Differentiating Glaucomatous Optic Neuropathy from Non-Glaucomatous Optic Neuropathies Using Deep Learning Algorithms

Short title: Deep Learning for Optic Neuropathies

Mahsa Vali[1], Massoud Mohammadi[2], Nasim Zarei[2], Melika Samadi[2], Amir Atapour-Abarghouei[3], Wasu Supakontanasan[4], Yanin Suwan[4], Prem S. Subramanian[5], Neil R Miller[6], Rahele Kafieh[7], Masoud Aghsaei Fard[2]

[1]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran

[2]Farabi Eye Hospital, Tehran University of Medical Sciences, Tehran, Iran

[3]Department of Computer Science, Durham University, Durham, UK.

[4] Glaucoma Service, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[5] Departments of Ophthalmology, Neurology, and Neurosurgery, Sue Anschutz-Rodgers University of Colorado Eye Center, Aurora, CO

[6] Department of Neuro ophthalmology, Johns Hopkins Hospital, Wilmer Eye institute, Baltimore, MD

[7]Department of Engineering, Durham University, South Road, Durham, UK

Surname indicated by underline.

Corresponding author: Masoud Aghsaei Fard, MD, FICO, Farabi Eye Hospital, Qazvin Sq, Tehran, Iran, Postal code: 13138, Email: masood219@gmail.com

Introduction:

Glaucoma is an optic neuropathy characterized in part by cupping of the optic disc with corresponding visual field changes.[1] Non-glaucomatous optic neuropathies (NGON) eventually produce optic disc changes that can appear similar to the changes seen in glaucomatous optic neuropathy (GON), including cupping.[2] It is important to differentiate GON from NGON, as the evaluation and management of these two entities is completely different. GON is a danger to vision and requires ocular therapy alone, whereas NGON often have systemic and neurologic associations that not only are vision-threatening but also life-threatening.

The clinical differentiation between GON optic disc changes and NGON-related disc changes can be difficult, particularly when there is optic disc cupping.[3-6] Whereas optic disc cupping may be seen in NGON, such as methanol toxicity, intracranial tumors, optic neuritis, dominant optic atrophy, periventricular leukomalacia, and arteritic anterior ischemic optic neuropathy,[3-6] the depth of cupping is reported to be more profound in eyes with glaucoma.[7, 8] In addition, NGON that include cupping have a greater degree of neuroretinal rim pallor than GON.[3, 5, 6] Conversely, moderate-to-severe glaucoma may produce not only cupping but also secondary optic disc pallor that easily can be confused with that seen in NGON.[9]

Artificial intelligence (AI) has been used to detect a variety of optic disc abnormalities in neuro-ophthalmology.[10] Using deep learning (DL), The Brain and Optic Nerve Study with Artificial Intelligence (BONSAI) consortium was able to identify papilledema and other optic disc abnormalities.[11] Several studies also have used DL for glaucoma detection and found that the performance of DL in detecting glaucoma compares favorably with that of experts.[12-16] However, only one study has used DL for differentiating GON from NGON. Yang et al.[17] applied a ResNet-50 to the problem and reported a sensitivity of 93.4% and a specificity of 81.8% in distinguishing GON from NGON; however, the number of GON and NGON cases that were used in this study was fewer than the number of normal optic discs. In addition, the authors did not perform optic disc segmentation, nor did they perform external validation of the DL data set. Therefore, we elected to use much larger data sets using transfer learning and optic disc segmentation to see if we could corroborate the findings of Yang et al.[17] or even improve the precision with which our algorithm differentiated GON from NGON. We also externally validated our algorithm with data from several other centers with patients of different ethnicities.

Methods:

We performed a training, validation, and external-testing study using digital color ocular fundus photograph after approval by the institutional review board of Tehran University of Medical Sciences and at each contributing institution. All investigations adhered to the tenets of the Declaration of Helsinki. We retrospectively obtained fundus photographs from one or both eyes of patients with well-documented GON and NGON as well from persons with no evidence of ocular disease. For training and evaluation of classification networks, we used subject-wise 5-fold cross-validation on images from the Farabi Eye Hospital as our “Single-Center data set”. The trained network then was evaluated using previously unseen data from four different data sets as “Multi-Center data set for external validation of classification algorithm”. A diagnosis was provided from medical records by three neuro-ophthalmologists (M.A.F, N.R.M, P.S.S) and two glaucoma specialists (M.M, Y.S). Fundus photographs were classified into three groups: normal optic disc, GON, and NGON. GON was defined as an enlarged vertical cup-to-disc ratio, diffuse or focal thinning of the neuroretinal rim, a pattern standard deviation on automated perimetry using a Humphrey Field Analyzer (Carl Zeiss Meditec AG, Jena, Germany) outside 95% normal limits (confirmed on at least two consecutive, reliable tests), a glaucoma hemifield test outside normal limits, and a with history of an intraocular pressure reading of >20 mm Hg on more than one visit.[18] All severities of GON, from mild to severe, were included. NGON were defined as evidence of optic nerve dysfunction associated with optic disc pallor (with or without cupping) with intraocular pressures of <20 mm Hg and caused by an identifiable specific pathogenetic process, such as inflammation (ie, optic neuritis), ischemia, compression, or a genetic mutation known to cause an NGON (eg, Leber hereditary optic neuropathy, dominant optic atrophy). Clinically normal optic discs were defined as those from patients with no clinical evidence of optic nerve or retinal dysfunction (ie, normal visual acuity, color vision, and visual fields), intraocular pressures <20 mmHg, and no clinical evidence of optic disc pallor. Normal discs were defined as those from patients without any clinical evidence of optic nerve dysfunction, regardless of the size of their cups; we did not perform optical coherence tomography

Single-Center data collection for training and validation: To design and validate a learning-based supervised system, we initially used a data set of 1,916 digital color ocular fundus photographs of GON (n=682), NGON (n=716), and normal discs (N=518) collected from Farabi Eye Hospital between 2011 and 2020. From these images, 94 were excluded due to low contrast and significant artifacts in the images. The final data set for training and validation consisted of 1,822 digital color ocular fundus photographs of GON (n eyes= 676, n subjects=190), NGON (n eyes=660, n subjects=112), and normal optic discs (n eyes=486, n subjects=214). Examples from each group and different artifacts leading to exclusion of images are demonstrated in Figure 1A. Briefly, images with clear and identifiable structures despite minor low-quality artifacts (e.g., blur, insufficient illumination, shadows) were categorized as “usable”. Images with serious quality issues that could not be reliably diagnosed by an ophthalmologist were categorized as “reject”. No data were removed due to the proximity of the optic disc to the image border, and images were not excluded solely because of low image quality.

Multi-Center data set for external validation: For extremal validation of the proposed model, we used the following data sets, the images from which had been obtained using different imaging devices and from different populations and ethnicities.

1. A publicly accessible data set [external data set 1] that included fundus images from 38 healthy, 13 GON, and 12 NGON eyes of 63 subjects.[19]

2. A glaucoma data set [external data set 2] from Ramathibodi Hospital (Mahidol University, Bangkok) that included 100 mild-to-severe GON fundus cases of 100 subjects. Ninety-nine images were used from this set.

3. A data set of 159 NGON eyes from the Wilmer Eye Institute [external data set 3], from which 123 eye images of 97 subjects were used, the rest not meeting the quality requirement for the segmentation stage.

4. A data set of 96 eye images from the University of Colorado [external data set 4], from which eight healthy, 40 GON, and 28 NGON eyes of 78 subjects were enrolled in the analysis.

One senior glaucoma specialist (W.S) who was not involved in the original analysis reviewed the GON and NGON fundus photographs from all four external-testing data sets and his prediction was compared with the findings using AI.

Data set for training segmentation algorithm: To prepare a robust algorithm to detect the location of the optic disc as the first step of our algorithm, we used a number of publicly accessible data sets, including 52 images form RIM-ONE r1,[20] 159 from RIM-ONE r3,[21] 277 from MESSIDOR, and 165 from Bin Rushed.[22]

Data split, augmentation and implementation: We used K-Fold cross-validation to ensure that every observation from the original data set had a chance of appearing in the training and test set. For classification, we applied the K-fold in a subject-wise manner. Specifically, we considered the possibility that multiple images might be acquired from a single subject. It was important that all images of an individual be only in the training or test data sets. We employed data augmentation to enable better generalization and prevent overfitting.[23] Data augmentation techniques used in this study included rotation (15-degree range), zoom (0.05-degree range), width shift (0.2-degree range), height shift (0.2-degree range), and shear (0.05-degree range). After augmentation, we resized all images to the default input size of 224×224 pixels and performed 5-fold cross-validation to avoid biased selection of test images. We did not perform any manual annotation or additional image preprocessing (Figure 1-B). We performed implementation using TensorFlow (https://www.tensorflow.org/) 2.8.2. Our models were trained and validated using an NVIDIA Tesla T4 GPU with 15GB memory. All source codes used to develop the optimal model are publicly available on GitHub ((https://github.com/mahsavali/NGOP-GON-Classification).

Optic disc segmentation (OD-SEG) network: The OD-SEG network was designed to crop the original fundus images to make a cropped image that contained the informative ROI around the optic disc.[24] Figure 1C shows a modified U-net[25] model that we used with an input image size of 256×256, initial channels of 16, and kernel size of 3×3. Using these 16 convolutional layers, we could recognize the location of the OD with an acceptable accuracy. The contracting, bottleneck and expansive paths of the U-Net consist of eight convolutional neural networks (CNN), two CNN with 0.2 dropout and six CNN blocks. We used dice coefficient loss for training. The model was trained and evaluated with 5-fold cross-validation on the “Data set for training segmentation algorithm”. To compensate for off-centered masks, we used a morphological mending strategy. In this strategy, the output of the OD-SEG is evaluated, and if the number of pixels in the mask is less than 200 (determined empirically), a morphological operation is applied to fit a circle to the segmented area and mend the miss-segmented OD (Figure supplement 1). We then evaluated its performance on unseen data on a limited number of manually segmented images from the “Single-Center data set”. After the localization step, we cropped a bounding box with a height of 3 times the optic disc radius, resized it to 224×224 pixels, and used it as input to the next classification network.

Classification network: As part of the proposed pipeline, we retrained and evaluated six architectures: VGG,[26] ResNet,[27] Inception,[28] MobileNet,[29] DenseNet[30] and Vision Transformer.[31] We pre-trained all models on ImageNet[32] and performed transfer learning on the smaller data set of medical images by adapting pre-trained features as initialization values and fine-tuning models using the new data set. To classify the fundus images into three classes, (healthy, GON, and NGON), we removed the last layer of each of the models and replaced it with a global average pooling layer with a softmax activation function. We used transfer learning to share the weight parameters (Figure 1D). We fine-tuned the transfer learning model by freezing the low layers but trained the top layers to update weight parameters. During the feature extractor and fine-tuning stages, we considered the number of frozen layers and the number of epochs as hyperparameters and changed them in accordance with the model's performance on the validation set to converge earlier and get higher recognition accuracy. In addition, we used ablation studies to measure the impact of each component of the overall system on the predictive performance of the pipeline by removing individual steps of the proposed framework. Finally, to provide visual explanations of image regions having a greater influence on the final predictions, we used gradient-weighted class activation mapping (Grad-CAM)[33].

Statistical analysis:

Evaluation of OD-SEG network: We used dice coefficient (F1 score for segmentation) as a measure of overlap between the ground truth and the predicted mask by the network. This measure yields a value ranging from 0.0 to 1.0, where the value of 1.0 denotes perfect overlap.

Evaluation of Classification network: To assess the performance of the classification network, we provided confusion matrices. In the counting matrix, the columns represent the numbers of samples in the true class, and the rows contain the predictions of the network. Each diagonal element gives the number of correctly classified images in the corresponding class (i.e., column) and off-diagonal elements represent the number of misclassified images. The overall accuracy was determined by dividing the total number of correctly classified images by the total number of images. The normalized values (by dividing the counts in each cell by the total number of samples in that class) were presented in the confusion matrix. Furthermore, we used the one-versus-rest strategy and reported the sensitivity, specificity, and F_1 score.

Results:

Performance of OD-SEG network on the data sets for training segmentation algorithm: An average dice value of 94.2% was achieved for 5-fold cross-validation on this data set.

Figure 2 (a) depicts examples of the utilized data sets (first row), ground truth masks (second row), and performance of the trained OD-SEG network (third row).

Performance of OD-SEG network on the Single-Center data set: To evaluate the performances of the trained OD-SEG, 128 images (42 normal eyes, 31 GON eyes and 55 NGON eyes) from the "Single-Center data set" were manually segmented by an ophthalmologist and used as ground truth. Figure 2-(b) shows examples of unseen data from the "Single-Center data set" (first row) and the ground truth masks (second row). It is evident that the performance of the channel-wise thresholding method [17] (third row) is low, due to less similarity of the masks to the ground truth. However, the masks predicted by the trained OD-SEG network (fourth row) show confirming results and masks very similar to the ground truth. This also is numerically proven where the average dice coefficient between the ground truth masks and the predicted masks by OD-SEG is 93.8%, compared with the channel-wise thresholding method that gives 65.2%.

Performance of the classification network on the Single-Center data set: Figure 3 demonstrates the confusion matrices for each of the six fine-tuned CNNs with 5-fold subject-wise cross-correlation. Table 1 is a numerical summary of sensitivity, specificity, precision, and F1-score for each network. DenseNet121 achieves the best performance with a sensitivity of 95.36%, precision of 95.35%, specificity of 92.19% and F1 score of 95.40%. This network thus was selected as the best-performing network for the reported external validation results.

Ablation study: Based on the Table supplement, we used the best network (DenseNet 121) in this stage and studied by ablation analysis the two main components of the proposed framework, component 1 (augmentation step) and component 2 (using OD-SEG for cropping the ROI). Table supplement shows the performance of the best network without both components (first row), by adding only component 1 (second row), by adding only component 2 (third row) and, finally, by adding both components (the last row). The higher values of the classification metrics (Precision, F1-score, Sensitivity, and Specificity) in the last row demonstrate the effectiveness of both components on the final results obtained by the classifier.

Performance of the classification network on the "Multi-Center data set for external validation of classification algorithm": Table 2 shows sensitivity, specificity, precision, and F1-score of the DenseNet121 on Multi-Center data set for external validation, and Figure 1 shows confusion matrices for external validation data sets in detecting GON, NGON, and healthy eyes.

A comparison of network performance against a glaucoma specialist in detecting GON and NGON from multicenter data for external validation: Of the 152 GON optic discs included in the Multi-Center data set, the network correctly diagnosed 130 (85.5 %), whereas a senior glaucoma specialist correctly diagnosed 108 (71.05 %), misclassifying 44 (28.94%) as consistent with NGON. Of 163 NGON discs, the network diagnosed 145 (88.94 %) correctly, whereas the glaucoma specialist diagnosed 134 (82.20 %) correctly, with 29 (17.79 %) being misclassified as GON discs by the specialist. The glaucoma specialist thus had an overall sensitivity of 71.05%

and a specificity of 82.21%, whereas the DL method had a sensitivity of 85.53% and a specificity of 89.02%. The sensitivity of the CNN network exceeded that of the senior specialist (P=0.002, chi-square). The specificity was not different between them (P= 0.08, chi-square).

Feature visualization

Figure supplement 2 shows gradient class activation heatmaps for the best-performing-network on six images (two each of normal, GON and NGON). It is clear from this figure that the network is classifying the images according to the optic disc area by contrasting the normal, GON, and NGON images. The region that has the greatest influence on the decisions is located on the optic disc.

Discussion:

The aim of this research was to assess the performance of a deep-learning system to differentiate GON from NGON using fundus images. In our study, DenseNet121 showed high performance with a sensitivity of 95.36% and specificity of 97.63% in the training data set. In the four external-testing data sets, the total precision of the network was 86.07%, which shows the generalizability of the method. In differentiating GON from NGON from multicenter data for external validation, the DL method had significantly better sensitivity (85.53%) than that of the glaucoma specialist (71.05%).

Despite the differences in the pathogeneses of GON and NGON, several studies have indicated that differentiating NGON from GON based on the appearance of the optic disc can be challenging.[3-8] Even though a major ophthalmoscopic hallmark of GON is optic disc cupping, NGON also can present with an increased cup-to-disc ratio. Trobe et al.[4] showed that GON and NGON were difficult to distinguish. Indeed, these investigators noted that of 29 eyes with NGON, 13 (44%) were misdiagnosed as having GON by at least one observer. Consistent with this observation is that in our study, a senior glaucoma specialist who reviewed the GON and NGON fundus photographs from all four data sets in a masked fashion misclassified 28.9% of GON and 17.8% of NGON discs. To address this issue, AI has been used to distinguish among different optic neuropathies. Specifically, DL models have been used to distinguish between patients with and without glaucoma.[12-16] Normal and abnormal discs (with atrophy, hypoplasia, papilledema, anterior ischemic optic neuropathy) also can be identified using DL.[34-36] Only one study has used DL for differentiating GON from NGOP. Yang et al.[17] achieved a sensitivity of 93.4% and specificity of 81.8% in distinguishing GON from NGON using ResNet-50; however, ROI selection and evaluation of the generalizability of this technique were not considered in that study. Furthermore, the number of GON and NGON cases was much lower than the healthy controls, thus producing a biased result with high accuracy for normal-appearing discs and, consequently, a high averaged accuracy. Finally, Yang et al. included only GON eyes with a cup-to-disc ratio >0.7 in their study and also did not use an external validation data set. Both factors could lead to poor generalization of the results to all glaucomatous eyes. For ROI and disc segmentation, instead of basic thresholding models, we used a modified U-net model in our study. We also included more GON (from mild to severe glaucoma) and NGON cases. In addition, in our study, we tested six different fine-tuned CNNs for classification with 5-fold subject-wise cross-correlation (Table 1). As seen in Table 1, the Vision transformer was not performing well. This is most likely due to the large number of parameters. It also cannot be fine-tuned on a hierarchical basis layer by layer and therefore requires a larger data set and longer training time. DenseNet121 and MobileNet have better performance compared with other models. The main element in the MobileNet is the depth-wise separable convolution. This has resulted in the MobileNet network having a higher convergence speed and acceptable accuracy compared with other networks despite MobileNet having 4.2 million parameters compared with the ResNet18 network that has 11 million and the VGG16 that has 14 million. DenseNet121 works best, likely due to its connections among different layers, thus providing better feature transfer with different levels of granularity. Accordingly, we used DenseNet121 for external validation testing. Finally, and most importantly, we also compared the network performance against the ability of a senior glaucoma specialist to distinguish between GON and NGON using the same fundus photographs and found that the sensitivity and specificity of the CNN were higher than those of the specialist.

Our study does have some limitations. Even though the total number of photographs in our external data set was much larger than that of Yang et al.,[17] it still was small. We plan to remedy this in future work. Also, although our network differentiated three classes (GON, NGON, and normal optic discs) in the four external-testing data sets, we asked the glaucoma specialist to differentiate between only GON and NGON in those data sets.

In summary, using DL, we showed a high performance in correctly identifying GON and NGON on color fundus photographs—higher than that of a senior glaucoma expert. The performance of our system in the external-testing data is reliable, demonstrating the generalization capabilities of the proposed framework. In particular, we believe that this model can assist general ophthalmologists, glaucoma specialists, and neuro-ophthalmologists in correctly differentiating GON from NGON, resulting in appropriate and timely management.

Acknowledgment:

Funding/Support: None

Financial Disclosures: No financial disclosures.

REFERENCES:

1. Mantravadi AV, Vadhar N. Glaucoma. Prim Care. 2015;42(3):437-449.

2. Lee AG, Chau FY, Golnik KC, et al. The diagnostic yield of the evaluation for isolated unexplained optic atrophy. Ophthalmology 2005;112(5):757-9.

3. Waisberg E, Micieli JA. Neuro-Ophthalmological Optic Nerve Cupping: An Overview. Eye and Brain 2021;13:255.

4. Trobe JD, Glaser JS, Cassady J, et al. Nonglaucomatous excavation of the optic disc. Archives of Ophthalmology 1980;98(6):1046-50.

5. Ambati BK, Rizzo III JF. Nonglaucomatous cupping of the optic disc. International ophthalmology clinics 2001;41(1):139-49.

6. Fraser CL, White AJ, Plant GT, Martin KR. Optic nerve cupping and the neuro-ophthalmologist. J Neuroophthalmol. 2013;33(4):377-389.

7. Hata M, Miyamoto K, Oishi A, et al. Comparison of optic disc morphology of optic nerve atrophy between compressive optic neuropathy and glaucomatous optic neuropathy. PLoS One 2014;9(11):e112403.

8. Fard MA, Moghimi S, Sahraian A, Ritch R. Optic nerve head cupping in glaucomatous and non-glaucomatous optic neuropathy. British Journal of Ophthalmology 2019;103(3):374-8.

9. Ramm L, Schwab B, Stodtmeister R, et al. Assessment of optic nerve head pallor in primary open-angle glaucoma patients and healthy subjects. Current eye research 2017;42(9):1313-8.

10. Leong Y-Y, Vasseneix C, Finkelstein MT, et al. Artificial intelligence meets neuro-ophthalmology. The Asia-Pacific Journal of Ophthalmology 2022;11(2):111-25.

11. Milea D, Najjar RP, Jiang Z, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. New England Journal of Medicine 2020;382(18):1687-95.

12. Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 2018;125(8):1199-206.

13. Wu J-H, Nishida T, Weinreb RN, Lin J-W. Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: A meta-analysis. American Journal of Ophthalmology 2021.

14. Chaurasia AK, Greatbatch CJ, Hewitt AW. Diagnostic Accuracy of Artificial Intelligence in Glaucoma Screening and Clinical Practice. Journal of Glaucoma 2022;31(5):285-99.

15. Nakahara K, Asaoka R, Tanito M, et al. Deep learning-assisted (automatic) diagnosis of glaucoma using a smartphone. British Journal of Ophthalmology 2022;106(4):587-92.

16. Al-Aswad LA, Kapoor R, Chu CK, et al. Evaluation of a deep learning system for identifying glaucomatous optic neuropathy based on color fundus photographs. Journal of glaucoma 2019;28(12):1029-34.

17. Yang HK, Kim YJ, Sung JY, et al. Efficacy for differentiating nonglaucomatous versus glaucomatous optic neuropathy using deep learning systems. American Journal of Ophthalmology 2020;216:140-6.

18. Hodapp E, Parrish RK, Anderson DR. Clinical decisions in glaucoma: Mosby Incorporated, 1993.

19. Cen L-P, Ji J, Lin J-W, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature communications 2021;12(1):1-13.

20. Fumero F, Alayón S, Sanchez JL, et al. RIM-ONE: An open retinal image database for optic nerve evaluation. 2011 24th international symposium on computer-based medical systems (CBMS): IEEE, 2011.

21. Fumero F, Sigut J, Alayón S, Silvia S, González-Hernández M, González de la Rosa M. Interactive Tool and Database for Optic Disc and Cup Segmentation of Stereo and Monocular Retinal Fundus Images. Václav Skala - UNION Agency, 2015.

22. Almazroa A, Alodhayb S, Osman E, et al. Retinal fundus images for glaucoma analysis: the RIGA dataset. Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications: SPIE, 2018; v. 10579.

23. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:171204621 2017.

24. Orlando JI, Prokofyeva E, del Fresno M, Blaschko MB. Convolutional neural network transfer for automated glaucoma identification. 12th international symposium on medical information processing and analysis: SPIE, 2017; v. 10160.

25. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention: Springer, 2015.

26. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014.

27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition2016.

28. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition2015.

29. Howard AG, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:170404861 2017.

30. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition2017.

31. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929 2020.

32. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition: Ieee, 2009.

33. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision2017.

34 Yang HK, Oh JE, Han SB, et al. Automatic computer-aided analysis of optic disc pallor in fundus photographs. Acta Ophthalmologica 2019;97(4):e519-e25.

35. Liu TA, Ting DS, Paul HY, et al. Deep learning and transfer learning for optic disc laterality detection: Implications for machine learning in neuro-ophthalmology. Journal of Neuro-Ophthalmology 2020;40(2):178-84.

36. Milea D, Singhal S, Najjar RP. Artificial intelligence for detection of optic disc abnormalities. Current Opinion in Neurology 2020;33(1):106-10.

Figure legends:

Figure 1: General view of the proposed method for automated Fundus image classification, (A): Examples from Helathy, GON, and NGON fundus images, (B): Steps of the proposed framework, (C): OD-SEG network, and (D): Transfer Learning architecture.

Figure 2*:* (a) Example fundus images from 4 different data sets used for training OD-SEG network (first row), ground truth masks (second row) and predicted masks by trained OD-SEG network (third row). Figure shows the acceptable similarity indicating good performance of the algorithm. (b) Examples from fundus images collected in the current study (first row), ground truth masks (second row), the performance of channel-wise thresholding method (third row), and the performance of the OD-SEG network (last row).

Figure 3: Confusion matrices (average result between 5- folds) for different models.

Figure 4: Confusion matrices for the external validation data sets.

Figure supplement 1: A flowchart indicates image quality assessment of the optic disc segmentation (OD-SEG) network.

Figure supplement 2: Features shown by the Grad-CAM algorithm to demonstrate that the correctly performance of network is functioning properly.