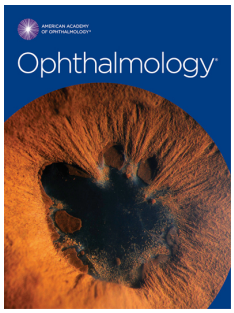


Explaining the Rationale of Deep Learning Glaucoma Decisions with Adversarial Examples

Jooyoung Chang, MD, Jinho Lee, MD, Ahnul Ha, MD, Young Soo Han, MD, Eunoo Bak, MD, Seulggie Choi, MD, Jae Moon Yun, MD, Uk Kang, PHD, Il Hyung Shin, PHD, Joo Young Shin, MD, Taehoon Ko, PHD, Ye Seul Bae, MD, Baek-Lok Oh, MD, Ki Ho Park, MD PHD, Sang Min Park, MD PHD

PII: S0161-6420(20)30579-0

DOI: https://doi.org/10.1016/j.ophtha.2020.06.036

Reference: OPHTHA 11328

To appear in: *Ophthalmology*

Received Date: 16 January 2020

Revised Date: 14 June 2020

Accepted Date: 15 June 2020

Please cite this article as: Chang J, Lee J, Ha A, Han YS, Bak E, Choi S, Yun JM, Kang U, Shin IH, Shin JY, Ko T, Bae YS, Oh B-L, Park KH, Park SM, Explaining the Rationale of Deep Learning Glaucoma Decisions with Adversarial Examples, *Ophthalmology* (2020), doi: https://doi.org/10.1016/j.ophtha.2020.06.036.

Explaining the Rationale of Deep Learning Glaucoma Decisions with Adversarial Examples

Authors: Jooyoung Chang, MD1; Jinho Lee, MD2,3; Ahnul Ha, MD 2,4; Young Soo Han, MD 2,4; Eunoo Bak, MD 2,4; Seulggie Choi, MD 1; Jae Moon Yun, MD 5; Uk Kang, PHD6,7; Il Hyung Shin, PHD 6; Joo Young Shin, MD8; Taehoon Ko, PHD9; Ye Seul Bae, MD 5,9 Baek-Lok Oh, MD 2,4; Ki Ho Park, MD PHD*2,4; Sang Min Park, MD PHD*1,5;

1. Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea

2. Department of Ophthalmology, Seoul National University College of Medicine, Seoul, Korea

3. Department of Ophthalmology, Hallym University Chuncheon Sacred Heart Hospital, Chuncheon, Korea

4. Department of Ophthalmology, Seoul National University Hospital, Seoul, South Korea

5. Department of Family Medicine, Seoul National University Hospital, Seoul, South Korea

6. InTheSmart Co., Ltd., Seoul, South Korea

7. Seoul National University Hospital Biomedical Research Institute, Seoul, South Korea

8. Department of Ophthalmology, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea

9. Office of Hospital Information, Seoul National University Hospital, Seoul, South Korea

* Co-correspondence:

Sang Min Park, Department of Family Medicine and Biomedical Sciences, College of Medicine, Seoul National University, 101 Daehak-ro, Jongno-gu, Seoul, South Korea

Tel.: 82-2-2072-3331

Fax: 82-2-766-3276

Email: smpark.snuh@gmail.com

Ki Ho Park, Department of Ophthalmology, Seoul National University College of Medicine, Seoul National University, 101 Daehak-ro, Jongno-gu, Seoul, South Korea

Tel: 82-2-760-2438,

Fax: 82-2-741-3187,

Email: kihopark@snu.ac.kr

Funding: InTheSmart, Co. Ltd (Grant Number: 0620180650) and Seoul National University Hospital Research Fund (Grant Number: 0320190160). 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.

Running head: Explaining the Rationale of Deep Learning Glaucoma Decisions

Word Count: 3,615 words

Abstract

Objective: To illustrate what is inside the "black box" of deep learning models (DLMs) so that clinicians can have greater confidence in the conclusions of artificial intelligence by evaluating adversarial explanation on its ability to explain the rationale of deep learning model decisions for glaucoma and glaucoma related findings. Adversarial explanation generates adversarial examples (AEs), or images which have been changed to gain or lose pathology-specific traits, to explain the DLM's rationale.

Design: Evaluation of explanation methods for DLMs.

Participants: Health screening participants (n=1,653) at the Seoul National University Hospital Health Promotion Center.

Methods: We evaluated 6,430 retinal fundus images for referable glaucoma (RG), increased cup-to-disc ratio (ICDR), disc rim narrowing (DRN), and retinal nerve fiber layer defect (RNFLD), and trained DLMs for each diagnosis and findings. Surveys consisting of explanations using AE and gradient-weighted class activation mapping (GradCAM), a conventional heatmap-based explanation method, were generated for 400 pathologic and normal patient-eyes. For each method, board-trained glaucoma specialists rated the location explainability, the ability to pinpoint decision-relevant areas in the image, and rationale explainability, the ability to inform the user on the model's reasoning for the decision based on pathological features. Scores were compared by paired Wilcoxon signed-rank test.

Main Outcome: Area under the receiver operating characteristic curve (AUC), sensitivities, and specificities of DLMs. Visualization of clinical pathology changes of AEs. Survey scores for locational and rationale explainability.

Results: The AUCs were 0.90, 0.99, 0.95, and 0.79, and sensitivities were 0.79, 1.00, 0.82, and 0.55 at 0.90 specificity, for RG, ICDR, DRN, and RNFLD DLMs, respectively. Generated AEs showed valid clinical feature changes, and survey results for location explainability was 3.94±1.33 and 2.55±1.24 using AEs and GradCAMs, respectively, out of a possible maximum score of 5 points. The score for rationale explainability was 3.97±1.31 and 2.10±1.25 for AE and GradCAM, respectively. AE provided better location and rationale explainability than GradCAM (p-value<0.001).

Conclusions: Adversarial explanation increased the explainability over GradCAM, a conventional heatmap-based explanation method. Adversarial explanations may help medical professionals more clearly understand the rationale of DLMs when using them for clinical decisions.

Introduction

Deep learning models (DLMs) can detect glaucoma using retinal fundus images. [1, 2] However, the "black-box" nature of DLMs[3] and the lack of valid explanations following DLM decisions limit the application and acceptance of these models in the medical field.

Currently, the explainability of DLMs using retinal fundus imaging is limited to highlighting pixel areas of importance in the form of heatmaps using gradient-based saliency techniques [4, 5] or occlusion-testing. [6] In the context of glaucoma-related findings, heatmaps can identify the location of an increased cup-to-disc ratio (ICDR) or disc rim narrowing (DRN) [7] but cannot explain the DLM's rationale, or the understanding of the pathology-specific feature.

Understanding DLM decisions is important for clinical decision support, [8] and the identification of important regions alone may be insufficient for understanding the model's rationale or for assisting clinical decisions in the real world. Highlights of important regions are implicit, and users must first assume that the model's highlight identifies the pathology by cause and not by coincidence nor by identification of other correlating features of the same region. Furthermore, determining the model's understanding using heatmaps alone assumes the user's ability to infer pathology from location, which may not always be the case with varying user expertise or clinical settings. Thus, location maps alone may fail to support clinical decisions by contributing no more than what the clinician can already identify. However, an explanation method which explicitly shows the model's understanding of pathology with examples of added or removed pathology could provide a basis for reasonable and educational interactions between the clinician, the patient, and artificial intelligence, and thus enhance interaction, experience, and decision-making.

Adversarial explanation is a method which uses the prediction DLM to make counterexamples as a means to explain the DLM's understanding of pathology-specific features.[9] For example, a valid adversarial explanation will consist of a "negative" adversarial example (AE) with decreased cup-to-disc ratio or a "positive" AE with increased cup-to-disc ratio given a retinal fundus image and a DLM for ICDR. Previously, AEs provided valid explanations for non-medical images,[9] however, whether they provide valid explanations for retinal fundus images and whether they significantly increase the explainability of a model over current methods have not been reported. Thus, the clinical validation of adversarial explanation as a method for explaining deep learning models for glaucoma and glaucoma-related findings is warranted.

Methods

Study Population and Image Data

Data at Health Promotion Center of Seoul National University Hospital (SNUH),[10] from 2005 to 2016, were used for this study. Patients were anonymized before analysis was performed, and the need for patient consent was waived by the institutional review board at SNUH (IRB#:H-1703-044-837). This work adhered to the tenets of the Declaration of Helsinki.

Annotation of Findings and Diagnosis

To minimize inter-observer variability, each image was inspected by three physicians from a team of nine board-trained ophthalmologists. Following ISGEO[11] guidelines and other well designed studies, [2, 12, 13] referable glaucoma was defined as having vertical cup-to-disc ratio ≥0.7, disc rim narrowing of ≤0.1 disc diameter, disc hemorrhage, or retinal nerve fiber layer (RNFL) defect. Respective findings for ICDR, DRN, and RNFL defect were also annotated. Supplementary Table 1 shows the details of the criteria of glaucoma diagnosis, which was classified into normal, glaucoma suspect (GS), and glaucoma, most likely (GM), based on the criteria of previous population studies.[2, 12, 13] Referable glaucoma was defined as having either GS or GM.

Fundus images were taken using a digital non-mydriatic fundus camera (CR-2; Canon Inc., Tokyo, Japan). Images were carefully focused on the posterior pole of the retina (macula-centered) at a 45° field of view using the built-in split-line focusing device. Image quality was assessed based on field definition. Adequate field definition was confirmed only when (1) the entire macula and optic disc were visible, and (2) both main temporal arcades were completely visible.[14] Retinal photographs of poor image clarity and poor field definition were regarded as non-gradable. If any reviewer deemed a photograph as non-gradable, it was excluded from the study.

Information on the age at visit, sex, diabetes medication, hypertension medication, and historical images were given for the most informed evaluation. We developed the interface program for annotation (Supplementary Figure 1), and identical, color-calibrated monitors were distributed for annotation.

Development of Deep Learning Models

Patients were randomized into training, validation, and testing sets. To ensure that each subset contained a similar proportion of pathologic images, patient-randomized sampling was done until the proportion of

pathologic images had at least the prevalence of the least prevalent finding, namely that of RNFL defect (~2.5%). The training set was used for model training, the validation set was used to determine when to end training, and the testing set was used for performance measurements and explanations. The resulting training, validation, and testing sets consisted of 4,867, 544, and 605 images, respectively.

Deep learning models were separately trained to predict referable glaucoma and each glaucoma related finding; ICDR, DRN, and RNFL defect. To resolve class-imbalance, we up-sampled pathologic images in the training set such that one in three images were pathologic. Images were resized to 224 by 224 pixels during training and testing.

A ResNet-50 was used as the backbone of our models, [15] which outputs two numbers representing pathologic or non-pathologic, which add up to 1. Training was done for 100 epochs, using transfer learning from a model trained on the ImageNet database, [16] saving the model with the best validation set cross-entropy loss or area under the receiver operating characteristic curve (AUC). During training, we used random crop, random vertical flip, MixUp, [17] and random horizontal flip. Batches comprised of half adversarial examples and half original images, otherwise known as "half-half" training. [18]

Explanation Using Adversarial Examples

We produced adversarial examples (AE) using adversarial explanation. [9] In their work, Woods and colleagues use backpropagation from a DLM to iteratively alter the noise, or the adversarial noise (AN), given to an image which maximizes the model's prediction for a predetermined category. They introduced $g_{explain+/-,}$ where the shape of the noise was optimized to reveal salient, perceptually different examples. Simply put, adversarial explanation is a method of "repainting" the image until the model regards it as either completely pathologic or completely normal. We denoted adversarial examples which were encouraged to be pathologic as positive adversarial examples (AE+), and those which were encouraged to be normal as negative adversarial examples (AE-). Adversarial examples were made by 100 iterations of backpropagation.

For a comparison with a conventional heatmap-based explanation method, we produced saliency heatmaps using gradient-weighted class activation mapping (GradCAM). [4] To directly compare GradCAM with AEs, saliency heatmaps were also generated for the adversarial explanations by taking the absolute value of the pixel values in the adversarial noise. This is akin to identifying the areas of the AEs where most change occurred.

Because GradCAM heatmaps are small, usually 7 by 7 pixels depending on the DLM, we used bicubic interpolation when overlaying on the original image.

Expert Survey of Deep Learning Explanation Methods

A survey was conducted by three board-trained ophthalmologists to compare the explainability of AEs and GradCAM. From the testing set, we sampled 400 patient eyes which consisted of 100 images each for referable glaucoma, ICDR, DRN, and RNFL defect models. Due to the small prevalence of pathologic images, all pathologic images in the testing set were included in the survey and normal images were randomly sampled from the testing set. For each image, we produced an explanation using adversarial examples and an explanation using GradCAM (Supplementary Figure 2). For each explanation, two questions were asked: “Q1. How well does this explanation method explain the location of the normal or pathologic finding?” and "Q2. How well does this explanation method explain the rationale of the normal or pathologic finding?" Each question was given a score between 1 and 5. Because one adversarial example explanation and one GradCAM explanation was shown sequentially for the same image, the difference in scores was compared in pairs to determine whether one explanation produced better explainability than the other with respect to location or rationale.

Statistical Methods

The associations between findings and diagnoses was determined by $\chi 2$-test. Fleiss Kappa for three raters was used to calculate interobserver agreement. Survey results were analyzed using paired Wilcoxon signed-rank test, which tests the significance of the difference of scores without the need for any assumption in the distribution of scores. Alpha values of less than 0.05 was considered statistically significant. For deep learning, we used Fastai with Pytorch backend on Python 3.6.

Results

From 6,430 images that were annotated 3-fold, 417 were excluded for poor image quality. Our final dataset consisted of 1,653 patients consisting of 3,091 distinct visits and 6,013 images. The prevalence of referable glaucoma, ICDR, DRN, and RNFL defect were 9.6%, 4.4%, 6.0%, and 2.6%, respectively, and interobserver agreement kappa were 0.461, 0.499, 0.367, and 0.188, respectively (Table 1). Overall, images consisted of 54.6% males and had a mean age of 55.7 ± 9.6 years at the time of visit (Table 1). Patient-randomized training, validation, and testing subsets consisted of at least 15 or more than 2.5% of each pathology, except for disc hemorrhage which was not used for training (Table 1). There was significant association between referable

glaucoma and pathologic findings (Supplementary Table 2). The referable glaucoma, ICDR, DRN, and RNFL defect models had AUC 0.899, 0.986, 0.950, and 0.790, respectively for the testing subset. When stratified by greater severity of glaucoma, namely most-likely glaucoma diagnosis, models for referable glaucoma, ICDR, DRN, and RNFL defect had AUC of 0.989, 0.991, 0.987, and 0.900, respectively. The receiver operating characteristics stratified by severity of glaucoma is shown in Figure 1. Sensitivities were 0.79, 1.00, 0.82, and 0.55 for referable glaucoma, ICDR, DRN, and RNFL defect, respectively, at the pre-set specificity of 0.90.

Adversarial Examples

Figure 2 shows adversarial explanations and GradCAM heatmaps for referable glaucoma, RNFL defect, ICDR, and DRN DLM decisions for selected images. Figure 2a depicts adversarial explanations and GradCAM heatmaps for the referable glaucoma DLM on a positive sample. The GradCAM+ shows that referable glaucoma is due to the disc area and GradCAM- highlights non-disc regions. $AE_{glaucoma-}$ has decreased the cup-to-disc ratio and widened the superiotemporal disc rim, which the DLM predicts as referable glaucoma negative. $AE_{glaucoma+}$, which results in a higher prediction for glaucoma, increases cup-to-disc ratio and narrows the superiotemporal disc rim. Figure 2b depicts explanations for the referable glaucoma DLM on a negative sample. GradCAM- highlights the fovea region and parts outside the field-of-view (FOV), and GradCAM+ highlights several non-FOV areas. $AE_{glaucoma+}$ adds an increased cup-to-disc ratio finding, which the original image lacks. Figure 2c depicts a sample with RNFLD. GradCAM+ highlights regions outside the FOV and GradCAM- highlights the region inferior to the disc. AE_{RNFLD}- shows diminished RNFLD and AE_{RNFLD+} shows a darker defect. Adversarial noises show that the most changes have been made at the area of RNFLD. Figure 2d depicts a sample without RNFLD, for which GradCAM highlights regions outside the FOV. AE_{RNFLD}- does not make apparant changes to the original image, whereas AE_{RNFLD}+ shows darkened nerve fiber regions temporally and supriotemporally to the disc. Figures 2e and 2f depicts the ICDR positive sample and ICDR negative samples, respectively. GradCAM correctly highlights disc region. Relevant changes in cup-to-disc ratio are shown explicitly in $AE_{ICDR+,}$ which increases cup-to-disc ratio, and AE_{ICDR-}, which decreases cup-to-disc ratio. Figure 2g depicts a positive DRN image. The location of the disc is correctly identified in the GradCAM+ for this DRN positive sample. The widening and narrowing changes to superiotemporal disc rim width can be seen in the AE_{DRN-} and AE_{DRN+}, respectively. Figure 2h depicts a negative DRN image. The GradCAM heatmaps fails to locate the disc. However, the AE_{DRN+} shows a superiotemporal disc rim narrowing, which the original lacks.

A zoomed-in and annotated adversarial explanations are shown in Figure 3. Adversarial examples shows valid clinical feature changes including changes to peri-papillary atrophy, retinal nerve fiber layer defect, cup-to-disc ratio, and disc rim narrowing. Figure 3a shows an example with ICDR, for which $AE_{glaucoma}$- has removed peripapillary atrophy (white arrows) and reduced cup size (black arrow), whereas $AE_{glaucoma}$+ has exaggerated the peripapillary border (white arrows) and increased cup size (black arrow). Figure 3b shows an example with only RNFL defect, for which $AE_{glaucoma}$- has replaced the RNFL defect with disconnected vessel-like patterns (white arrows), and $AE_{glaucoma}$+ darkens and widens the RNFL defect (black arrows). Figure 3c shows an example with ICDR and DRN, for which AE_{ICDR}- has reduced cup size (white arrows), and AE_{ICDR}+ has increased cup size (black arrows) and added bright peaks near cup margins (black arrows). Figure 3d shows an example with DRN, for which AE_{DRN}- increases the superiotemporal rim width (black arrows), and AE_{DRN}+ makes the disc rim narrower. Figure 3c and Figure 3d appear defocused because they are zoomed-in sections of images of size of 224 by 224 pixels, a size requirement for input into the DLM. Figure 3e shows an example of RNFL defect, for which AE_{RNFLD}- fills the RNFL defect with discontinuous fiber-like patterns (green arrows), and AE_{RNFLD}+ produces a defect with darker contrast (green arrow) and thins the bridging vessel (red arrow).

Survey Results

Table 2 shows the survey results for location explainability (Q1) and Table 3 shows the survey results for rationale explainability (Q2). The average score for location explainability was 3.94±1.33 and 2.55±1.24 for AE and GradCAM, respectively, out of a possible maximum score of 5 points. When rating location explainability, raters gave on average 1.39 points higher for AE than GradCAM (p-value<0.001). The average score for rationale explainability was 3.97±1.31 and 2.10±1.25 for AE and GradCAM, respectively. When rating rationale explainability, raters gave on average 1.87 points higher for AE than GradCAM (p-value<0.001). AE received higher scores for both location and rationale explanations among normal images (1.59±1.56 and 2.03±1.60), more so than among pathologic images (1.00±1.45 and 1.58±1.74). Survey scores for referable glaucoma, DRN, and ICDR resulted in favor of AE, with 1.13±1.35, 1.91±1.38, and 2.19±1.23 points difference for location explainability and 1.98±1.52, 2.20±1.42, and 2.69±1.31 points difference for rationale explainability. However, absolute grade scores were low for RNFL defect at below 3 points and resulted in minimal explainability differences between AE and GradCAM. For location explainability, GradCAM scored marginally higher than AE among RNFLD positive images.

Discussion

To date, the implementation of explainable DLMs on glaucoma prediction using retinal fundus images using adversarial explanations has not been reported. Explainability in deep learning-based predictions have largely been focused on identifying the region of importance, [4, 5] using, for example, saliency maps. [7] This can explain where specific pathologies exist but not provide an explanation of the rationale for its decision. Adversarial explanations can provide model-derived positive- and negative-examples as a means to explain the model's rationale. [9] Here, we showed that explanations using adversarial examples may enhance the location and rationale explainability of a model over traditional saliency techniques such as GradCAM. Furthermore, an analysis of select adversarial examples shows that adversarial explanation can add or remove referable glaucoma and glaucoma-related findings to explain DLM decisions.

In this study, trained DLMs achieved similar AUCs to previous studies, [1, 2, 7, 19-23] however comparisons should consider the differences in severity of disease from differing study populations. Generally, clinic-based datasets have higher prevalence of and greater severity in cases compared to a screening-based dataset, and the difference in severity and population is known to affect AUC results. For example, a previous study reported an AUC of 0.945 for referable glaucoma in a clinic-based validation set but 0.855 in an external screening-based validation set. [1] One study showed that the AUC was 0.89 for mild cases but 0.97 for moderate to severe cases of glaucoma. [20] Most studies use clinic-based datasets to report AUCs for referable glaucoma, which range from 0.89 to 0.986 [2, 19-22]. These studies vary in the way data is collected and may not represent "real-world" populations due to variable inclusion and exclusion criteria of images. In community-based testing samples, previous work reported AUCs for referable glaucoma between 0.855 [1] and 0.942. [23] Few studies developed and validated models for specific glaucoma-related findings. [1, 7] A study using a clinic-based dataset reported AUCs of 0.922, 0.946, and 0.778 for ICDR, DRN, and RNFLD, [1] respectively, and a study using a screening-based dataset reported AUCs of 0.982 and 0.983 for glaucomatous disc change and RNFLD, respectively. [7] Our study, despite being based in a screening setting having lower prevalence and severity of disease, reported an AUC of 0.989 for referable glaucoma, and comparable AUCs for glaucoma-related findings. When stratified by disease severity, the AUC for our models improved when detecting findings among most-likely glaucoma patients.

There were large differences in the visualizations of what the model used by using adversarial explanation and GradCAM strategies. These differences occurred due to how the explanations were made. With GradCAM+, areas of the original image that contributed to a positive prediction were highlighted, whereas with AE+, the

example with added pathology was shown along with an AN+ heatmap of areas with the most change. Thus, GradCAM+ did not highlight anything at all if the image was totally normal, as seen in Figure 2b and Figure 2d, whereas AN+ highlighted the area that changed most in AE+.

Survey results showed that adversarial explanations provided better location and rationale explainability over GradCAM. The reasons for these improvements are multifactorial. First, GradCAM had no fundamental logic for explaining the rationale of the model's decision. The information contained in a saliency map is purely locational, highlighting the regions contributing towards or against a decision. On the contrary, AE provided a logic for explaining the rationale behind the model's decision, namely, an example with removed pathology (AE-) and an example with added pathology (AE+). Second, GradCAM explanations performed poorly when given normal images. Gradient-based saliency techniques have had trouble interpreting heatmaps for normal findings.[24, 25] Similarly, in our study, the survey results showed lower explainability for GradCAM among normal images. It is probable that for normal cases, the GradCAM+ heatmaps did not correctly highlight the region of interest because the image features did not make any contributions toward a pathologic prediction. AEs however performed well for both pathologic and normal images because they were able to add pathology-specific features to normal images to give a more definitive location.

Expert survey showed successful explanations for referable glaucoma, ICDR, and DRN. The validation of adversarial explanation has potential applications such as deep learning-based medical devices with greater explainability, medical education with generated AEs, and predictive visualization of glaucoma progression.

Our study includes several strengths. First, our work validates a deep learning explanation method which has not been applied and validated for retinal fundus images and provides a guide for validating explanation methods for other medical imaging modalities. Second, our dataset was annotated prospectively for deep learning purposes using a specific criterion for referable glaucoma, self-developed annotation software, and unified color-calibrated monitors, which minimizes misclassification bias. Third, our survey assessed two different aspects of explainablity, namely location and rationale, for which previous studies have not addressed separately. [8]

Limitations

Our data has limited diversity compared to other work[1, 2] and did not validate on external datasets, which may limit the generalizability of our DLMs. However, the primary objective of our work was to validate the

explainability of an explanation technique rather than the diagnostic performance of a DLM. While our model may behave differently on external datasets, the adversarial explanation technique can be generalized and be used for DLMs developed elsewhere.

Our model for RNFL defect did not produce a high AUC (0.790) and resulted in only a small increase in explainability by adversarial explanations. Previous studies report similar[1] or higher[7] AUC, and we think that RNFL defect did not perform as well as other glaucoma parameters for the following reasons. First, the number of RNFL defect training examples were few (122 images) compared to other studies (1168 images [7]) It is widely believed that a large dataset is crucial in medical deep learning,[26] and the number of RNFL defect images may have been too few for the DLM to be successfully trained. Second, the characteristics of the existing RNFL defect images of our database may have been subtle and ambiguous because the study population was based on a health screening setting and not a glaucoma clinic setting. The screening setting represents relatively earlier stages of glaucoma in which RNFL defects are smaller or harder to detect. This is evident in our analysis of interobserver agreement, where RNFL defect had lower agreement (0.19) compared to other glaucoma parameters (>0.35). Third, the RNFL defect model performed poorly when detecting less severe cases. When tasked with detecting RNFL defect without less severe, suspected glaucoma patients, the AUC for RNFL defect increased to 0.900, whereas excluding more severe, most-likely glaucoma patients, the AUC dropped to 0.705. Because our population consists of health screening participants, the existence of less severe cases resulted in lower performance. Finally, the lower performance of the RNFL defect model likely resulted in lower explainability. Compared to all other models, RNFL defect had low explainability scores for both AE and GradCAM. While the differences between the two methods were insignificant or marginally better for GradCAM in RNFLD+ subgroups, the low absolute scores indicate the inability of the model to consistently produce accurate and valid explanations rather than the outperformance of one method over the other. This warrants further work using more robust and accurate DLMs for RNFLD.

In conclusion, the expert survey showed significant improvements in location and rationale explainability of DLMs when using AE compared to GradCAM, and AE produced pathology-specific feature changes to explain deep learning-based decisions for referable glaucoma and glaucoma-related findings. Adversarial explanation may help medical professionals to better understand DLM-based decisions.

Funding

Our work was funded by InTheSmart, Co. Ltd (Grant Number: 0620180650) and Seoul National University

Hospital Research Fund (Grant Number: 0320190160). All authors declare no conflict of interest. Funding sources were not involved with any aspect pertinent to the study. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Acknowledgement

Author Contributions: A Ha, BL Oh, E Bak, J Chang, J Lee, JM Yun, JY Shin, KH Park, S Choi, SM Park, and Y S Han contributed to analysis and interpretation of data. IH Shin, J Chang, SM Park, T Ko, U Kang, and Y S Bae contributed to study concept and design. A Ha, BL Oh, E Bak, IH Shin, J Chang, J Lee, JM Yun, JY Shin, KH Park, S Choi, SM Park, T Ko, U Kang, Y S Bae, and Y S Han contributed to critical revision of the manuscript. J Chang, J Lee, and SM Park contributed to drafting of the manuscript. J Chang contributed to statistical analysis. A Ha, E Bak, IH Shin, J Lee, JM Yun, JY Shin, KH Park, S Choi, SM Park, T Ko, U Kang, Y S Bae, and Y S Han contributed to administrative, technical, or material support.

Declaration of Interest

All authors declare no conflict of interest.

References

1. Phene S, Dunn RC, Hammel N, et al. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. Ophthalmology 2019;126(12):1627-39.
2. 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.
3. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology 2019;103(2):167-75.
4. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision 2019.
5. Tulio Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv e-prints 2016:arXiv:1602.04938.
6. Mitani A, Huang A, Venugopalan S, et al. Detection of anaemia from retinal fundus images via deep learning. Nature Biomedical Engineering 2020;4(1):18-27.
7. Son J, Shin JY, Kim HD, et al. Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images. Ophthalmology 2020;127(1):85-94.
8. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology 2019;103(2):167.
9. Woods W, Chen J, Teuscher C. Adversarial explanations for understanding image classification decisions and improved neural network robustness. Nature Machine Intelligence 2019;1(11):508-16.
10. Yoon C, Goh E, Park SM, Cho B. Effects of smoking cessation and weight gain on cardiovascular disease risk factors in Asian male population. Atherosclerosis 2010;208(1):275-9.
11. Foster PJ, Buhrmann R, Quigley HA, Johnson GJ. The definition and classification of glaucoma in prevalence surveys. The British journal of ophthalmology 2002;86(2):238-42.
12. Iwase A, Suzuki Y, Araie M, et al. The prevalence of primary open-angle glaucoma in Japanese: The Tajimi Study. Ophthalmology 2004;111(9):1641-8.
13. Foster PJ, Buhrmann R, Quigley HA, Johnson GJ. The definition and classification of glaucoma in prevalence surveys. Br J Ophthalmol 2002;86(2):238-42.
14. Fleming AD, Philip S, Goatman KA, et al. Automated assessment of diabetic retinal image quality based on clarity and field definition. Invest Ophthalmol Vis Sci 2006;47(3):1120-5.

15. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv e-prints2015.

16. Ting DS, Liu Y, Burlina P, et al. AI for medical imaging goes deep. Nature medicine 2018;24(5):539.

17. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. arXiv e-prints2017.

18. Tsipras D, Santurkar S, Engstrom L, et al. Robustness May Be at Odds with Accuracy. arXiv e-prints 2018:arXiv:1805.12152.

19. Liu S, Graham SL, Schulz A, et al. A Deep Learning-Based Algorithm Identifies Glaucomatous Discs Using Monoscopic Fundus Photographs. Ophthalmology Glaucoma 2018;1(1):15-22.

20. Christopher M, Belghith A, Bowd C, et al. Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. Scientific Reports 2018;8(1):16685.

21. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. Scientific Reports 2018;8(1):14665.

22. Asaoka R, Tanito M, Shibata N, et al. Validation of a Deep Learning Model to Screen for Glaucoma Using Images from Different Fundus Cameras and Data Augmentation. Ophthalmology Glaucoma 2019;2(4):224-31.

23. Ting DSW, Cheung CY-L, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 2017;318(22):2211-23.

24. Lam C, Yu C, Huang L, Rubin D. Retinal lesion detection with deep learning using image patches. Investigative ophthalmology & visual science 2018;59(1):590-6.

25. Quellec G, Charrière K, Boudi Y, et al. Deep image mining for diabetic retinopathy screening. Medical image analysis 2017;39:178-93.

26. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. Radiographics 2017;37(7):2113-31.

Figure Legends

Figure 1. Receiver Operating Characteristic Curves for Referable Glaucoma, ICDR, DRN, and RNFLD stratified by Severity of Glaucoma. Abbreviations; AUC, area under curve; ICDR, increased cup-to-disc ratio; DRN, disc rim narrowing; RNFLD, retinal nerve fiber layer defect.

Figure 2. Adversarial Explanations of Deep Learning Models for Referable Glaucoma, RNFL Defect, Increased Cup-to-Disc Ratio, and Disc Rim Thinning. Abbreviations; Pred, prediction; AE, adversarial example; AN, adversarial noise; RNFLD, retinal nerve fiber layer defect; ICDR, increased cup-to-disc ratio; C/D, cup-to-disc; DRN, disc rim narrowing; FOV, field of view; All prediction values indicate probability for positive pathology. First row shows the original image and GradCAM for positive and negative pathology. The second row is the negative adversarial example with respective adversarial noise. The third row is the positive adversarial example with respective adversarial noise. Adversarial examples were produced using only the predictive deep learning model and the original sample image. Adversarial noise is shown superimposed on the original image.

Figure 3. Examples of Zoomed-In Adversarial Explanations for Deep Learning Models for Referable Glaucoma, Increased Cup-to-Disc Ratio, Disc Rim Narrowing, and RNFL Defect. Abbreviations; AE, adversarial example; ICDR, increased cup-to-disc ratio; DRN, disc rim narrowing; RNFLD, retinal nerve fiber layer defect. The first column is the original image, the second column is the negative adversarial example, and the third column is the positive adversarial example. Figure 3a shows an example with ICDR, for which $AE_{glaucoma}$- has removed peripapillary atrophy (white arrows) and reduced cup size (black arrow), whereas $AE_{glaucoma}$+ has exaggerated the peripapillary atrophy (white arrows) and increased cup size (black arrow). Figure 3b shows an example with only RNFL defect, for which $AE_{glaucoma}$- has replaced the RNFL defect with disconnected vessel-like patterns (white arrows), and $AE_{glaucoma}$+ darkens and widens the RNFL defect (black arrows). Figure 3c shows an example with ICDR and DRN, for which AE_{ICDR}- has reduced cup size (white arrows), and AE_{ICDR}+ has increased cup size (black arrows) and added bright peaks near cup margins (black arrows). Figure 3d shows an example with DRN, for which AE_{DRN}- increases the superiotemporal rim width (black arrows), and AE_{DRN}+ makes the disc rim narrower. Figure 3c and Figure 3d appear defocused because they are zoomed-in sections of images of size of 224 by 224 pixels, a size requirement for input into the deep learning model. Figure 3e shows an example of RNFL defect, for which AE_{RNFLD}- fills the RNFL defect with discontinuous fiber-like patterns (green arrows), and AE_{RNFLD}+ produces a defect with darker contrast (green arrow) and thins the bridging vessel (red arrow).

Tables

Table 1. Sample characteristics, associations between findings and diagnoses, interobserver agreement, and deep learning model results.

N (%*), or otherwise	Total	Pathology				Randomization Subsets		
		Referable Glaucoma†	Increased C/D ratio†	Disc rim narrowing†	RNFL defect†	Training Set	Validation Set	Testing Set
Data								
Number of Images	6013 (100)	579 (9.6)	267 (4.4)	358 (6)	157 (2.6)	4864 (80.9)	544 (9)	605 (10.1)
Number of Patient-visits	3091 (100)	402 (13)	171 (5.5)	253 (8.2)	132 (4.3)	2504 (81)	276 (8.9)	311 (10.1)
Number of Patients	1653 (100)	223 (13.5)	82 (5)	138 (8.3)	76 (4.6)	1322 (80)	165 (10)	166 (10)
Characteristics								
Age, years (mean, std)	55.7 (9.59)	55.8 (9.7)	56.1 (9.0)	56.2 (10.0)	56.0 (10.1)	55.76 (9.60)	54.23 (9.36)	56.18 (9.60)
Sex, male	3,284 (54.6)	374 (64.6)	196 (73.4)	240 (67.0)	102 (65.0)	2657 (54.6)	312 (57.3)	315 (52.1)
Referable Glaucoma†	579 (9.6)	579 (100)	265 (99.3)	354 (98.9)	153 (97.5)	468 (9.6)	53 (9.7)	58 (9.6)
Glaucoma suspect	472 (7.8)	472 (81.5)	187 (70.0)	255 (71.2)	72 (45.9)	388 (8.0)	43 (7.9)	41 (6.8)
Glaucoma most-likely	107 (1.8)	107 (18.5)	78 (29.2)	99 (27.7)	81 (51.6)	80 (1.6)	10 (1.8)	17 (2.8)
Increased C/D ratio†	267 (4.4)	265 (45.8)	267 (100)	179 (50)	70 (44.6)	214 (4.4)	28 (5.1)	25 (4.1)
Disc rim narrowing†	358 (6.0)	354 (61.1)	179 (67)	358 (100)	102 (65)	288 (5.9)	36 (6.6)	34 (5.6)
RNFL defect†	157 (2.6)	153 (26.4)	70 (26.2)	102 (28.5)	157 (100)	122 (2.5)	15 (2.8)	20 (3.3)
Interobserver Agreement								
Kappa statistic‡		0.461	0.499	0.367	0.188			
Model Performance among Testing Set								
AUC, units (95% CI)		0.899 (0.870-0.928)	0.986 (0.977-0.995)	0.950 (0.930-0.970)	0.790 (0.712-0.868)			
AUC among no glaucoma and GS subgroup, units (95% CI)		0.862 (0.820-0.904)	0.985 (0.974-0.996)	0.925 (0.889-0.961)	0.705 (0.574-0.836)			
AUC among no glaucoma and GM subgroup, units (95% CI)		0.989 (0.981-0.997)	0.991 (0.983-0.999)	0.987 (0.978-0.996)	0.900 (0.838-0.962)			

Abbreviations: RNFL, retinal nerve fiber layer defect; C/D, cup-to-disc; AUC, area under receiver operating characteristic curve; CI, confidence interval; GS, glaucoma suspect; GM, glaucoma most-likely. *Percentage of data using proportion of subset among total set (rows) and percentage of characteristics using proportions of images within the total set or subsets (columns). † Significant associations ($p<0.05$) between all combinations of Glaucoma, Increased C/D ratio, RNFL defect, and Disc rim narrowing using $\chi 2$-test was observed and is shown in Supplemental Table 1. ‡Using Fleiss Kappa for 3 raters.

Table 2. Comparison of scores for "Q1. How well does this explanation method explain the location of the normal or pathologic finding?" of each explanation method.

Models	Subgroups	Images, N	Total Surveys, N	Score for Adversarial Explanation, Mean±Std	Score for GradCAM, Mean±Std	Paired Score Difference, Mean±Std	p-value*
All Models (R. Glaucoma, DRN, ICDR, RNFLD)							
	All	400	1200	3.94±1.33	2.55±1.24	1.39±1.55	<0.001
	Pathologic	137	411	4.14±1.25	3.14±1.20	1.00±1.45	<0.001
	Normal	263	789	3.83±1.36	2.24±1.14	1.59±1.56	<0.001
Referable glaucoma							
	All	100	300	3.99±0.93	2.86±1.17	1.13±1.35	<0.001
	R.Glaucoma+	58	174	4.17±0.88	2.91±1.22	1.25±1.40	<0.001
	R.Glaucoma-	42	126	3.75±0.95	2.79±1.11	0.96±1.26	<0.001
DRN							
	All	100	300	4.76±0.56	2.85±1.34	1.91±1.38	<0.001
	DRN+	34	102	4.87±0.41	3.77±1.05	1.10±1.16	<0.001
	DRN-	66	198	4.70±0.61	2.37±1.21	2.33±1.29	<0.001
ICDR							
	All	100	300	4.87±0.47	2.68±1.09	2.19±1.23	<0.001
	ICDR+	25	75	4.84±0.59	3.41±0.87	1.43±1.14	<0.001
	ICDR-	75	225	4.88±0.42	2.44±1.04	2.45±1.16	<0.001
RNFLD							
	All	100	300	2.12±0.92	1.80±1.03	0.32±1.51	<0.001
	RNFLD+	20	60	1.93±1.16	2.37±1.15	-0.43±1.58	0.049
	RNFLD-	80	240	2.17±0.84	1.66±0.95	0.50±1.44	<0.001

Abbreviations; N, number; R. Glaucoma, referable glaucoma; DRN, disc rim narrowing; ICDR, increased cup-to-disc ratio; RNFLD, retinal nerve fiber layer defect. * For Wilcoxon Signed-Rank Test

Table 3. Comparison of scores for " Q2. How well does this explanation method explain the rationale of the normal or pathologic finding?" of each explanation method.

Models	Subgroups	Images, N	Total Surveys, N	Score for Adversarial Explanation, Mean±Std	Score for GradCAM, Mean±Std	Paired Score Difference, Mean±Std	p-value*
All Models (Glaucoma, DRN, ICDR, RNFLD)							
	All	400	1200	3.97±1.31	2.10±1.25	1.87±1.66	<0.001
	Pathologic	137	411	4.19±1.24	2.61±1.39	1.58±1.74	<0.001
	Normal	263	789	3.86±1.33	1.84±1.08	2.03±1.60	<0.001
Referable glaucoma							
	All	100	300	4.20±0.94	2.22±1.25	1.98±1.52	<0.001
	R.Glaucoma+	58	174	4.35±0.89	2.33±1.31	2.02±1.58	<0.001
	R.Glaucoma-	42	126	4.00±0.96	2.06±1.15	1.94±1.44	<0.001
DRN							
	All	100	300	4.62±0.76	2.42±1.42	2.20±1.42	<0.001
	DRN+	34	102	4.72±0.67	3.10±1.53	1.62±1.46	<0.001
	DRN-	66	198	4.57±0.80	2.07±1.22	2.51±1.30	<0.001
ICDR							
	All	100	300	4.80±0.63	2.11±1.16	2.69±1.31	<0.001
	ICDR+	25	75	4.73±0.79	2.75±1.36	1.99±1.60	<0.001
	ICDR-	75	225	4.83±0.57	1.90±1.00	2.93±1.11	<0.001
RNFLD							
	All	100	300	2.27±0.97	1.66±1.01	0.61±1.63	<0.001
	RNFLD+	20	60	2.15±1.29	2.40±1.18	-0.25±1.63	0.213
	RNFLD-	80	240	2.30±0.88	1.47±0.87	0.83±1.56	<0.001

Abbreviations; N, number; R. Glaucoma, referable glaucoma; DRN, disc rim narrowing; ICDR, increased cup-to-disc ratio; RNFLD, retinal nerve fiber layer defect. * For Wilcoxon Signed-Rank Test

Figure 1. Receiver Operating Characteristic Curves for Referable Glaucoma, ICDR, DRN, and RNFLD stratified by Severity of Glaucoma.

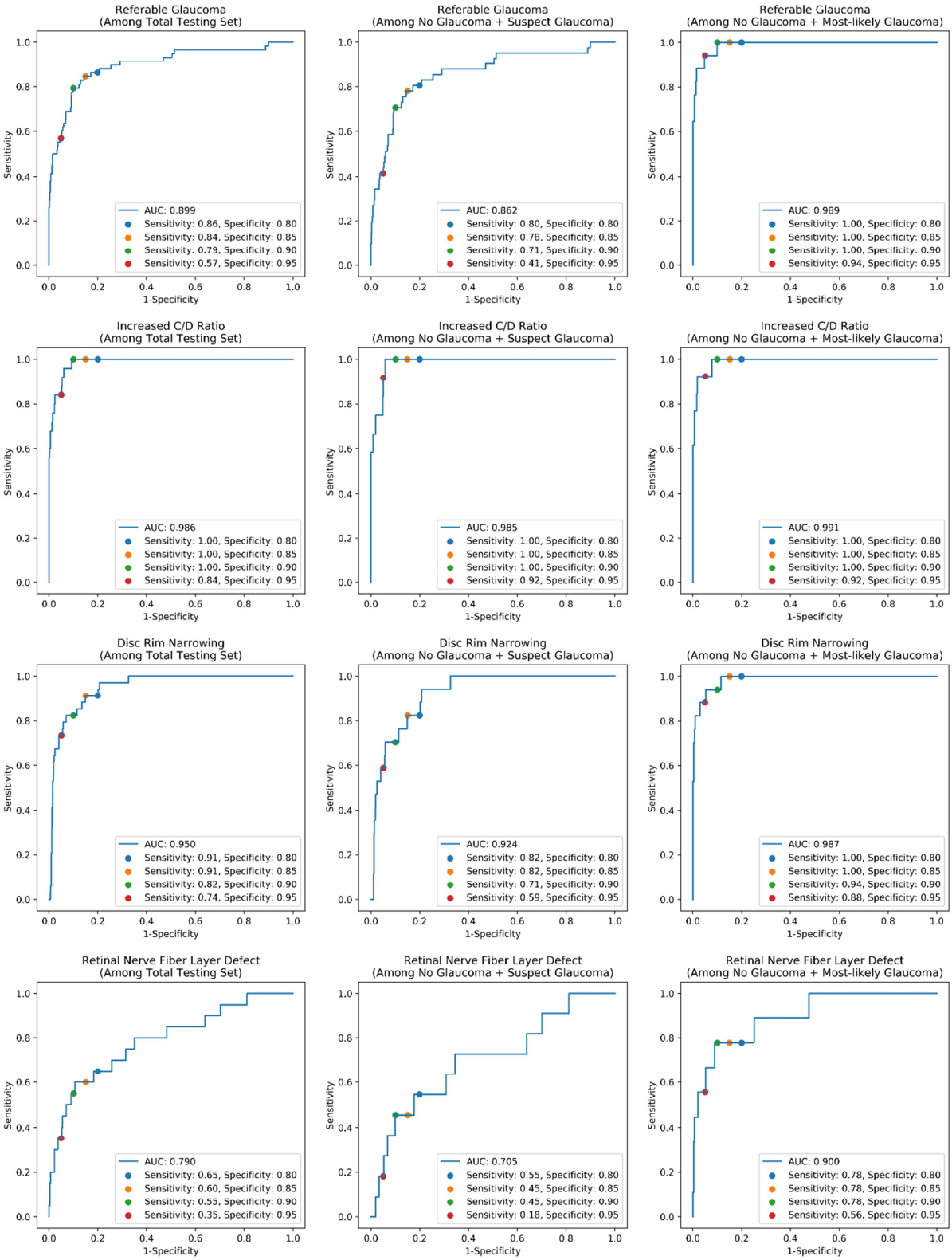


Figure 2. Adversarial Explanations of Deep Models for Referable Glaucoma, RNFL Defect, Increased Cup-to-disc Ratio, and Disc Rim Thinning.

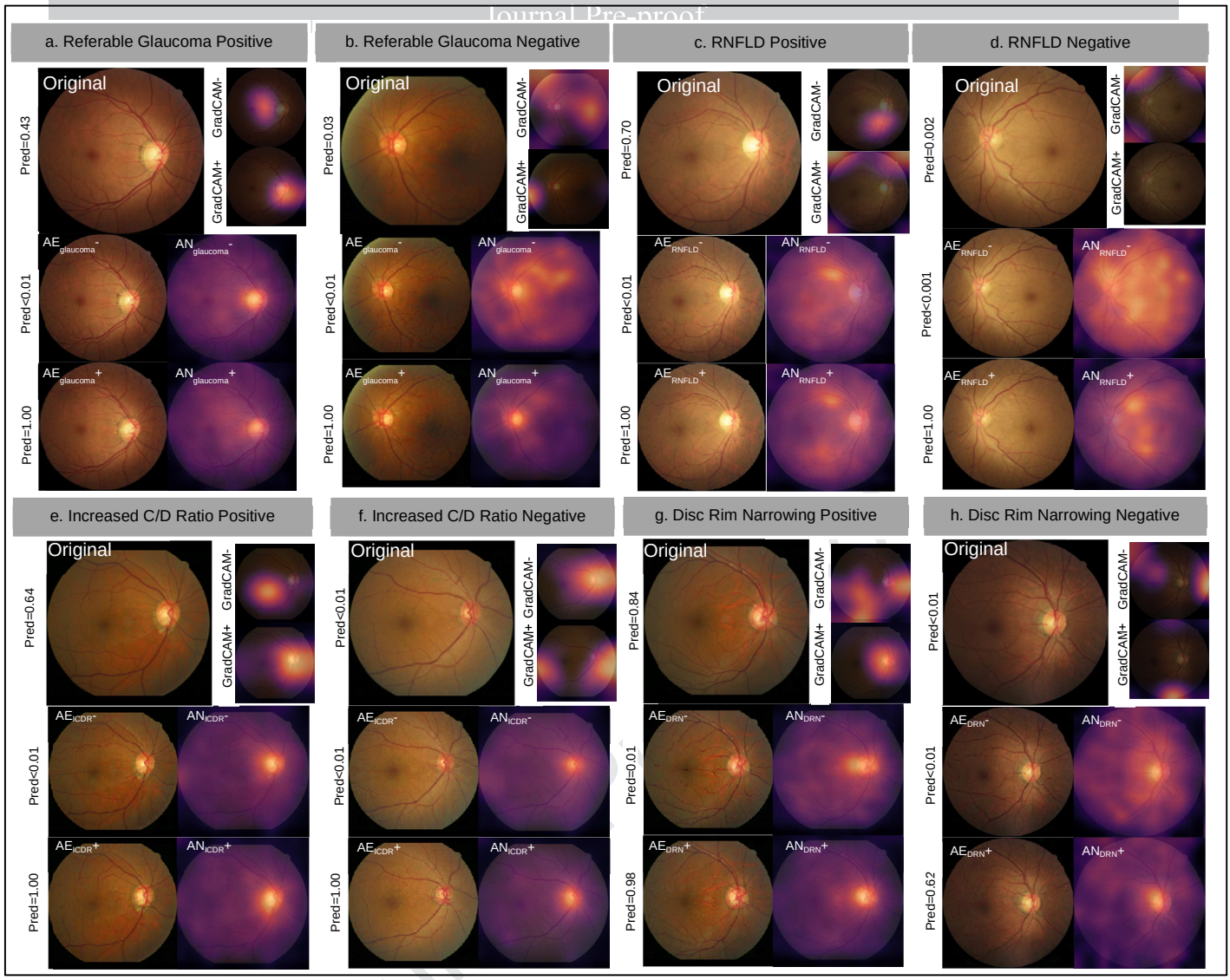
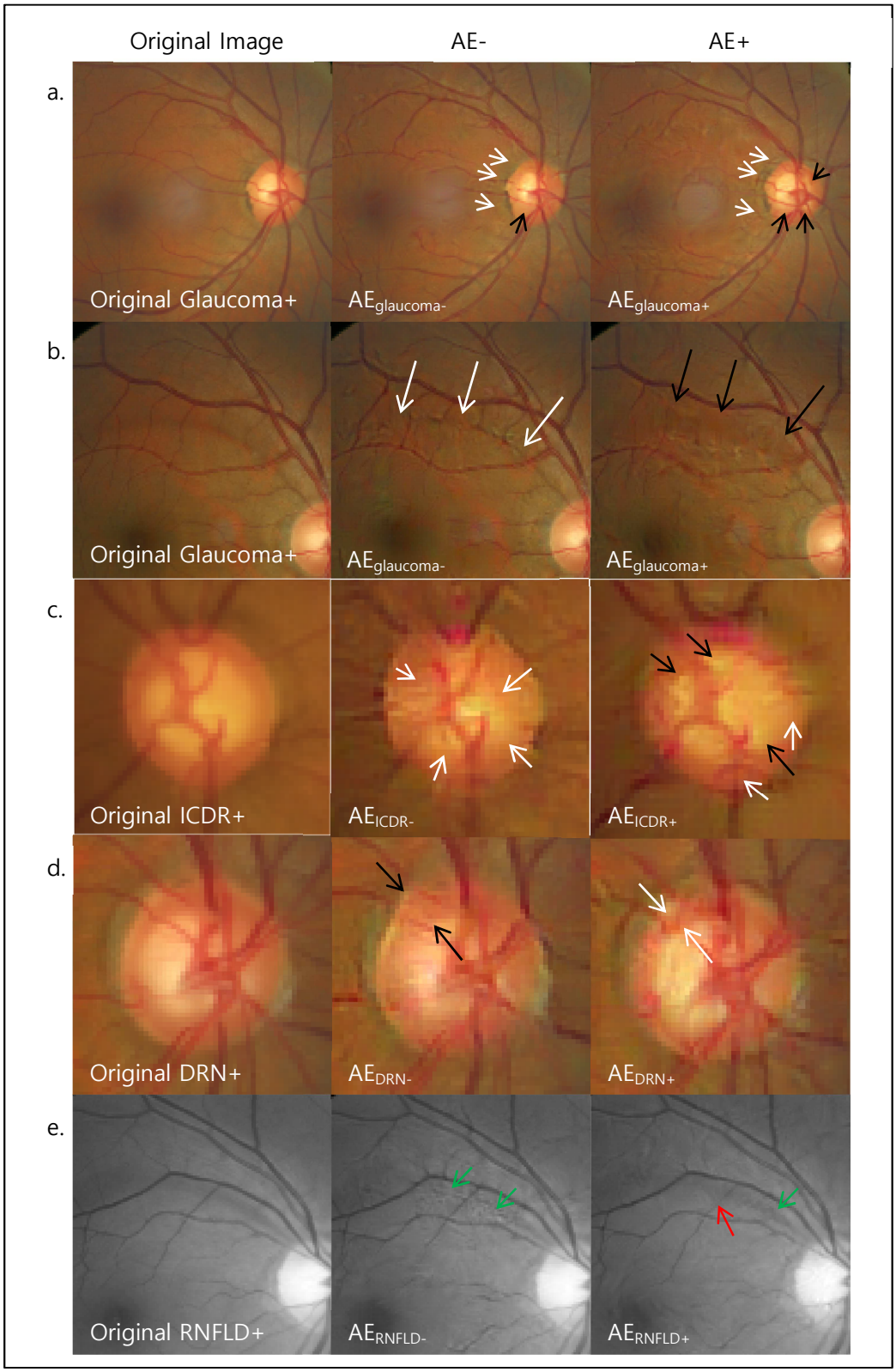


Figure 3. Examples of zoomed-in adversarial explanations for deep learning models for referable glaucoma, increased cup to disc ratio, disc rim narrowing, and RNFL defect.



An evaluation by board-trained glaucoma specialists showed that, compared to a conventional heatmap-based explanation method, adversarial explanation improved the explainability of glaucoma decisions made by deep learning models.