subjective, since they require a comprehensive judgment of shapes and structures of optic disc (OD) and optic cup (OC)\textsuperscript{10,11}. As such, several tools incorporating computer vision and machine learning techniques have been developed to perform automated OD and OC segmentation for large-scale data analysis.

Recently, deep learning techniques have been shown to perform very well in a wide variety of medical imaging tasks\textsuperscript{12,13}, including diabetic retinopathy screening\textsuperscript{14–16}, and age-related macular degeneration detection\textsuperscript{17–19}. Automated glaucoma detection from fundus photos has also received increasing attention\textsuperscript{20–23}. However, most of the studies have focused on predicting glaucoma directly from the fundus photos, without any visualization result. By contrast, OD and OC segmentation could be helpful for calculating the risk factors (e.g., CDR, Rim-to-Disc Ratio\textsuperscript{24}), and providing a segmentation visualization result. Moreover, although some automated segmentation methods appear to perform well on the small datasets\textsuperscript{25–28}, they have not been compared to performance by the practicing ophthalmologists.

In this study, we developed a deep learning system for automated OD and OC segmentation in fundus photos and evaluated its performances compared against seven ophthalmologists for OD and OC segmentation and glaucoma discriminating based on CDR calculations.
METHODS

Data acquisition

The fundus photos were collected from Zhongshan Ophthalmic Center, Sun Yat-sen University, China. The fundus photos were captured by using Zeiss Visucam 500 and Canon CR-2 machines. We included 1496 fundus photos from 748 subjects. As long as the diagnosis of both eyes are determined, both eyes of the same subjects were included. After a quality assessment, the low-quality fundus photos (e.g., low-contrast, blurry) are excluded. Finally, a total of 1200 fundus photos are selected in our study with 120 glaucoma and 1080 non-glaucoma cases (Inclusion criteria: 1. Age ≥ 18 years old; 2. Clear images without artifacts or overexposure; 3. Definite diagnoses acquired). Diagnoses were based on the comprehensive evaluation of the subjects' medical records, including fundus photos, IOP measurements, optical coherence tomography images, visual fields (VF). The fundus photos came from previous clinical studies, and all the participants signed informed consent before enrollment. IRB/Ethics Committee ruled that approval was not required for this study.

The dataset was split into a training set (400 photos with 40 glaucoma cases), a validation set (400 photos with 40 glaucoma cases, female: 52%, mean age: 25.3 ± 11.5), and a test set (400 photos with 40 glaucoma cases, female: 55%, mean age: 23.7 ± 9.0), following the REFUGE challenge. The photos from the same patient were assigned to the same set. The training set was used to learn the algorithm parameters, the validation set was used to choose model, and the test set was used to evaluate the algorithm, as well as the ophthalmologists.

Diagnostic criteria for glaucoma

Patients with glaucomatous damage in the ONH area and reproducible glaucomatous VF defects were included in our study. A glaucomatous VF defect is defined as a reproducible reduction in sensitivity compared to the normative dataset, in reliable tests, at: (1) two or more contiguous locations with p-value < 0.01, (2) three or more contiguous locations with p-value <
ONH damage is defined as CDR > 0.7, thinning of RNFL (an RNFL defect in the optic nerve head shown on the OCT reports), or both, without a retinal or neurological cause for VF loss. Specifically, First, the diagnostic criteria was based on the trial in glaucoma, i.e., UKGTS30. Second, if the points exist on the rim, there could be false-positive cases. However, as mentioned in our manuscript, the included subjects in our study received repeated VF tests to ensure reliability. If the defects exist all the time, we consider them as glaucomatous defects.

All OD and OC annotations were manually labeled by seven licensed ophthalmologists (average experience: 8 years, range: 5-10 years). All ophthalmologists independently reviewed and marked OD and OC in each photo as the tilted ellipses using a free image labelling tool with capabilities for image review, zoom, and ellipse fitting. Ophthalmologists did not have access to any patient information or knowledge of disease prevalence in the data. The final standard reference labels of OD and OC were created by merging the annotations from multiple ophthalmologists using majority voting. Specifically, a senior specialist with more than 10 years of experience in glaucoma performed a quality check afterward, analyzing the resulting masks to account for potential mistakes. When errors in the annotations were observed, this additional reader analyzed each of the seven segmentations, removed those that were considered failed in his/her opinion and repeated the majority voting process with the remaining ones. Only a few cases had to be corrected using this protocol.

Algorithm development

In this study, we proposed a deep learning system for automated OD and OC segmentation in fundus photos (Figure 1). The proposed system included two main stages: (1) OD region detection, which first localized the OD center within the whole fundus photo, and then cropped the OD region to remove the background; and (2) OD and OC segmentation, which segmented the OD and OC jointly via a multi-label deep network in the cropped OD image. We employed a U-Net network for OD region detection, which was based on encoder–decoder architecture to achieve satisfactory performances in many biomedical image tasks31. The encoder path consisted of the multiple convolutional layers with various filter banks to produce a set of
feature representations for the inputs, while the decoder path aggregated the feature
representations to predict the probability map of the OD region in the fundus photo.
Additionally, skip connections were used to concatenate the feature representations from the
encoder path to the corresponding decoder path. The final output of the U-Net network was a
probability map, indicating the OD region and background for each pixel in the fundus image, as
shown as Figure 1 (c). The implementation details of the U-Net network for OD detection were
given in the Supplement A. With the probability map of OD localization, we used a thresholding
of 0.5 to obtain the mask for the OD region, and cropped a local image around the OD for the
following OD and OC segmentation stage.

In the second stage of our algorithm, a multi-label network was utilized to segment OD
and OC simultaneously in the cropped OD region image\textsuperscript{26}. Similar to the U-Net network, the
multi-label network also consisted of an encoder and a decoder path based on convolutional
layers. The difference is that the multi-label network employed the average pooling layers to
naturally down-sample the images as multi-scale inputs to the corresponding encoder path, while
the multi-scale outputs from each scale of decoder path were fused together as the final
probability map. Additionally, the multi-label loss function was used to learn the binary
classifier of each class (i.e., OD and OC), and assign multiple labels to each pixel for
segmentation of OD and OC jointly. The implementation details of the multi-label network for
OD and OC segmentation were given in the Supplement B. In the fundus photo, the size ratio of
OC region is less than the OD and background, which could lead overfitting of deep model
during training. To address this, we map the OD region image into the polar coordinates, before
being fed into the multi-label network. Polar transformations were carried out using the OD
center as the origin and the local image width as the radius (see Figure 1 (e)). The
implementation details of polar transformations were given in the Supplement C. After passing
through the multi-label network, an inverse polar transformation reverted the predicted map back
to the original coordinates.

The U-Net network for OD detection and multi-label network for OD and OC
segmentation were trained separately. The U-Net network was trained based on the whole fundus images resized to 800 by 800 pixels, with the OD reference label, while the multi-label network was trained based on the OD region images resized to 400 by 400 pixels, with the OD and OC reference labels. Random flips and rotations were applied to all training photos before they were fed into the networks for data augmentation. These two networks were implemented with Python (version 3.6) based on Keras (version 2.2) with a Tensorflow (version 1.12) backend. All network parameters of the networks were optimized by using stochastic gradient descent with a learning rate of 0.0001 and a momentum of 0.9. In order to prevent the networks from overfitting, early stopping was performed, which saved the network model after each epoch and chose the final model with the lowest loss on the validation set. Each stage of training required around 2 hours for completion, on a single NVIDIA Titan XP.

**Statistical analysis and evaluation**

For segmentation evaluation, we reported three performance metrics, namely, OD dice, OC dice, and CDR mean absolute error (MAE). The dice scores measured the overlap ratio between the target regions of the reference label and segmented result, while CDR MAE was the mean absolute error between the calculated CDR values from the reference label and segmented result. We also determined the standard deviation (SD) and 95% Bayesian confidence interval (CI)\(^{32}\) for each segmentation metric.

In addition to evaluating the segmentation performance, we also compared the algorithm against ophthalmologists for discriminating glaucoma from non-glaucoma photos based on CDR calculations. The performances across different diagnostic thresholds of CDR were assessed in terms of the area under receiver operator characteristic curve (AUC). To convert the CDR to a binary prediction, we chose the highest point on the ROC curve, which offers minimal trade-off between sensitivity and specificity, as the final discriminating threshold. Moreover, the 95% bootstrapping CI\(^{33}\) was provided for each discriminating metric as: computing 10,000 bootstrap replicates from the set, and each metric was computed for algorithm and reference label on the same bootstrap replicate. p-values were reported by comparing the AUC with the algorithm and
ophthalmologist predictions. All statistical analyses were performed using Python (version 3.6) with SciPy (version 1.2) and Scikit-learn (version 2.20). Figures were created using Matplotlib (version 3.0) and Seaborn (version 0.9).

RESULTS

The segmentation performances of our algorithm and annotations of the seven ophthalmologists, for the test set, were listed in Table 1. For glaucoma data, the algorithm obtained an OD dice of 0.941 (SD, 0.057, 95% CI, 0.926-0.956), OC dice of 0.864 (SD, 0.089, 95% CI, 0.841-0.887), and CDR MAE of 0.065 (SD, 0.056, 95% CI, 0.051-0.080). For non-glaucoma data, the algorithm predicted an OD dice of 0.937 (SD, 0.040, 95% CI, 0.934-0.941), OC dice of 0.794 (SD, 0.096, 95% CI, 0.786-0.803), and CDR MAE of 0.079 (SD, 0.050, 95% CI, 0.074-0.083). The segmentation performances of the algorithm on the whole test set achieved an OD dice of 0.938 (SD, 0.041, 95% CI, 0.934-0.941), OC dice of 0.801 (SD, 0.097, 95% CI, 0.793-0.809), and CDR MAE of 0.077 (SD, 0.051, 95% CI, 0.073-0.082). For OD segmentation, the algorithm performed better than Ophthalmologist 2, who reported an OD dice of 0.928 (SD, 0.046, 95% CI, 0.925-0.932), and Ophthalmologist 3, who determined the OD dice to be 0.924 (SD, 0.039, 95% CI, 0.921-0.927). For OC segmentation, the algorithm performed better than Ophthalmologist 1, who obtained an OC dice of 0.705 (SD, 0.121, 95% CI, 0.695-0.715), and Ophthalmologist 7, who got an OC dice of 0.670 (SD, 0.138, 95% CI, 0.658-0.681). The OD and OC dice scores of inter-agreement for seven ophthalmologists were given in Figure 2. Boxplots for the calculated CDRs of the reference label, the ophthalmologist annotations and the algorithm outputs, for glaucoma and non-glaucoma data on test set, were plotted in Figure 3. The average CDRs of the reference labels for the glaucoma and normal cases were 0.656 and 0.453, respectively for test set.

Figure 4 showed the visual results of automated OD and OC segmentation, for both glaucoma and non-glaucoma data. Several failure cases were also provided in Figure 4 (E, F). One common failure case for OD segmentation was confusion when peripapillary atrophy (PPA)
was present, since this looks similar to the OD (green arrow in Figure 4 (E)). Failure cases also occurred due to the low-quality of the fundus photos, where poor illumination and low-contrast often made it difficult to determine the boundary of the OC (green arrow in Figure 4 (F)). However, this could be relieved using additional image enhancement pre-processing.

The performances of discriminating glaucoma from non-glaucoma subjects based on CDR, for test set, were shown in Figure 5 and Table 2. The algorithm obtained an AUC of 0.948 (95% CI, 0.920-0.973), with a sensitivity of 0.850 (95% CI, 0.794-0.923) and specificity of 0.853 (95% CI, 0.798-0.918). The algorithm obtained the rank 2 discriminating performance, only lower than ophthalmologist 2, who got an AUC of 0.956 (95% CI, 0.933-0.975, p-value < 0.0001). Moreover, Figure 4 (E) and (F) showed the false negative and false positive samples, respectively.

DISCUSSION

The purpose of this study was to develop a deep learning algorithm for automated OD and OC segmentation in fundus photos and compare its performance to ophthalmologist annotations. The results demonstrated that the proposed deep learning algorithm achieved satisfactory performances on the OD and OC segmentation task and the glaucoma discriminating task based on CDR calculations.

OD and OC segmentation are fundamental for fundus analysis, especially for CDR calculations during discriminating glaucoma from non-glaucoma subjects. Developing an automated system for this task is crucial. First, as briefly mentioned, manual fundus photo labelling is highly time-consuming, with the average ophthalmologist requiring 40 seconds to annotate a single photo. Because our algorithm could reduce this time to 2 second, it would be highly beneficial for accelerating processing time and analyzing large-scale datasets. Second, manual annotations are highly subjective. In fact, the segmentations carried out by the ophthalmologists were easily affected by both fundus resolution and image quality. The inter-
agreement rating between the various ophthalmologists, for both OD and OC dice scores on the test set, were shown in Figure 2. As can be seen, there was slight variation between the OD segmentation results, with inter-agreement scores ranging from 0.89 to 0.95. However, the OC segmentation task suffered a larger variability, with inter-agreement scores ranging from 0.47 to 0.85. The boundary of OD was clear and definite enough to determine in fundus photo, which produced a high inter-agreement score between the by ophthalmologists, as shown in Figure 2 (A). Different from the OD, the boundary of OC was more difficult to identify, which was influenced by many factors such as tilted disc, illumination, and low contrast, etc. These factors may result in the clinical uncertainty during different ophthalmologists and a variable OC segmentation. Moreover, OC segmentation by an ophthalmologist was a highly subjective task, which was related to individual bias and clinical experiences. This also led a low inter-agreement score (See Figure 2 (B)). By contrast, the automated algorithm provided a consistent result for the same photo with freezing the trained parameters and model. Moreover, due to limited GPU memory capabilities and parameter size constraints, input fundus photos had to be down-sampled for training, thus removing the requirement for high-resolution photos. Another observation is that the performances of algorithm on glaucoma cases (OD dice of 0.941, cup dice of 0.864, CDR MAE of 0.065) was better than its on non-glaucoma cases (OD dice of 0.937, cup dice of 0.794, CDR MAE of 0.079). One reason is that the advanced glaucoma cases with severe cupping usually present more clear interfaces between the OD and OC.

Over the decades, many automated deep learning algorithms have been proposed for glaucoma diagnosis in fundus photos, optical coherence tomography (OCT), and anterior segment OCT (AS-OCT). However, while many of these produce diagnostic results from fundus photos directly, they lacked clinical interpretability and analyticity. By contrast, segmentation based algorithms generate a visible segmentation result and have more potential for clinical assistant and analysis. Some automated algorithms based on various visual features and machine learning techniques have been developed for segmenting OD and OC. Cheng et al. classified each superpixel in fundus image with various hand-crafted...
features as OD and OC segmentation and reported an OD dice of 0.905 and OC dice of 0.759. Zheng et al.\textsuperscript{41} integrated the OD and OC segmentation within a graph-cut framework. However, they only utilized hand-crafted features, which were affected by the low quality of fundus photos. In our study, a multi-label deep network was employed to obtain highly discriminative representations and segment the OD and OC jointly with the multi-label loss. The results demonstrated that the proposed method enabled automated OD and OC segmentation with a comparable performance to ophthalmologists. We also evaluated the model for discriminating glaucoma from non-glaucoma subjects based on CDR calculations, which were calculated based on the segmentation results as an important glaucoma indicator. The proposed algorithm performed extremely well in comparison to ophthalmologists for glaucoma discriminating.

One limitation of this study was a specific Chinese population was evaluated and the results may not apply to other ethnic groups. Another potential limitation of our study was that the fundus photos were only taken using Zeiss Visucam 500 and Canon CR-2 cameras. This could possibly have a negative effect on the quality and performance when the algorithm was applied to images from other fundus acquisition devices. Third, in our study, we added CDR calculations as one of the clues for glaucoma diagnosis. However, some patients were shown to have a small CDR despite significant visual field loss, while others displayed a large CDR without reporting any VF loss\textsuperscript{7}. The dataset contained 10\% of glaucoma subjects and most of these glaucoma subjects were at moderate or advanced stage, the difficulty of discriminating glaucoma from non-glaucoma subjects based on CDR calculation was relatively lower. The performance of the algorithm may go down in another larger dataset. Future studies were needed to explore whether other annotations such as RNFL defects would further enhance the performance of the algorithm. Besides, early-stage glaucoma is very hard to diagnose through fundus photos. It would be interesting to add more photos from early-stage patients and train the algorithm to make diagnosis. We may try to find new clues other than CDR or RNFL defects in glaucoma discriminating based on fundus photos.

In summary, we developed and investigated a deep learning system for OD and OC
segmentation in fundus images. Deep learning technique was shown to be a promising
technology for helping clinicians to reliably and rapidly identify OD and OC regions. Moreover,
we also evaluated discriminating glaucoma from non-glaucoma subjects based on the CDR
calculations, where the proposed algorithm performed extremely well in comparison to
ophthalmologists, obtaining an AUC of 0.946. As such, our technique shown high potential for
assisting ophthalmologists in fundus analysis and glaucoma screening.
REFERENCES


20. Zhao H, Li H, Maurer-Stroh S, Cheng L. Synthesizing retinal and neuronal images with


Figure Captions:

Figure 1. Deep learning system for automated segmentation in fundus images. The algorithm included two stages: Optic disc (OD) region detection and OD and optic cup (OC) segmentation. For a given fundus image (a), the U-Net network (b) was utilized to detect the OD region (c). With the cropped OD region (d), a polar transformation was used to map the image into polar coordinate (e). A multi-label network (f) segmented the OD and OC jointly, and an inverse transformation returned the output map (g) back to original coordinates (h).

Figure 2. Dice scores of inter-agreement for seven ophthalmologists on the test set for (A) optic disc, and (B) optic cup.

Figure 3. Boxplots of the calculated cup-to-disc ratio (CDR) from segmentation results on test set.

Figure 4. The visual results of segmentation. The segmented optic disc and optic cup regions were labeled by red and yellow colors, respectively. (a, b) Glaucoma cases, (c, d) non-glaucoma cases, (e, f) failure cases.

Figure 5. (A) The average receiver operating characteristic curves (AUC) for glaucoma diagnosis based on cup to disc ratio (CDR) on test set. (B) The precision-recall curves for glaucoma diagnosis based on cup to disc ratio (CDR) on test set.
Figure 1. Deep learning system for automated segmentation in fundus images. The algorithm included two stages: Optic disc (OD) region detection and OD and optic cup (OC) segmentation. For a given fundus image (a), the U-Net network (b) was utilized to detect the OD region (c). With the cropped OD region (d), a polar transformation was used to map the image into polar coordinate (e). A multi-label network (f) segmented the OD and OC jointly, and an inverse transformation returned the output map (g) back to original coordinates (h).
Figure 2. Dice scores of inter-agreement for seven ophthalmologists on the test set for (A) optic disc, and (B) optic cup.

Figure 3. Boxplots of the calculated cup-to-disc ratio (CDR) from segmentation results on test set.
Figure 4. The visual results of segmentation. The segmented optic disc and optic cup regions were labeled by red and yellow colors, respectively. (a, b) Glaucoma cases, (c, d) non-glaucoma cases, (e, f) failure cases.

Figure 5. (A) The average receiver operating characteristic curves (AUC) for glaucoma diagnosis based on cup to disc ratio (CDR) on test set. (B) The precision-recall curves for glaucoma diagnosis based on cup to disc ratio (CDR) on test set.
Table 1: Segmentation performances of ophthalmologists and algorithm on test set. OD = optic disc; OC = optic cup; CDR = cup-to-disc ratio; MAE = mean absolute error; CI = confidence interval; SD = standard deviation.

<table>
<thead>
<tr>
<th></th>
<th>OD dice (SD, 95% CI)</th>
<th>OC dice (SD, 95% CI)</th>
<th>CDR MAE (SD, 95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Glaucoma data (40 images)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ophthalmologist 1</td>
<td>0.963 (0.962-0.970)</td>
<td>0.888 (0.900-0.894)</td>
<td>0.057 (0.055-0.059)</td>
</tr>
<tr>
<td>Ophthalmologist 2</td>
<td>0.945 (0.936-0.954)</td>
<td>0.867 (0.843-0.890)</td>
<td>0.057 (0.045-0.049)</td>
</tr>
<tr>
<td>Ophthalmologist 3</td>
<td>0.942 (0.936-0.952)</td>
<td>0.887 (0.864-0.899)</td>
<td>0.039 (0.035-0.045)</td>
</tr>
<tr>
<td>Ophthalmologist 4</td>
<td>0.953 (0.946-0.956)</td>
<td>0.894 (0.818-0.902)</td>
<td>0.052 (0.044-0.052)</td>
</tr>
<tr>
<td>Ophthalmologist 5</td>
<td>0.947 (0.939-0.955)</td>
<td>0.865 (0.849-0.889)</td>
<td>0.055 (0.054-0.056)</td>
</tr>
<tr>
<td>Ophthalmologist 6</td>
<td>0.954 (0.946-0.962)</td>
<td>0.904 (0.864-0.921)</td>
<td>0.048 (0.042-0.051)</td>
</tr>
<tr>
<td>Ophthalmologist 7</td>
<td>0.954 (0.947-0.962)</td>
<td>0.719 (0.638-0.878)</td>
<td>0.149 (0.091-0.197)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>0.941 (0.907-0.956)</td>
<td>0.864 (0.829-0.887)</td>
<td>0.065 (0.056-0.076)</td>
</tr>
<tr>
<td><strong>Non-glaucoma data (360 images)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ophthalmologist 1</td>
<td>0.956 (0.953-0.959)</td>
<td>0.686 (0.107-0.766)</td>
<td>0.157 (0.151-0.163)</td>
</tr>
<tr>
<td>Ophthalmologist 2</td>
<td>0.926 (0.924-0.931)</td>
<td>0.842 (0.834-0.849)</td>
<td>0.053 (0.049-0.057)</td>
</tr>
<tr>
<td>Ophthalmologist 3</td>
<td>0.922 (0.918-0.925)</td>
<td>0.827 (0.831-0.844)</td>
<td>0.041 (0.038-0.044)</td>
</tr>
<tr>
<td>Ophthalmologist 4</td>
<td>0.949 (0.946-0.952)</td>
<td>0.821 (0.791-0.866)</td>
<td>0.056 (0.043-0.059)</td>
</tr>
<tr>
<td>Ophthalmologist 5</td>
<td>0.945 (0.942-0.948)</td>
<td>0.870 (0.843-0.876)</td>
<td>0.041 (0.036-0.044)</td>
</tr>
<tr>
<td>Ophthalmologist 6</td>
<td>0.953 (0.950-0.956)</td>
<td>0.903 (0.898-0.909)</td>
<td>0.034 (0.031-0.037)</td>
</tr>
<tr>
<td>Ophthalmologist 7</td>
<td>0.955 (0.952-0.958)</td>
<td>0.664 (0.652-0.676)</td>
<td>0.130 (0.063-0.156)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>0.937 (0.934-0.941)</td>
<td>0.794 (0.786-0.803)</td>
<td>0.079 (0.050-0.074)</td>
</tr>
<tr>
<td><strong>All data (400 images)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ophthalmologist 1</td>
<td>0.956 (0.954-0.959)</td>
<td>0.705 (0.121-0.695)</td>
<td>0.147 (0.141-0.153)</td>
</tr>
<tr>
<td>Ophthalmologist 2</td>
<td>0.928 (0.925-0.932)</td>
<td>0.844 (0.837-0.851)</td>
<td>0.053 (0.050-0.056)</td>
</tr>
<tr>
<td>Ophthalmologist 3</td>
<td>0.924 (0.921-0.927)</td>
<td>0.843 (0.836-0.849)</td>
<td>0.041 (0.038-0.044)</td>
</tr>
<tr>
<td>Ophthalmologist 4</td>
<td>0.949 (0.947-0.952)</td>
<td>0.809 (0.800-0.819)</td>
<td>0.056 (0.052-0.059)</td>
</tr>
<tr>
<td>Ophthalmologist 5</td>
<td>0.945 (0.942-0.948)</td>
<td>0.870 (0.863-0.876)</td>
<td>0.043 (0.040-0.046)</td>
</tr>
<tr>
<td>Ophthalmologist 6</td>
<td>0.953 (0.950-0.956)</td>
<td>0.903 (0.898-0.909)</td>
<td>0.035 (0.032-0.038)</td>
</tr>
<tr>
<td>Ophthalmologist 7</td>
<td>0.955 (0.952-0.957)</td>
<td>0.670 (0.658-0.681)</td>
<td>0.132 (0.066-0.170)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>0.938 (0.934-0.941)</td>
<td>0.801 (0.793-0.809)</td>
<td>0.077 (0.051-0.082)</td>
</tr>
</tbody>
</table>

Table 2: Diagnosis performances of experts and algorithm on test set. AUC = area under the receiver operating characteristic curve; CI = confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>AUC (95% CI)</th>
<th>p-value</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Precision (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference label</td>
<td>0.946 (0.911-0.971)</td>
<td>&lt; 0.0001</td>
<td>0.875 (0.811-0.927)</td>
<td>0.867 (0.819-0.923)</td>
<td>0.700 (0.588-0.791)</td>
</tr>
<tr>
<td>Ophthalmologist 1</td>
<td>0.922 (0.884-0.955)</td>
<td>&lt; 0.0001</td>
<td>0.875 (0.800-0.897)</td>
<td>0.858 (0.812-0.887)</td>
<td>0.585 (0.464-0.694)</td>
</tr>
<tr>
<td>Ophthalmologist 2</td>
<td>0.956 (0.933-0.975)</td>
<td>&lt; 0.0001</td>
<td>0.875 (0.826-0.930)</td>
<td>0.869 (0.834-0.928)</td>
<td>0.700 (0.577-0.876)</td>
</tr>
<tr>
<td>Ophthalmologist 3</td>
<td>0.935 (0.902-0.964)</td>
<td>&lt; 0.0001</td>
<td>0.850 (0.791-0.931)</td>
<td>0.850 (0.795-0.908)</td>
<td>0.630 (0.536-0.771)</td>
</tr>
<tr>
<td>Ophthalmologist 4</td>
<td>0.916 (0.871-0.954)</td>
<td>&lt; 0.0001</td>
<td>0.850 (0.791-0.930)</td>
<td>0.847 (0.796-0.898)</td>
<td>0.575 (0.438-0.667)</td>
</tr>
<tr>
<td>Ophthalmologist 5</td>
<td>0.911 (0.866-0.951)</td>
<td>&lt; 0.0001</td>
<td>0.850 (0.756-0.892)</td>
<td>0.856 (0.755-0.886)</td>
<td>0.625 (0.525-0.755)</td>
</tr>
<tr>
<td>Ophthalmologist 6</td>
<td>0.947 (0.916-0.972)</td>
<td>&lt; 0.0001</td>
<td>0.900 (0.824-0.930)</td>
<td>0.881 (0.832-0.922)</td>
<td>0.659 (0.556-0.773)</td>
</tr>
<tr>
<td>Ophthalmologist 7</td>
<td>0.916 (0.876-0.951)</td>
<td>&lt; 0.0001</td>
<td>0.825 (0.762-0.882)</td>
<td>0.825 (0.772-0.878)</td>
<td>0.625 (0.469-0.740)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>0.948 (0.920-0.973)</td>
<td>0.0001</td>
<td>0.850 (0.794-0.923)</td>
<td>0.853 (0.798-0.918)</td>
<td>0.700 (0.600-0.800)</td>
</tr>
</tbody>
</table>