Classification for referable glaucoma with fundus photographs using multimodal deep learning

Yeong Chan Lee[1], **Hyun Bin Cho**[2], **and Yoon Ho Choi**[2]

[1]Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea
[2]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Samsung Medical Center, Seoul, Republic of Korea
*All authors contributed equally to this work.

Glaucoma is a silent eye disease leading to blindness. Artificial intelligence models for predicting glaucoma using fundus photographs have been developed, however, it is important to evaluate robustness for outliers with high performance to classify glaucoma. We joined The AIROGS challenge, developed, and evaluated a multi-modal deep learning model to predict glaucoma and assess ungradability which is uncertainty for classifying certain class given images. We achieved 0.7635 for the partial AUROC, 0.6125 for sensitivity at 95% specificity, 0.5316 for ungradability kappa, and 0.8057 for ungradability AUROC in preliminary test phase 2.

Correspondence: ***Yeong Chan Lee, conan_8th@naver.com***

Introduction

Glaucoma is a progressive eye disease leading to blindness due to chronic damage to the optic nerve. It may be asymptomatic in early glaucoma, therefore, the key to preventing loss of vision from glaucoma is early detection and treatment by ophthalmologists [1]. Artificial intelligence models have been widely used to predict various eye diseases from fundus images [2]. However, the models were not evaluated with robustness for dealing with out-of-distribution data when maintaining high performance to predict glaucoma. To be used in real-world scenarios, it is important to determine whether a fundus image has enough information to diagnose diseases.

Challenge objectives and constraints. The AIROGS challenge was organized to predict glaucoma with a real-world fundus photograph using computational algorithms. The algorithms will be evaluated for screening performance and robustness. The participants could not access the test set including fundus photographs and ungradable images which cannot be decided as glaucoma. Glaucoma in the test set should be predicted within 10 seconds per single image. In the final test, the participants submitted the predicted results and a short paper only once.

Methods

Dataset. We obtained a total of 101,442 fundus photographs including 98,172 non-referable glaucoma (NRG) and 3,270 referable glaucoma (RG) from The Rotterdam EyePACS AIROGS dataset [3] RG accounts for only 3.2% in the imbalanced dataset. Thus, we chose an undersampling strategy for preventing overfitting to NRG during the all three phases (Table 1). A phase means a process to train and validate a model. The photographs of NRG were randomly selected same as the number of images of RG (3,270). We constructed three datasets with undersampled NRG images and RG images. Few images were excluded if optic disc had not been detected in the image.

Table 1. Dataset configuration. In phases 1–3, we selected all different images for non-referable glaucoma.

	Phase 1	**Phase 2**	**Phase 3**
Training set			
Referable glaucoma	2,588	2,588	2,614
Non-referable glaucoma	2,550	2,575	2,605
Validation			
Referable glaucoma	631	631	640
Non-referable glaucoma	660	644	649

Image preprocessing. Overall, we converted images to grayscale, resized the images with 608 pixels of width and 608 pixels of height, and applied the contrast limited adaptive histogram equalization (CLAHE) to fundus photographs for emphasizing the features of the fundus. We rotated the images within 45 degrees, enlarged within range of 70% and 130%, flipped horizontally and vertically, shifted within 200 pixels, and brightened within range of 30% and 170%. We augmented the images brighter or darker because the brightness of fundus photographs was diverse.

Segmentation of an optic disc using weakly supervised learning. It is essential to identify the glaucomatous changes at the optic disc. So, we decided to develop a model for detecting and segmenting the optic disc in the fundus photograph. We annotated the position of the optic disc in the 101 images (TRAIN000000 to TRAIN000100) roughly. Then we did not label the position with white (RGB, [255, 255, 255]), but label the position with gradient through multivariate normal distribution. The probability density function of the multivariate normal distribution $\mathbf{X} = [X_1, X_2]$ was generated with mean of $\boldsymbol{\mu} = [0,0]$ and standard deviation of

$\boldsymbol{\sigma} = [0.7, 0.7]$ in the rectangle annotation. The generated values $\mathbf{x}$ are rescaled by

$$\frac{\mathbf{x}}{\max(\mathbf{x}) \times 255}$$

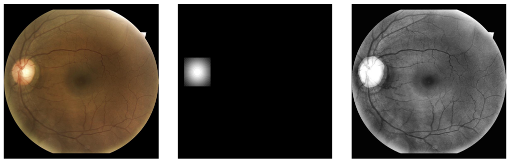

Fig. 1. An example of paired an original image (left), a roughly labeled masked image for the optic disc (center) and the predicted result (right).

We developed a basic U-net architecture [4] combined with DenseNet121 [5] as an encoder for a segmentation task. The shape of input was width of 608 and height of 608 and the channel of input was 1. We used a stochastic gradient descent algorithm with an initial learning rate of 0.001 dropped the rate of 0.9 at every 20 epochs for optimization. We defined a loss function with a dice coefficient metric. We chose the best model with the lowest validation loss within 10,100 steps per 100 epochs.

Then we extracted the predicted optic disc with the identical size (608 for width, 608 for height) from the centroid of the segmented object from each original fundus photograph. If the segmented object was not detected in an image, the image was excluded from the dataset. A total of 684 images were excluded from 101,442 images. We, as non-specialized ophthalmologists, discovered 18 false positives not fully including optic disc in 1,001 fundus photographs (TRAIN000000 to TRAIN001000).

The architecture of our multi-modal model. Our multi-modal neural network comprises two MobileNetV2[6] models which feed two inputs with full fundus photograph and the extracted optic disc. First, we added a fully connected layer with 128 neurons and a dropout layer with a dropout rate of 0.2 after the last convolution layer of each MobileNetV2 model. Then we concatenated the two dropout layers from the two MobileNetV2. Lastly, we additionally added a fully connected layer, a dropout layer, a prediction layer. Before we combined the two MobileNetV2 networks, we pretrained the models with the dataset in phase 1, respectively. Then, after building the multi-modal model, the model was trained with categorical cross-entropy loss and stochastic gradient descent with a learning rate of 0.001 and a momentum of 0.9 using datasets of phases 1–2. In the last phase, we specifically used F1 loss with weight of 0.7 and categorical cross-entropy with weight of 0.3.

Uncertainty. Using dropout layers, we measured uncertainty for each image to assess the robustness of the model [7]. We obtained 20 predicted probabilities for each image, then we tested statistically whether the mean of values is 0.5 or not using Wilcoxon one-sample test. We defined ungradability for predicting glaucoma as a common logarithm of the p-value for the Wilcoxon test.

Evaluation. Screening performance was evaluated using the partial area under the receiver operator characteristics curve (AUROC) over the specificity of 0.9 and sensitivity at the specificity of 0.95. Ungradability for images was evaluated with Cohen's kappa and AUROC using human references and predictions for robustness.

Results

We obtained 0.7635 for the partial AUROC, 0.6125 for sensitivity at 95% specificity, 0.5316 for ungradability kappa, and 0.8057 for ungradability AUROC in preliminary test phase 2. Final test results are shown in https://airogs.grand-challenge.org/evaluation/final-test-phase/leaderboard/.

References

1. Joshua D Stein, Anthony P Khawaja, and Jennifer S Weizer. Glaucoma in adults—screening, diagnosis, and management: A review. *Jama*, 325(2):164–174, 2021.
2. Zhaoran Wang, Pearse A Keane, Michael Chiang, Carol Y Cheung, Tien Yin Wong, and Daniel Shu Wei Ting. Artificial intelligence and deep learning in ophthalmology. *Artificial Intelligence in Medicine*, pages 1–34, 2020.
3. Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. Rotterdam eyepacs airogs train set, December 2021. The previous version was split into two records. This new version contains all data and the second record is deprecated.
4. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
5. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
6. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
7. Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.