This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge

Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, Devika R G, Hrishikesh P S, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikan, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, Clara I. Sánchez

***Abstract*— The early detection of glaucoma is essential in preventing visual impairment. Artificial intelligence (AI) can be used to analyze color fundus photographs (CFPs) in a cost-effective manner, making glaucoma screening more accessible. While AI models for glaucoma screening from CFPs have shown promising results in laboratory settings, their performance decreases significantly in real-world scenarios due to the presence of out-of-distribution and low-quality images. To address this issue, we propose the Artificial Intelligence for Robust Glaucoma Screening (AIROGS) challenge. This challenge includes a large dataset of around 113,000 images from about 60,000 patients and 500 different screening centers, and encourages the development of algorithms that are robust to ungradable and unexpected input data. We evaluated solutions from 14 teams in this paper and found that the best teams performed similarly to a set of 20 expert ophthalmologists and optometrists. The highest-scoring team achieved an area under the receiver operating characteristic curve of 0.99 (95% CI: 0.98-0.99) for detecting ungradable images on-the-fly. Additionally, many of the algorithms showed robust performance when tested on three other publicly available datasets. These results demonstrate the feasibility of robust AI-enabled glaucoma screening.**

***Index Terms*— Color fundus photography, glaucoma screening, out-of-distribution detection, retina, robustness.**

I. Introduction

Glaucoma is one of the main causes of irreversible blindness and impaired vision in the world. It affects the optic nerve, which connects the eye with the brain, and leads to progressive visual field damage. This damage initially passes unnoticed by the patient. Only in later stages will glaucoma patients experience visual loss. According to estimates, by 2040, over 110 million people will have varying degrees of visual impairment caused by glaucoma [1], with 10% becoming blind in both eyes and 25% in one eye [2]. Many people experience visual impairment from glaucoma because it is often not detected until later stages [3], [4]. Current treatments of glaucoma cannot repair the damage, but can only halt or slow the progression of the condition [5]. Implementing screening programs to identify patients early on for treatment can alleviate the consequences of the disease. Artificial intelligence (AI) may be the enabling technology for the cost-effective implementation of these programs by automatically detecting perimetric glaucoma (i.e., glaucoma in which there is already visual field damage) in color fundus photographs (CFPs) [6]–[10].

Existing AI solutions have been shown to drop in performance in real-world screening practice due to comorbidities, poor quality images, different ethnicities, or unexpected out-of-distribution (OOD) samples [9]. Ad-hoc quality check modules have been added to AI solutions to overcome this performance drop, but recent research has indicated that these quality checks are not sufficiently accurate when deployed in real-world settings [11]. To allow a safe and effective deployment in screening, the reliability and robustness of such solutions need to be assessed. Medical image analysis challenges often exclusively focuses on performance metrics that are potentially unrealistic and overestimated due to the use of test sets that do not represent real-world scenarios. Moreover, metrics to measure reliability and robustness are often neglected due to the difficulty of estimating them in the provided test sets.

To develop solutions that overcome the aforementioned issues related to robustness in glaucoma screening, we organized the Artificial Intelligence for RObust Glaucoma Screening (AIROGS) challenge. The goal of this challenge was to evaluate the feasibility of the development of a state-of-the-art, reliable AI solution that takes a CFP as input and provides as output the likelihood of referable glaucoma, accompanied with outputs for robustness (i.e., predicting whether the input image can be graded reliably or not). The screening task was to distinguish no referable glaucoma (i.e., either no glaucoma at all or non-referable glaucoma, that is suspected pre-perimetric glaucoma) from referable glaucoma. This is different from, for example, distinguishing between multiple glaucoma severity

Manuscript received on the 31st of January 2023; This research was funded in part by Eurostars grant E12712 and supported in part by Amazon Web Services. *(Corresponding author: Coen de Vente.)*

Please see the Acknowledgment section of this article for the author affiliations.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

TABLE I: Statistics of the Rotterdam EyePACS AIROGS dataset. # = number of. CFPs = color fundus photographs.

			Full set	Training set	Test set
# CFPs / # patients			112,732 / 60,071	101,442 / 54,274	11,290 / 5,797
Prevalence (# CFPs (% within set) / # patients)		RG	4,872 (4.3%) / 3,531	3,270 (3.2%) / 2,336	1,602 (14.2%) / 1,195
		NRG	106,306 (94.3%) / 58,388	98,172 (96.8%) / 53,400	8,134 (72.0%) / 4,988
		U	1,554 (1.4%) / 1,415	0 (0.0%) / 0	1,554 (13.8%) / 1,415
Age at encounter (mean ± std. dev.)		All classes	56.9 ± 10.3	56.7 ± 10.2	59.3 ± 10.9
		RG	63.3 ± 10.5	63.0 ± 10.5	63.9 ± 10.7
		NRG	56.5 ± 10.2	56.5 ± 10.1	57.5 ± 10.4
		U	63.7 ± 11.0	N/A	63.7 ± 11.0
# sites			500	486	376
Cameras (# CFPs (% within set) / # patients)	Canon	CR1	11,462 (10.2%) / 6,013	10,274 (10.1%) / 5,422	1,188 (10.5%) / 591
	Canon	CR2	10,523 (9.3%) / 5,538	9,179 (9.0%) / 4,866	1,344 (11.9%) / 672
	Canon	DGI	10,644 (9.4%) / 5,690	9,581 (9.4%) / 5,145	1,063 (9.4%) / 545
	Optovue	iCam 100	29,108 (25.8%) / 16,166	26,480 (26.1%) / 14,742	2,628 (23.3%) / 1,424
	TopCon	NW200	3,109 (2.8%) / 1,588	2,888 (2.8%) / 1,478	221 (2.0%) / 110
	TopCon	NW400	22,519 (20.0%) / 11,736	20,557 (20.3%) / 10,737	1,962 (17.4%) / 999
	Centervue	DRS	1,805 (1.6%) / 988	1,598 (1.6%) / 879	207 (1.8%) / 109
	Nidek	AFC300	61 (0.1%) / 31	53 (0.1%) / 27	8 (0.1%) / 4
	Crystalvue	NFC-700	8 (0.0%) / 4	6 (0.0%) / 3	2 (0.0%) / 1
	Unknown		23,493 (20.8%) / 12,323	20,826 (20.5%) / 10,981	2,667 (23.6%) / 1,342

stages.

To encourage the development of solutions that are robust to any kind of ungradable and unexpected input data and are equipped with inherent robustness mechanisms, the training set we provided was a subset of the full AIROGS dataset where only gradable images were included and ungradable images excluded. The test set, however, is unfiltered, containing all images found in screening settings (gradable and ungradable), representing a real-world scenario.

AIROGS was part of the International Symposium on Biomedical Imaging (ISBI) 2022 challenge program. It re-opened after presenting the results during ISBI 2022 and submissions can still be made on Grand Challenge[1].

Our challenge, along with the dataset we made publicly available, distinguishes itself from previous glaucoma challenges. First, to the best of our knowledge, our dataset is the largest publicly available CFP dataset with glaucoma labels by a large margin. In total, our dataset contains 112,732 CFPs, exceeding the size of other publicly available datasets containing CFPs with glaucoma, of which the sizes range from 22 to 2,000 CFPs [12]–[20]. It is a highly diverse dataset as it originates from 500 screening centers across the United States of America and was acquired with a large variety of cameras. Second, the AIROGS challenge is the first challenge to emphasize robustness in glaucoma screening. Third, AIROGS is one of the first types of challenges on grand-challenge.org that requires participants to submit an algorithm (a *Type 2* challenge), rather than a file with their predictions on the test set (a *Type 1* challenge), as is done in more traditional challenges. This makes human intervention in the generation of test set results impossible, reducing the possibility of cheating. Moreover, it greatly improves reproducibility, allowing everyone to reuse the trained algorithms that were submitted and apply them to new data in a cloud-based environment. Fourth, the reproducibility enabled testing of the participating algorithms on three external datasets: two for evaluating the screening task and one for evaluating robustness.

[1]https://airogs.grand-challenge.org

II. Datasets

A. The Rotterdam EyePACS AIROGS dataset

The Rotterdam EyePACS AIROGS dataset contains 112,732 CFPs from 60,071 subjects and 500 different sites with a heterogeneous ethnicity. The images were originally acquired for a diabetic retinopathy screening program [21]. For grading of the CFPs, all graders were trained and then selected for this task using the European Optic Disc Assessment Trial (EODAT) [22], containing 110 stereoscopic optic nerve photographs, in which all glaucomatous eyes had reproducible visual field defects on standard automated perimetry. 90 experienced ophthalmologists and optometrists were examined and those who scored at least 85% overall accuracy and 92% specificity were selected to label images for the present study. Eventually, 30 out of 90 candidates passed.

For each eye, three images were taken by the camera operators to reduce the number of ungradable eyes. When labeling the images, graders classified one eye at a time. The labeling tool first presented the first CFP for each eye, upon when graders could choose from the options “Referable glaucoma” (RG), “No referable glaucoma” (NRG), or “Ungradable” (U). If a grader selected U for the first image, the tool showed the consecutive CFP. The third image was presented if the second image was also deemed U. Each eye was scored by two separate graders, who were both unaware of the identity of the other grader. If the two graders agreed on the label of a CFP, this became the final label. If they disagreed, the image was scored by one of the glaucoma specialists who passed the EODAT test with at least 95% accuracy. The final label was then based on his judgment.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

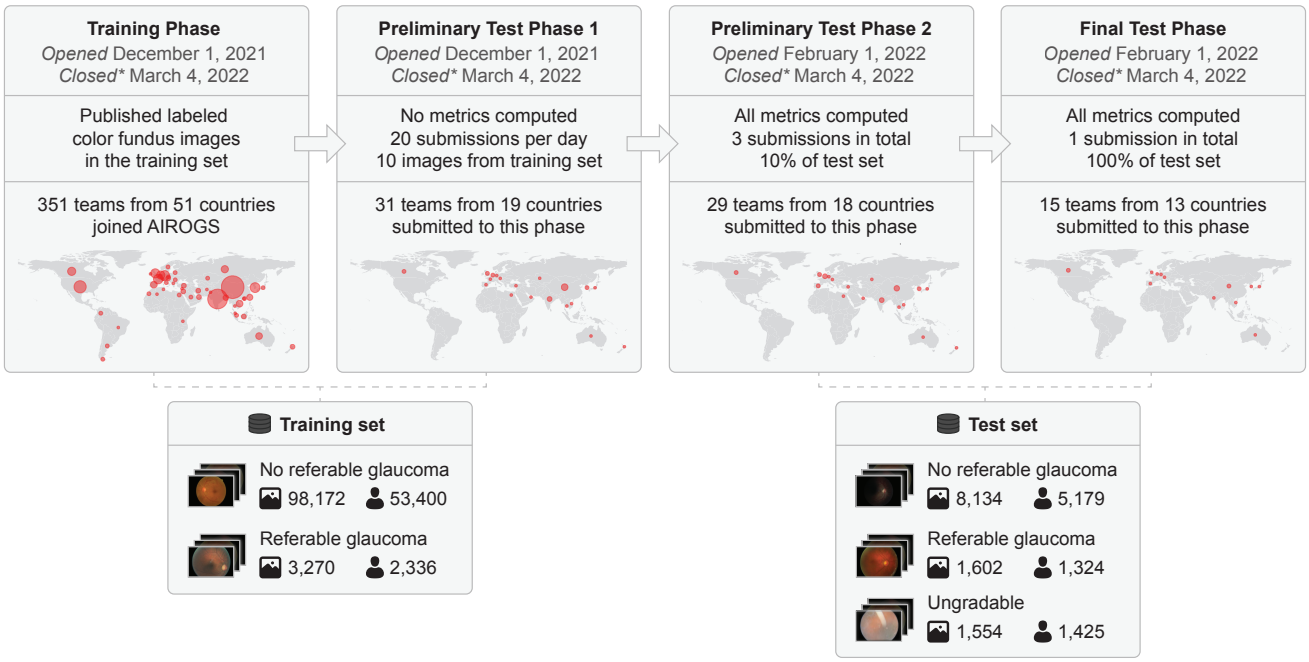


Fig. 1: Overview of all phases in the AIROGS challenge. A world map is shown for each phase that indicates with red circles from which countries the teams that participated in that phase originated. A circle is shown for each country from which at least one team participated and its size represents the number of teams that joined from that country. The relevant subset of the AIROGS dataset for each phase is shown at the bottom of the figure. *All phases reopened for new submissions after the winning teams were announced.

The graders were instructed to select RG if they found glaucomatous signs which they expected to be associated with visual field defects on standard automated perimetry. The signs that could be selected were "appearance neuroretinal rim superiorly", "appearance neuroretinal rim inferiorly", "baring of the circumlinear vessel superiorly", "baring of the circumlinear vessel inferiorly", "disc hemorrhage(s)", "retinal nerve fiber layer defect superiorly", "retinal nerve fiber layer defect inferiorly", "nasalization (nasal displacement) of the vessel trunk", "laminar dots" and "large cup". If the graders did not expect any glaucomatous visual field defects, NRG was to be selected, ignoring any comorbidities (e.g., age-related macular degeneration and diabetic retinopathy). If there was not enough information visible in the CFP to decide between RG and NRG, graders were instructed to select U. Since the goal of this study was to develop solutions for automated screening, glaucoma severity was not reported by the human graders.

The graders were not only evaluated at the start, but they were periodically monitored during the grading process, as well. If their sensitivity or specificity dropped below 80% or 95%, respectively, they were removed from the study and all images they labeled were re-graded by any of the remaining graders. In case a grader wrongly classified a CFP as U, while its final label was NRG or RG, their specificity or sensitivity went down, respectively. In the end, 20 graders remained.

Out of the three CFPs that were available for each eye, we only included the RG or NRG photograph in the dataset if it was available. Otherwise, only one of the U photographs was used. We split the data into a training set of 101,442 CFPs and a test set of 11,290 CFPs, ensuring that data from patients in the training set was not in the test set. We randomly sampled patients when making the split, oversampling patients with ungradable and RG CFPs for the test set, such that approximately 1,600 RG and 1,600 U photographs ended up in the test set. When making the split, we ensured that data from a single patient ended up in either the training set or the test set. Since we were interested in AI solutions that can identify ungradable data without training on ungradable data, we left out all U photographs that ended up in the AIROGS training set. Table I shows statistics about RG, NRG, and U prevalence, age, sites, and cameras for the full dataset, the training set, and the test set. For further information about the acquisition process, labeling, and dataset statistics, including the prevalence of ethnicity, please refer to the paper on the REGAIS dataset, of which AIROGS is a subset [23]. The AIROGS dataset includes 99% of the data from the REGAIS dataset. The difference in size is due to the exclusion of ungradable eyes of patients in the AIROGS set, which was not done in the REGAIS set. The dataset included data from "people of African descent, Whites, Asians, Latin Americans, native Americans, people from the Indian subcontinent, people of mixed ethnicity, and people of unspecified ethnicity" [23]. Approval from the Institutional Review Board of the Rotterdam Eye Hospital was obtained to conduct this research.

B. External datasets

The participants uploaded their trained algorithms, rather than a file with predictions on our test set, to our challenge platform. This enabled us to reuse the developed models on external data after the challenge ended. To evaluate model

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

generalization and to demonstrate this reusability, we applied all trained algorithms to three external datasets: Retinal Fundus Glaucoma Challenge (REFUGE) [18], Glaucoma grAding from Multi-Modality imAges (GAMMA) [20] and Diabetic Retinopathy Image Database (DRIMDB) [24]. The former two are datasets with positive and negative glaucoma CFPs, which we used for externally evaluating the screening performance. We used the latter dataset, which contained different types of ungradable images, to evaluate the robustness externally.

The REFUGE test set contained 400 CFPs, of which 40 CFPs showed glaucoma and 360 CFPs did not. The definition of glaucoma was glaucomatous damage in the optic nerve head area and reproducible glaucomatous visual field defects, which is similar to our definition of glaucoma described earlier [18].

The GAMMA dataset is a multi-modal dataset with optical coherence tomography scans and CFPs for each eye. We used the CFP data from the 100-sample training set as only that subset of the GAMMA dataset had publicly available labels. We defined positive glaucoma in the same way as Wu *et al.* [20], *i.e.*, as the union of the early, intermediate, and advanced glaucoma stages. These stages were defined using the mean deviation (MD) from the visual field reports as follows: an MD less than -6 dB the early stage, an MD between -6 dB and -12 dB for the intermediate stage, and an MD worse than -12 dB for the advanced stage [20]. This resulted in 50 negative and 50 positive glaucoma samples.

DRIMDB is a dataset with 125 "Good" CFPs, 69 "Bad" CFPs, and 22 "Outlier" CFPs. According to Şevik *et al.* [24], one of the criteria of the "Good" category was OD presence. We also manually confirmed the OD was visible in all CFPs that were labeled "Good" in the DRIMDB dataset. Therefore, we assumed the CFPs with th2e category "Good" were gradable. The images labeled "Bad" and "Outlier" were assumed to be ungradable. This resulted in 125 gradable and 91 ungradable images.

III. Challenge setup

The AIROGS challenge consisted of four phases (see Fig. 1). The *Training Phase* opened on the 1st of December 2021 and closed on the 4th of March 2022, providing the participants with approximately three months to develop their solutions. At the start of this phase, the training set was released and has since been available for download under the CC BY-NC-ND license on Zenodo[2].

To ensure fair competition and to encourage the development of inherent robustness mechanisms, teams were not permitted to use additional fundus image training data, including weights pre-trained on fundus image data or in preprocessing steps such as OD segmentation. Manually labeling the challenge data and using the resulting annotations during training was allowed.

To test the algorithms developed by participants, they needed to wrap their trained algorithm in a Docker[3] container and submit it to our challenge platform. This allows the submitted algorithms to be run on data that is not directly accessible by the participating teams. Example code for generating such a containerized submission can be found on GitHub[4]. *Preliminary Test Phase 1* opened and closed simultaneously with the *Training phase* and served as a check for whether the submitted algorithms could be run on the challenge platform and produced the output in the expected format. Algorithms were tested on 10 images from the training set for this check. All algorithms were executed on the challenge platform using an NVIDIA T4 GPU (16 GB VRAM) with 8 CPUs (32 GB RAM).

The test set was and still is closed, meaning the image data and the labels are private and cannot be downloaded. *Preliminary Test Phase 2* opened on the 1st of February 2022 and we allowed three submissions per team to this phase, as it used 10% of the test set for evaluation. All challenge metrics were also computed and reported back to the participants. The *Final Test Phase* opened simultaneously with *Preliminary Test Phase 2*, but algorithms were tested on 100% of the test data and only one submission per team was allowed. The challenge metrics computed for this phase were used for the final team ranking.

The algorithms were expected to produce four outputs, of which two were related to glaucoma screening performance (i.e., image classification of RG and NRG) and the other two to robustness (i.e., the identification of U). The glaucoma screening outputs were a likelihood score for RG (O_1) and a binary decision for RG (O_2, positive if RG and negative if NRG). The ungradability outputs were a binary decision on whether the image is ungradable (O_3, positive if ungradable and negative if ungradable) and a non-thresholded scalar value that is positively correlated with the likelihood for ungradability (e.g. the entropy of a probability vector produced by a machine learning model or the variance of an ensemble) (O_4). Output O_2 was not used in the evaluation pipeline for the challenge leaderboard, but it was requested by the challenge organizers for further analysis.

The evaluation was also based on the two aspects of screening performance and robustness, with two metrics per aspect. Screening performance was evaluated using the standardized partial area under the receiver operating characteristic curve [25] (90-100% specificity) for RG ($pAUC_S$), and the sensitivity at 95% specificity ($SE@95SP_S$). These metrics were based on these specificity ranges, as a high specificity is required for cost-effective glaucoma screening due to its relatively low prevalence [26], [27]. $pAUC_S$ and $SE@95SP_S$ are both based on output O_1. For evaluating the robustness, we determined the model's agreement with the human reference on ungradability using Cohen's kappa score (κ_U), calculated using output O_3. Furthermore, we calculated the area under the receiver operator characteristic curve using the human reference for ungradability as the true labels and output O_4 as the target scores (AUC_U).

To determine the final ranking, we first ranked all participants on the four individual metrics $pAUC_S$, $SE@95SP_S$, κ_U, and AUC_U resulting in the rankings R_{pAUC_S}, $R_{SE@95SP_S}$, R_{κ_U}, and R_{AUC_U}, respectively. The final score S_{final} was then calculated as the mean of those rankings:

[2]https://zenodo.org/record/5793241

[3]https://docker.com

[4]https://github.com/qurAI-amsterdam/airogs-example-algorithm

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

$$S_{final} = \frac{R_{pAUC_S} + R_{SE@95SP_S} + R\kappa_U + R_{AUC_U}}{4}. \quad (1)$$

The final ranking (later also referred to as *Mean position*), was based on S_{final}, where a lower value for S_{final} resulted in a higher ranking.

We calculated 95% confidence intervals (CIs) with non-parametric bootstrapping using 1000 iterations [28]. The code for evaluating submissions can be found on GitHub[5]. The performance of human graders was calculated by comparing the labels given by the individual graders (excluding the two glaucoma specialists, since the final labels were equal to their decision in case of disagreement) to the final labels as defined in Section II-A. To compute the performance of all human graders combined, each image was weighted equally in the calculation of the metrics. We also evaluated ensembles of participating algorithms, which were generated by averaging the outputs of these algorithms.

IV. Participating Methods

Fifteen teams submitted a working solution to the *Final Test Phase*, of which one team did not opt-in to contribute to the current paper. In this section, we present the methods of the fourteen participating teams. More extensive descripions are available on the AIROGS challenge website[6] and a selection of the participating methods were included in the ISBI challenge proceedings [37]–[39]. Table II and III summarize the participating methods in a structured manner.

A. PUMCH-eye [37]

The *PUMCH-eye* team proposed an approach with five trained models in their workflow. The first model (M_{disc}) was a segmentation model with ResNet101-UperNet [40] as the backbone that segmented the OD in the input CFP. For the development of this model, they manually labeled the OD in 40 images. In case M_{disc} successfully detected the OD, they computed the center c and the diameter d of the segmentation to crop the input image around c with size $3d$. This cropped image was then fed into a vision transformer [41] for the binary classification of RG and NRG. If the OD detection was unsuccessful, they fed the original input image to a different vision transformer for binary classification of RG and NRG.

The team also developed a vessel segmentation model with 40 images in which they manually annotated vessels (M_{vessel}). They trained a ResNet-18 (R_{vessel}) which took the output of M_{vessel} as input data, using the first 500 images in the training dataset and 100 manually selected images in the training set with relatively poor image quality. This classfication model served as one of the inputs for ungradability classification. The second input was taken from M_{disc}. The ungradability likelihood output (O_4) was then defined as the output likelihood of the binary classification model R_{vessel} (i.e., O_{vessel} or, if the M_{disc} could not detect an OD, as $R_{vessel} + 0.75$. O_3 was positive if O_4 was at least 0.95 and negative otherwise. The vessel segmentation model was evaluated on four randomly selected images, on which a Dice score of 0.787 was achieved. This was lower than the state-of-the-art for this problem [42]. However, this was expected to be sufficient for the downstream task of ungradability detection.

B. RWTH-CuP [38]

The *RWTH-CuP* team proposed an approach with two steps consisting of cropping around the OD by employing a detection network, followed by an ensemble of transformers (Swin Transformer-B [43] and DeiT-S [44]) and convolutional neural networks (CNNs) (EfficientNet-B4 [45] and EfficientNetV2-M [46]) that classifies the cropped image. They manually labeled the OD and its environment in 3,221 CFPs to develop this detection network, for which they trained a YOLOv5 [47] object detector network.

For ungradability classification, the team used a hybrid approach. As the probability that an image is ungradable is high if the OD could be found by the object detector network, they employed the confidence score of the YOLOv5 detection model as one of the ungradability measures. To capture other ungradability causes, such as blurred depictions of the OD, they trained an additional classifier on a manually selected subset of the CFPs in the development set. The team considered the 4000 CFPs with the lowest confidence score of the object detector and manually selected 600 images that were assumed by the team to be very close to being classified as ungradable. They used another set of 2,000 high-quality images to train an EfficientNet-B4 [45] ungradability classification model. O_4 was then defined as $(1 - c) + g$, where c is the object detection confidence and g the output the ungradability classification model. The binary O_3 output value was determined using a cut-off manually determined by a medical doctor in 20,000 images from the development set for which O_4 was computed.

C. Eyelab [48]

The *Eyelab* team employed a two-stage approach for glaucoma classification. The first step was to detect and crop the OD area and the second step was a vision transformer [49] that classified the cropped image from the first step. For the detection model, they trained a YOLOv5 [47] model using semi-automatically generated labels. Their method for ungradability detection was based on whether the optic disc detection model from the first step found an optic disc to be present.

D. Tien [50]

Tien used an ensemble of an EfficientNet [45] and DenseNet [51] for the classification of RG and NRG. For the ungradability task, they used an autoencoder network and a blending engine. They used the reconstruction error as a measure of the likelihood of ungradability. The higher the reconstruction error, the more likely it is that the image is ungradable. The blending engine fused the probability output from the binary classification model as a weight factor to the reconstruction

[5]https://github.com/qurAI-amsterdam/airogs-evaluation

[6]https://airogs.grand-challenge.org/evaluation/final-test-phase/leaderboard/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

TABLE II: Method overview from all participating teams for the screening task. OD = optic disc. #ODs = number of CFPs in which the OD was manually labeled. #vessels = number of CFPs in which the vessels were manually labeled.

Screening (part 1: Architecture, Pre-processing, Manual labels, Input size)

#	Team	Architecture																Pre-processing			Manual labels		
		Vision transformer	Swin-Transformer-B	EfficientNet	DeiT-S	MobileNet	ResNet	ResNet-RS	SeResNeXt	VGG	DenseNet	Inception-V3	ConvNeXt	SeNet	SeResNet	Inception-ResNet-v2	Ensemble	Crop field-of-view	Crop OD	Histogram equalization	#ODs	#vessels	Input size of classification model
1	PUMCH-eye	✓																	✓		40	40	384^2
2	RWTH-CuP		✓	✓	✓												✓		✓	✓	3,221		224^2
2	Eyelab	✓																	✓		1,500		384^2
4	Tien			✓							✓						✓	✓					$1,024^2$
5	UPF+AIML					✓											✓						512^2
6	FMS-CETCV						✓																512^2
7	ICT_HCI						✓											✓					512^2
8	SK							✓									✓	✓					256^2
9	SACM								✓	✓	✓	✓					✓		✓		735		$120^2/224^2$
10	UPRetina-UR			✓				✓										✓					512^2
11	OPTIMATeam											✓											224^2
11	MA			✓					✓		✓			✓	✓	✓		✓					512^2
13	YC										✓						✓		✓	✓	101		608^2
14	Mirazzak	✓											✓				✓						$224^2/384^2$

Screening (part 2: Loss function, Optimizer, Pre-training, Class imbalance solution)

#	Team	Loss function			Optimizer				Pre-training			Class imbalance solution	
		Cross-entropy	Focal loss	Kullback-Leibler	Adam [29]	AdamW [30]	SGD	SAM [31]	ImageNet	Masked auto-encoder [32]	Noisy student [33]	Balanced sampling	Weighted loss
1	PUMCH-eye	✓				✓				✓			
2	RWTH-CuP	✓			✓	✓			✓			✓	
2	Eyelab	✓					✓					✓	
4	Tien	✓			✓							✓	
5	UPF+AIML	✓				✓		✓	✓				
6	FMS-CETCV		✓				✓		✓				
7	ICT_HCI	✓			✓				✓			✓	
8	SK	✓			✓				✓				✓
9	SACM	✓				✓	✓		✓				✓
10	UPRetina-UR		✓		✓				✓			✓	
11	OPTIMATeam			✓	✓				✓			✓	
11	MA	✓			✓				✓			✓	
13	YC	✓					✓				✓		✓
14	Mirazzak	✓			✓				✓				✓

Screening (part 3: Data augmentation during training)

#	Team	Brightness modifications	Color jitter	Elastic deformations	Flipping	Bluring	Rotation	Translation	Zooming	Shearing	Quality reduction	Sharpening	Erasing	AutoAugment [34]	RandAugment [35]	TrivialAugment [36]
1	PUMCH-eye													✓		
2	RWTH-CuP			✓	✓						✓					
2	Eyelab	✓			✓		✓	✓	✓	✓			✓			
4	Tien				✓		✓		✓							
5	UPF+AIML															✓
6	FMS-CETCV				✓											
7	ICT_HCI														✓	
8	SK		✓			✓	✓	✓	✓							
9	SACM															
10	UPRetina-UR														✓	
11	OPTIMATeam	✓			✓	✓	✓	✓	✓							
11	MA	✓			✓	✓	✓	✓	✓			✓				
13	YC	✓			✓	✓	✓	✓	✓							
14	Mirazzak				✓		✓									

TABLE III: Method overview from all participating teams for the ungradability task and the deep learning frameworks they used. OD = optic disc. AE = autoencoder. VAE = variational autoencoder. rec. error = reconstruction error. OOD = out-of-distribution.

		Robustness				General
#	Team	Method	OD detection for uncertainty or confidence estimation	Threshold based on manual identification of low-quality images in the development set	Input size for ungradability approach	Deep learning framework
1	PUMCH-eye	Vessel and OD segmentation	✓	✓	384^2	PyTorch
2	RWTH-CuP	OD detection confidence	✓	✓	224^2	PyTorch
2	Eyelab	OD presence detection	✓	✓	384^2	PyTorch
4	Tien	AE rec. error + probability classification model			$1,024^2$	PyTorch
5	UPF+AIML	Synthetic image degradations		✓	512^2	PyTorch
6	FMS-CETCV	Interpolated gaussian descriptor			256^2	PyTorch
7	ICT_HCI	Probability classification model		✓	512^2	PyTorch
8	SK	Energy-based OOD + activation rectification			256^2	PyTorch+Lightning
9	SACM	AE and VAE rec. error + OD detection confidence	✓	✓	288^2	PyTorch
10	UPRetina-UR	Test-time augmentation + probability classification model			512^2	PyTorch+fast.ai
11	OPTIMATeam	Deep Dirichlet uncertainty			224^2	TensorFlow+Keras
11	MA	Ensemble			512^2	TensorFlow+Keras
13	YC	Monte-Carlo Dropout			608^2	TensorFlow
14	Mirazzak	Regret function			$224^2/384^2$	PyTorch

error. The highest weight was 1 (when the probability was 0.5) and the weight was lowest when the probability is certain (either 0 or 1).

E. UPF+AIML [52]

Team *UPF+AIML* trained two separate models, both based on the MobileNet-V2 [53] architecture for lightweight training, and both optimized with the Sharpness-Aware Minimization (SAM) [31] technique for better generalization. The first model was trained on the available training set for the screening task. The second model was tasked with identifying out-of-distribution data, i.e., ungradable images in this case. For this, ungradable images were simulated by applying four image transformations (brightness, gamma, saturation, and blur) on-line to the data, with such a strength that they would destroy image content and turn images useless for diagnostic purposes. The ungradability detection model was trained on a mixture of gradable (sampled directly from the original training set)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

and ungradable (simulated). After training, this model was applied to the training set, where all images were expected to be gradable. The threshold that would classify 0.1% of the training set as ungradable was selected for ungradabiltiy detection.

F. FMS-CETCV [54]

The *FMS-CETCV* team used a binary classifier with ResNet-50 as the backbone for classifying RG and NRG. They used focal loss [55] to account for the class imbalance in the training set.

For the classification of ungradable images, they used a self-supervised learning approach, inspired by the work of Oza *et al.* [56], where a one-class classification method was presented for unsupervised anomaly detection. The one-class classifier builds a feature space by extracting the features of the training sample which contain only the positive samples (i.e., gradable images). They used an encoder with ResNet-18 as the backbone, which is trained on the AIROGS test set that only contains gradable images. The feature space produced by this encoder is then used by a Gaussian anomaly classifier to distinguish gradable and ungradable images.

G. ICT_HCI [57]

Team *ICT_HCI* used ResNet-50 for their RG and NRG classification model. During inference, they made five random 512x512 crops of the image and then provided all crops separately to the model and get five scores. If the maximum of five scores was greater than 0.9, they let it be the output of the model, otherwise, they took the mean of the five scores as the output of the model.

The team used the minimum class probability of the two classes RG and NRG as the likelihood for ungradability. If and only if the ungradability likelihood was greater than 0.1, they set O_3 to be positive.

H. SK [58]

The *SK* team employed ResNet-RS [59] for RG and NRG classification. They replaced the final linear layer of ResNet-RS with a single linear layer with two channel outputs.

For ungradability classification, they used an inference-time OOD energy-based method [60] combined with activation rectification [61]. The energy-based method uses a scoring function based on energy, instead of softmax, to discriminate in-distribution (ID) and OOD data. In activation rectification, the outsized activation of a few layers can be attenuated by rectifying the activations at an upper limit. After rectification, the output distributions for ID and OOD data become much more well-separated. It is based on the observation that the mean activation for ID data is well-behaved with a near-constant mean and standard deviation, and the mean activation for OOD data has significantly larger variations across units and is biased towards having sharp positive values.

I. SACM [62]

Team *SACM* used the YOLOv5 [47] detection model to crop the optic disc in the CFP input image. In a semi-automated process, they manually labeled the locations of 735 optic discs and trained the detection model with 4,088 in total. The cropped image was then passed through an ensemble of classifiers (SeResNext-50 [63], VGG-16 [64], DenseNet-161 [51], EfficientNet-B5 [45], EfficientNet-B7 [45] and Inception-V3 [65]) to make the final prediction. They also used test-time augmentation.

For the robustness task, they used the detection model confidence, an autoencoder, and a variational autoencoder (VAE) [66]. They combined these three aspects of their pipeline using this formula to achieve a final ungradability score for O_4 as $(1-c)\cdot s\cdot p_{autoencoder}\cdot p_{vae}$, where c refers to the detection network confidence and $p_{autoencoder}$ and p_{vae} refer to the mean squared error between the input and output of the autoencoder and VAE, respectively.

J. UPRetina-UR [67]

The *UPRetina-UR* team used ResNet-RS-50 [59] for the classification of RG and NRG. They oversampled cases with RG during training to account for the class imbalance.

They employed a closed-set classification approach for the ungradability task based on the method proposed by Vaze *et al.* [68]. They applied test-time augmentation to obtain five predictions that are averaged to produce O_4.

K. OPTIMATeam [39]

OPTIMATeam used the first two blocks from the Inception-V3 [65] network for the classification of RG and NRG. They only used these two blocks to reduce the receptive field size, which was necessary for their ungradability approach.

The ungradability approach was based on the direct modeling of the uncertainty following the evidential deep learning approach [69]. They used Deep Dirichlet uncertainty estimation as the ungradability score O_4. To set a threshold for getting a binary value for O_3 based on O_4, they assumed that diagnosis is only possible if the OD has enough image quality for diagnosis, as glaucoma's main structural manifestation occurs in that region. They applied Grad-CAM [70] on the trained model for the screening task and occluded out the region where Grad-CAM was greater than 0.5. This allowed them to produce ID and OOD samples in their validation set, with which they computed the threshold for the binary ungradability decision. In particular, they constructed a receiver operating characteristic (ROC) curve using their values for O_4 with these ID and OOD samples. The ROC threshold where the sensitivity was 0.5 was set to calculate O_3.

L. MA [71]

Team *MA* used an ensemble of these twelve different architectures for the glaucoma screening task: SeNet-154 [63], SeResNet-101 [63], SeResNeXt-101 [63], EfficientNet-B1 [45], EfficientNet-B2 [45], EfficientNet-B3

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

[45], EfficientNet-B4 [45], EfficientNet-B5 [45], EfficientNet-B6 [45], EfficientNet-B7 [45], DenseNet-201 [51], Inception-ResNet-v2 [72]. The RG likelihood was computed by averaging the likelihoods of all respective models in the ensemble.

The ungradability output O_4 was the sum of the variances between all models in the ensemble for the positive and negative class probabilities. O_3 was positive if O_4 exceeded 0.2 and negative otherwise.

M. YC [73]

The *YC* team used two DenseNet-121 networks to classify RG and NRG in the CFPs. The first network was trained with the full CFP as input and the second network used a version of the CFP that was cropped around the optic disc as input. After the last convolutions of these networks, a fully connected layer with dropout was added. The outputs of these fully connected layers were then concatenated and used as the input to another fully connected layer with dropout, which was followed by the final layer of the network. For cropping the CFPs around the optic disc, they trained a U-Net [74] with a DenseNet-121 [51] backbone. To train this segmentation network, they first roughly annotated the position of the optic disc in 101 CFPs in the training set. Subsequently, they generated reference segmentation maps using a probability density function of the multivariate normal distribution around the annotated optic disc position.

They used Monte-Carlo drop-out [75] with 20 predicted probabilities per image for the robustness task. Then they statistically tested a Wilcoxon one-sample test whether the mean of the predicted probabilities was equal to 0.5. The team defined ungradability for predicting glaucoma as the logarithm of the p-value for the Wilcoxon test.

N. Mirazzak [76]

Team *Mirazzak* used an ensemble of ConvNeXts [77] and a vision transformer for the screening performance task.

For the ungradability task, they employed the *regret* function, which was proposed by Bibas *et al.* [78] as the generalization error of an explicit expression of the predictive normalized maximum likelihood learner. If the value of *regret* function was high, the samples were considered OOD and they were marked as ungradable.

V. Results

This section presents the glaucoma screening performance and robustness of the fourteen participating teams. The final rankings and mean positions of the teams are shown in the first plot of Fig. 2. Four teams shared a rank with another team, since their mean positions were exactly equal, causing there to be two teams for each of the ranks #2 and #11.

A. Glaucoma screening performance

The glaucoma screening performance of the participating teams is summarized in Fig. 2, showing $pAUC_S$ and $SE@95SP_S$ in the second and third plot, respectively. The highest scores for $pAUC_S$ and $SE@95SP_S$ were 0.90 (95% CI: 0.89 – 0.91) and 0.85 (95% CI: 0.83 – 0.87), respectively. These scores were both achieved by team *PUMCH-eye*.

Fig. 3a and Fig. 3b show the $pAUC_S$ and $SE@95SP_S$ for the ensembles when averaging the RG likelihood output O_1 of the best M participants in terms of the relevant metric. An optimal $pAUC_S$ of 0.91 (95% CI: 0.90 – 0.92) was achieved at $M = 3$. At $M = 2$, an optimal value for $SE@95SP_S$ was reached, which was 0.87 (95% CI: 0.85 – 0.89).

Fig. 4a shows the partial ROC curves between 90% and 100% specificity for all participants. The plot also presents the sensitivity and specificity of the human graders with a 95% CI. These were 0.86 (95% CI: 0.84 – 0.87) and 0.94 (95% CI: 0.94 – 0.95), respectively.

In Fig. 5, we compare the performance on the REFUGE test set of the final AIROGS algorithms, which were trained on the AIROGS train set, to the performance of the algorithms that were submitted to the REFUGE challenge, which were trained on the REFUGE train set. The top three participants of the REFUGE algorithms achieved AUCs of 0.99, 0.98 and 0.96. For the AIROGS algorithms, the best three AUCs were 0.98, 0.97 and 0.97. The mean $\pm$ std. dev. AUC of all REFUGE and AIROGS algorithms were 0.94 $\pm$ 0.04 and 0.95 $\pm$ 0.02, respectively. Fig. 6 presents the relation between the two glaucoma screening performance metrics of all participating AIROGS algorithms on the AIROGS test set and that performance on the REFUGE test set. For both metrics, almost all AIROGS algorithms (except for team *PUMCH-eye* for $SE@95SP_S$) scored higher on REFUGE than on AIROGS. Of all AIROGS participants, the best $pAUC_S$ and $SE@95SP_S$ on REFUGE were 0.94 and 0.88, respectively.

In Fig. 7, the relation between the glaucoma performance of all participating AIROGS algorithms on the AIROGS test set and that performance on GAMMA is shown. For both screening metrics, all AIROGS algorithms scored higher on GAMMA than on AIROGS. Of all AIROGS participants, the best $pAUC_S$ and $SE@95SP_S$ on GAMMA were 1.0 and 1.0, respectively.

B. Robustness

The robustness metrics of the participating teams are summarized in Fig. 2, showing κ_U and AUC_U in the fourth and fifth plot, respectively. The highest scores for κ_U and AUC_U were 0.82 (95% CI: 0.80 – 0.84) and 0.99 (95% CI: 0.98 – 0.99), respectively. These scores were achieved by team *Temirgali* and *RWTH-CuP*, respectively.

Fig. 3c and Fig. 3d show the κ_U and AUC_U for the ensembles when averaging output O_3 and output O_4, respectively, of the M best algorithms in terms of these respective metrics. An optimal κ_U of 0.85 (95% CI: 0.84 – 0.86) was achieved at $M = 6$. Also at $M = 6$, an optimal value for AUC_U was reached, which was 0.99 (95% CI: 0.99 – 0.99).

In Fig. 4b, ROC curves for robustness are shown for all participants. The plot also presents the sensitivity and specificity for separating ungradable from gradable images of the human graders with a 95% CI. These were 0.95 (95% CI: 0.94 – 0.96) and 0.97 (95% CI: 0.97 – 0.97), respectively.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

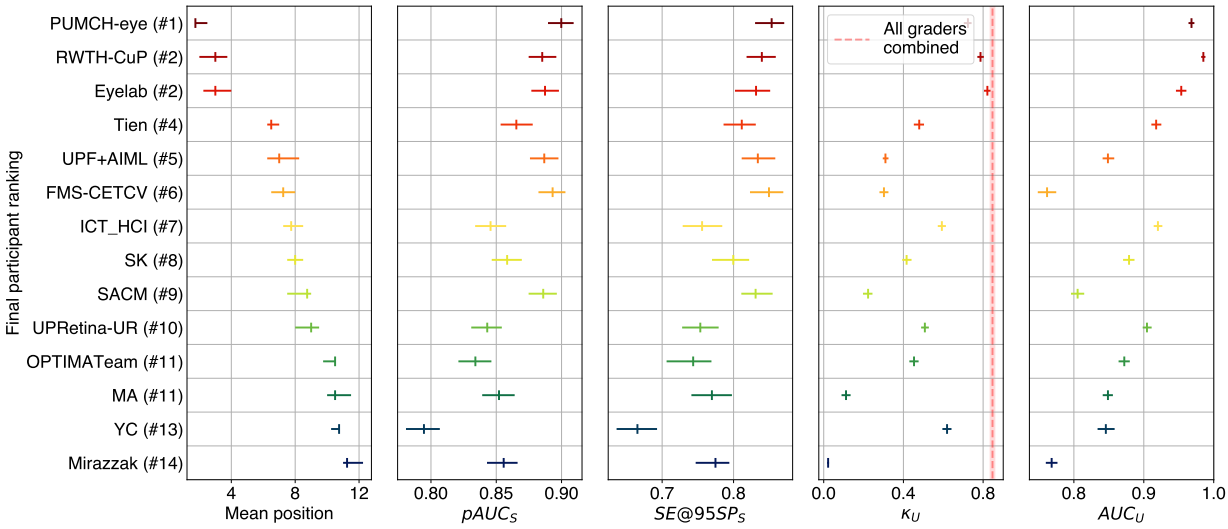


Fig. 2: Final rankings of all participating teams. The teams are sorted by their final ranking and therefore also by their mean position. The mean position is shown in the left plot and the four challenge metrics are shown in the other four plots. The κ_U of all human graders is indicated with a red dotted line. The width of the horizontal lines in all plots and the shaded area in the plot for κ_U are 95% CIs. We consistently use the same colors to refer to teams in other figures in this manuscript.

The results on the external DRIMDB dataset are shown in Fig. 8, indicating the relation between the ungradability metrics κ_U and AUC_U of all participating AIROGS algorithms on DRIMDB and those metrics on AIROGS. Of all AIROGS participants, the best κ_U and AUC_U on DRIMDB were 0.94 and 1.0, respectively.

C. Inference time

The time that each algorithm took to perform inference on the test set of the *Final Test Phase* is shown in Fig. 9. Please note that these results reflect the total time it took to run the Docker containers provided by the participants. The time it took for the software to load the model weights and to run other setup code defined by the teams was also included in this analysis. Since the test set was split into 38 separate chunks, this initialization and setup code was run at least 38 times for each participating team, as well.

VI. Discussion

AI models have been shown to be effective at detecting glaucoma in CFPs, but most studies lack evidence of robustness to real-world scenarios in which unexpected OOD data can be presented due to various causes. To this end, we relied on the community to develop robust AI solutions for glaucoma screening based on the largest multi-center real-world CFP dataset with glaucoma labels. We organized the AIROGS challenge around this dataset, ensuring the resulting algorithms are reusable in a cloud-based environment. We applied these algorithms to ungradable data, while the participants could only train on gradable data to ensure robustness to any kind of ungradable data, and to other publicly available datasets to assess their generalization.

A. Overall findings

The team with the highest $SE@95SP_S$ scored expert-level screening performance on the AIROGS test set with a sensitivity of 0.85 (95% CI: 0.83 – 0.87) at 95% specificity, similar to the sensitivity of 0.86 (95% CI: 0.84 – 0.87) at a specificity of 0.94 (95% CI: 0.94 – 0.95) of human graders. The highest $pAUC_S$ that was achieved by any of the teams was 0.90 (95% CI: 0.89 – 0.91). Ensembling the different participating methods improved the screening performance even further, to 0.91 (95% CI: 0.90 – 0.92) and 0.87 (95% CI: 0.85 – 0.89) for $pAUC_S$ and $SE@95SP_S$, respectively. Our analysis revealed that ensembling improved the performance for all metrics up to a certain point at which adding further models to the ensemble resulted in a decline in performance. A probable reason for this is that the models added after this point under-perform to such a degree that their outputs negatively impact the performance of the ensemble, instead of improving it. Seven out of fourteen teams exceeded the minimum performance of 80% sensitivity and 95% specificity that was required by human graders who were periodically monitored during the grading process. This shows these models can provide similar performance to human graders for glaucoma screening, suggesting that AI can potentially play a role in an automated screening process.

We also evaluated the screening performance of the algorithms on two external test sets. Even though the algorithms were trained on AIROGS data, they achieved very high performances on the two external test sets, showing reproducible results in different sets and populations. On average, the participating AIROGS algorithms scored slightly higher on the REFUGE dataset than the REFUGE participants. We found that the participating algorithms scored substantially higher on these external datasets than on the AIROGS test set, indicating the value of a challenging real-world dataset. This strong generalization of the developed solutions also shows the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

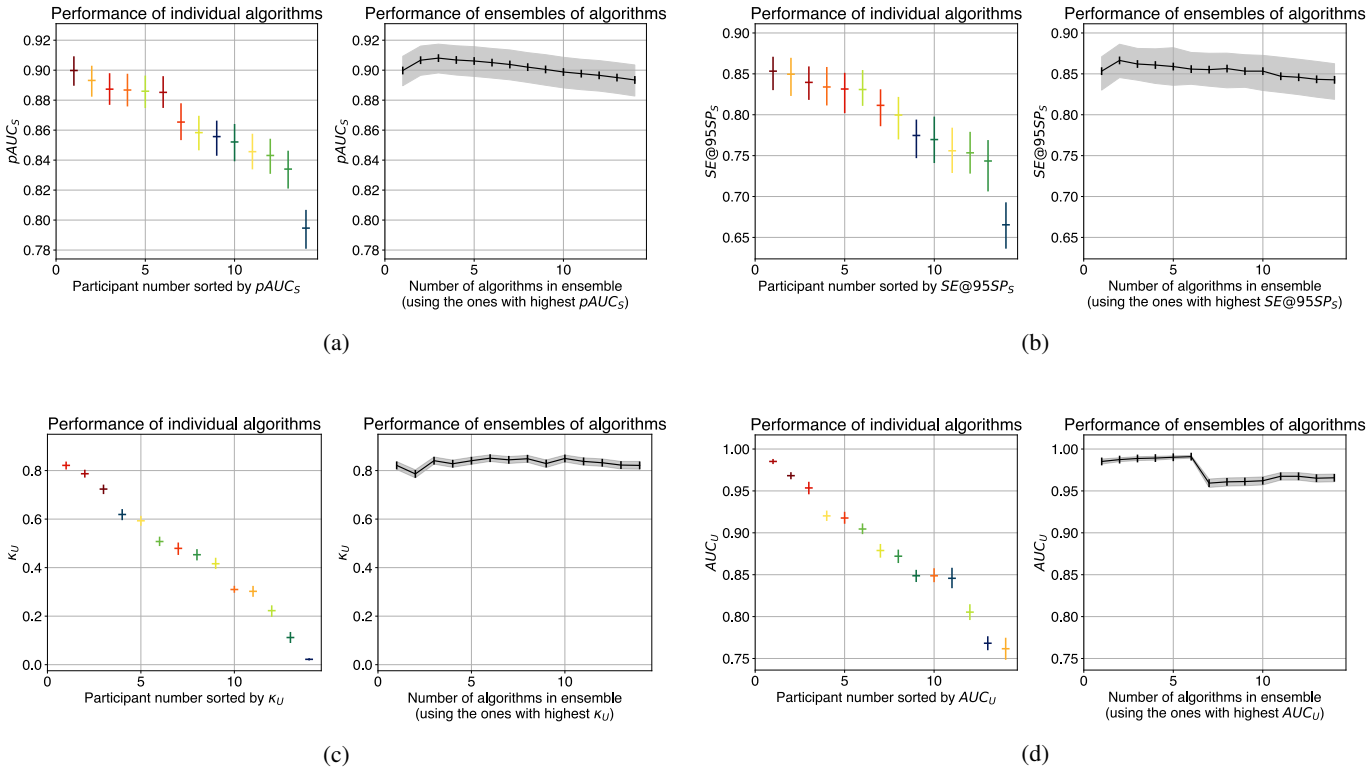


Fig. 3: The four challenge metrics (a) $pAUC_S$, (b) $SE@95SP_S$, (c) κ_U, and (d) AUC_U for the ensembles generated by incrementally fusing one algorithm at a time. The algorithms were fused by averaging the outputs of all algorithms in the ensemble. The vertical lines in the left plot and the shaded areas in the right plots indicate 95% CIs.

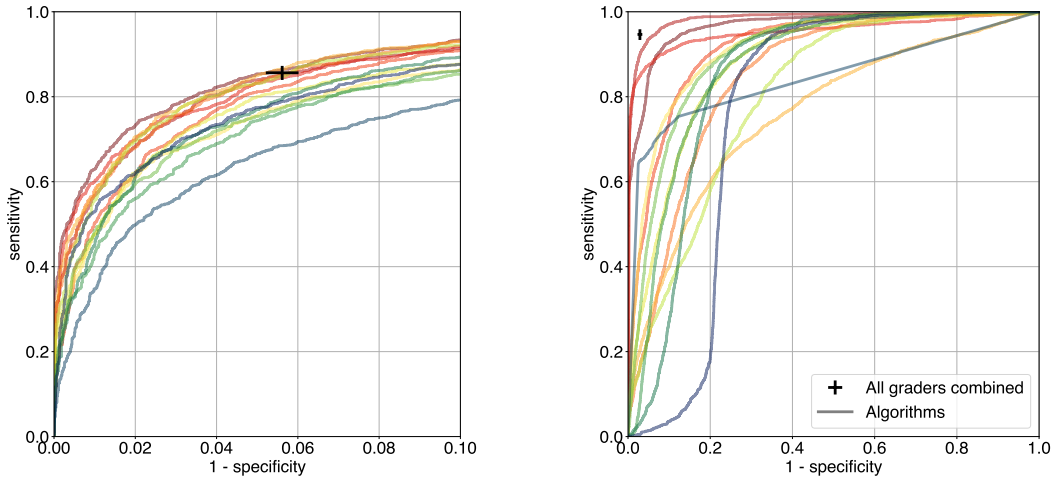


Fig. 4: ROC curves for both challenge tasks. The sensitivity and specificity of all human graders on the AIROGS test set combined are indicated with black lines. Respectively, the width and height of the black horizontal and vertical lines are 95% CIs. In (a), the partial ROC curve (90%-100% specificity) for screening is shown, with 1,602 positive (RG) and 8,134 negative (NRG) images from the AIROGS test set. In (b), the ROC curve for robustness is shown with 1,554 positive (ungradable) and 9,736 negative (gradable) images from the AIROGS test set.

potential of models trained on our dataset to be successfully implemented in screening programs with limited to no loss of performance. Unlike the external datasets, the AIROGS dataset represents a screening population, which likely consists of a relatively large amount of individuals with lower severity levels of glaucoma compared to clinical populations. Since more severe cases are expected to be picked up easier than less severe cases, the underlying ratio between less and more severe positive glaucoma cases could be a cause of the observation that the external test set performance was higher than the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

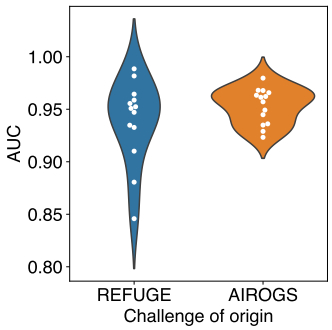
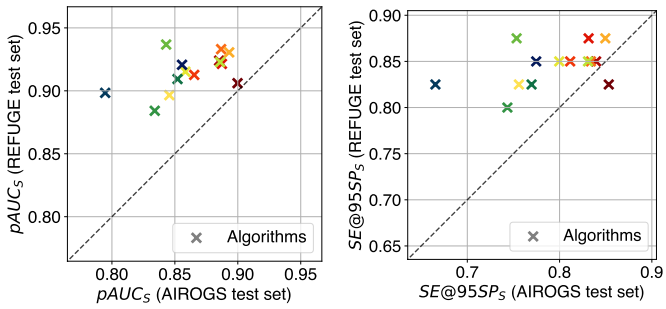
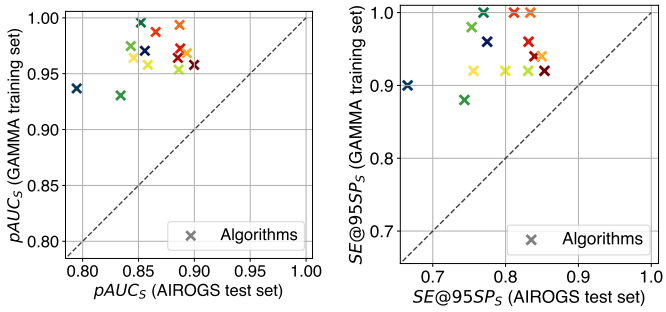


Fig. 5: Comparison of the AIROGS and REFUGE algorithms, tested on the REFUGE test set, visualized as violin and swarm plots. The final algorithms that were developed for the REFUGE challenge itself and for the AIROGS challenge are shown on the left and right, respectively. The AIROGS algorithms were only trained on the AIROGS train set and were not retrained with the REFUGE dataset.



(a) (b)

Fig. 6: Performance of the participating AIROGS algorithms on the REFUGE dataset, compared to their performance on the AIROGS dataset. Both screening metrics (a) $pAUC_S$ and (b) $SE@95SP_S$ are shown.

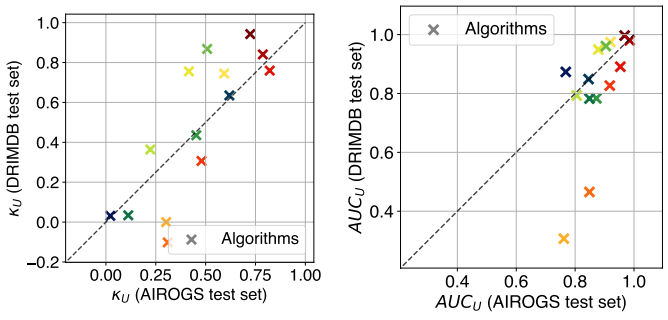


(a) (b)

Fig. 7: Performance of the participating AIROGS algorithms on the GAMMA dataset, compared to their performance on the AIROGS dataset. Both screening metrics (a) $pAUC_S$ and (b) $SE@95SP_S$ are shown.

internal test set performance.

The robustness to ungradable data in the AIROGS test set was evaluated for each team using the metrics κ_U and AUC_U. The teams that performed the best in terms of these metrics achieved 0.82 (95% CI: 0.80 – 0.84) and 0.99 (95% CI:



(a) (b)

Fig. 8: Performance of the participating AIROGS algorithms on the DRIMDB dataset, compared to their performance on the AIROGS dataset. Both robustness metrics (a) κ_U and (b) AUC_U are shown.

0.98 – 0.99) for κ_U and AUC_U, respectively. Human experts did reach a higher κ_U of 0.85 (95% CI: 0.84-0.86) for this task. Moreover, they achieved a sensitivity of 0.95 (95% CI: 0.94 – 0.96) and a specificity of 0.97 (95% CI: 0.97 – 0.97) for detecting ungradable cases, while the team with the best AUC_U achieved a lower sensitivity at 97% specificity of 0.90 (95% CI: 0.88-0.92). Although the teams achieved relatively high performances, they still achieved lower performance at the robustness task than human experts. This shows this task was especially challenging, possibly because the participating teams could not use ungradable development data or because their robustness approaches focused on specific forms of ungradability.

We also assessed robustness on the external DRIMDB dataset. The best-scoring team on this dataset scored very high performances; they achieved 0.94 and 1.0 for κ_U and AUC_U, respectively. These two metrics were lower on the AIROGS dataset for that team. This also indicates very strong generalization to other datasets for the robustness task. The high ungradability detection performance also indicates robustness

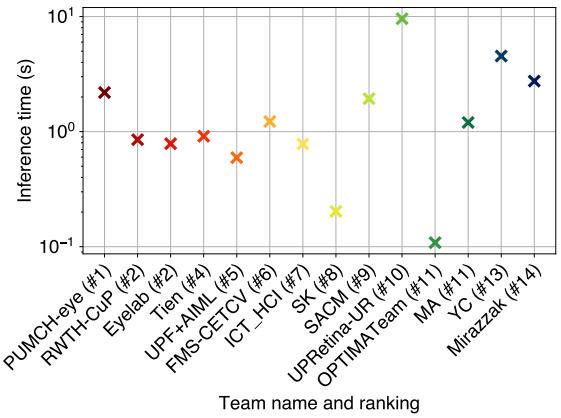


Fig. 9: Average inference time per CFP in the test set of the *Final Test Phase*. This time includes the actual inference time, model initialization, and other setup code executed by the submitted Docker containers.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

to other diseases in the image, as diabetic retinopathy was prevalent in the gradable subset of DRIMDB and the best algorithms did not classify these diseases as ungradable.

A large difference in performance between participating teams can be observed, both for the screening and the robustness task. Therefore, we think it is important to identify which methodological choices were made predominantly by top-performing teams. One of the most notable differences between the top three participants and the rest was the use of transformers. Outside of the top three, only the latest-placed team used a transformer. One of the possible reasons for this superiority that is achieved by transformers compared to CNNs could be their effectiveness at modeling long-range dependencies [79], [80]. This allows for a better understanding of contextual information, which is generally believed to be beneficial in medical imaging [80]. Empirically, transformers and methods that combine transformers with CNNs have previously also been shown to outperform CNNs in medical image analysis [81]–[83], which is in line with our findings.

Moreover, all best three participants manually labeled ODs for training either a segmentation or detection model to crop around the OD during pre-processing. Even though this was also done by two other teams, this seems like an effective strategy to achieve higher screening performance. A likely reason for the effectiveness of this approach is that most glaucoma-related imaging features can be found on or around the OD. This shows how a priori medical knowledge could still be of value even when a large amount of data is available. A less important factor appears to be the number of manually labeled ODs. A possible reason for this could be that the OD detection or segmentation network is not required to be extremely accurate as combining a rough localization of the OD with a large enough padding margin could also suffice to crop the image during pre-processing.

Since the development set only consisted of images that were labeled gradable (either RG or NRG) and the use of external fundus data was prohibited, all teams came up with an uncertainty or OOD detection method based on the gradable data for the robustness task. The ungradability methods of the top three participants in terms of mean position, κ_U, and AUC_U, all revolved around the confidence of a neural network that localized the OD. Of the other participants, only the ninth-placed team had such an approach. Apart from these methods based on OD detection, only team *UPF+AIML* implemented a different robustness technique that was also based on domain knowledge. This raises the impression that solutions based on domain knowledge are more effective for robustness than more general OOD detection solutions. However, it still needs to be evaluated if such approaches are robust for other general tasks (not glaucoma screening) or other sources of OOD data.

For calculating the κ_U metric, the participants were required to output a binary decision on ungradability. A popular approach, especially among the top participants, was to manually identify relatively low-quality images in the development set and base a threshold for this binary output on that subset. This technique was employed by the best three, fifth, tenth, and twelfth teams in terms of κ_U. This indicates that this could be a successful approach, although not in general as the accuracy of this binary value is also highly dependent on the quality of the scalar output for ungradability O_4 that is being thresholded. We found the difference between the ranking in terms of κ_U and AUC_U of one team, in particular, stood out. Team *YC* ranked only eleventh for AUC_U (which depended on the scalar output O_4), but ranked fourth in terms of κ_U (which depended on the binary output O_3), indicating the approach they used for thresholding their scalar value was highly effective. The difference between their AUC_U and κ_U was 7, while the next biggest value of this difference was only 3. Team *YC* indeed came up with a relatively sophisticated method for binarizing O_4 compared to others, based on a Wilcoxon one-sample test to statistically test whether the mean of the predicted probabilities from a Monte-Carlo drop-out approach was 0.5 or not.

B. Strengths and limitations

The dataset presented in this paper substantially exceeds what was publicly available before in terms of number of images and patients. The dataset is also highly diverse because of the large number of different sites, cameras that were used, and ethnicities. Comorbidities were not excluded from the dataset, as our goal was to develop tools that are robust to data with such conditions in real-world screening settings. These conditions should not have an influence on the classification of glaucoma or ungradability. The quality of the labels was controlled by the initial and periodical evaluation of human graders, the fact that each image was independently labeled twice by two trained graders, and, in case of disagreement, by a highly experienced reader. The participants submitted their solutions as containerized algorithms, allowing reproducibility, facilitating inference on other data, and preventing manual manipulation of the test set.

One of the rules of the AIROGS challenge was the prohibition of the use of external fundus data for development. A limitation of this work is the fact that we cannot be sure if any of the teams used such data in their development process. A possible approach to prevent this and make the process fairer is to have participants submit a containerized algorithm for training, which would be trained by the challenge organizers with private challenge training data. Nevertheless, with such an approach it would still be challenging and time-consuming for the challenge organizers to verify if the training containers do not contain any weights pre-trained on other data.

The teams that participated in the competition were permitted to create their own manual annotations on the data and employ them in the development of their models. We made a deliberate decision not to forbid this practice, as we believed that the possibility of achieving superior results outweighed the disadvantage of potentially introducing a slight unfairness due to some teams having access to larger manual labor workforces than others. Teams *PUMCH-eye*, *RWTH-CuP*, *Eyelab*, *UPF+AIML*, *ICT_HCI*, *SACM*, and *YC* took advantage of this opportunity and produced at least one of the subsequent manual annotations: OD detection or segmentation, vessel segmentation, and identification of low-quality images in the development set. As a result, the final rankings of the teams

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

have likely been influenced by this practice, and it is important to consider this potential bias when comparing solutions. We think it is also worth noting that the aforementioned solution for promoting fairness of submitting a containerized algorithm for training on private data would disable manual annotation of development data.

The dataset used for the challenge is diverse, but improvements could still be made in that respect. All screening sites were based across the United States of America, raising the question of whether a more generalizable model could be obtained with data from across the world. On the other hand, we showed that many algorithms trained on the AIROGS dataset performed at least as well on three external test sets, of which two originated from China and one from Turkey, as on our internal test set.

Not all research groups working in the field of retinal image analysis participated in this challenge and many teams that joined the challenge did not submit a solution to the *Final Test Phase*. Possible reasons for this include that many teams saw their results did not match to ones already present on the leaderboard, that the barrier for some teams was too high to get a solution wrapped in a Docker container, or that they were not able to finish in time. Therefore we would like to stress the challenge is still open and we are curious to see if the community can make further improvements. After all, especially for the robustness task, there seems to be room for improvement, given the gap with the human grader performance.

C. Future directions

Based on the solutions that were presented by the teams, we think it would be valuable to combine methodologies from different participants and to work further on their ideas. For example, as we mentioned before, team *YC* apparently had a highly effective method for thresholding their ungradability scores as their κ_U was very high compared to their AUC_U. A possible future direction would be to combine methods of high performance in terms of AUC_U with the binarization technique from team *YC*. Moreover, we observed that algorithms that scored high in terms of robustness, used domain knowledge for this aspect of the challenge. Possible future directions could be to explore other ways to incorporate domain knowledge into an ungradability method. This observation also leads to the question of whether there are more fields in medical image analysis in which domain knowledge can be leveraged for uncertainty estimation and OOD detection.

Next to a decision on RG and NRG presence, the graders were asked to provide which clinical, glaucomatous features were present in the eyes they classified as RG, as listed in Section II-A and further described by [84]. This information was not yet included in the dataset release for this challenge, as it fell outside the scope of this challenge. Future solutions and challenges could be developed with this information, possibly resulting in more explainable algorithms.

This challenge only focused on classification based on a single CFP. It may be interesting to explore the effect on screening performance and robustness of including various types of metadata in our dataset, which we have available but have not been published yet. This metadata, although missing for some images, includes the camera type, age, and anonymous patient identification (which can be used to link two eyes to a single patient).

In order to ensure the safe and effective implementation of the AI models for glaucoma screening described in this paper, several important steps need to be undertaken. González-Gonzalo *et al.* [85] provide valuable insights into the key aspects that are crucial for the integration of AI models in ophthalmic practice.

Among these aspects, additional retrospective validation studies play a significant role in validating the performance and generalization of these models. An external evaluation with substantially different data was already performed in this study. However, we think it is crucial to evaluate with additional large screening datasets that represent real-world scenarios before practical implementation. Prospective validation studies and cost-effectiveness analyses are also essential for evaluating the accuracy, reliability, and generalization of glaucoma screening AI models in real-world settings. These analyses are especially important for screening programs since screening solutions that are not specific enough can have substantial negative financial impacts, as they can lead to unnecessary hospital visits. It is also essential to identify and mitigate potential limitations such as data quality, which the AIROGS challenge aimed to address, model interpretability, integration with screening workflows, and potential biases. Finally, establishing mechanisms for post-market surveillance is vital to monitor and evaluate the performance and safety of these AI models after their regulatory approval.

Further implementation and real-world evaluation of these algorithms are needed, as described above. As this was considered out of the scope of the current manuscript, we leave the execution of these steps to future work.

VII. Conclusions

We presented the results of community-acquired algorithms tested on real-world data for robust glaucoma screening from CFP. The best algorithms performed similarly in terms of screening to the carefully trained and selected human graders, and were shown to be effective at flagging images that could not be graded. Methodological choices predominantly made by the best teams included, for the screening task, the use of vision transformers and the incorporation of optic disc detection models in pre-processing and, for the robustness task, out-of-distribution detection approaches based on domain knowledge. We hope the unprecedented size and real-world nature of the dataset we released and the algorithms that were developed using this dataset will help towards implementing robust AI for glaucoma screening.

References

[1] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

[2] P. Mokhles, J. S. Schouten, H. J. Beckers, A. Azuara-Blanco, A. Tuulonen, and C. A. Webers, "A systematic review of end-of-life visual impairment in open-angle glaucoma: an epidemiological autopsy," *Journal of Glaucoma*, vol. 25, no. 7, pp. 623–628, 2016.

[3] P. J. Ernest, M. J. Busch, C. A. Webers, H. J. Beckers, F. Hendrikse, M. H. Prins, and J. S. Schouten, "Prevalence of end-of-life visual impairment in patients followed for glaucoma," *Acta Ophthalmologica*, vol. 91, no. 8, pp. 738–743, 2013.

[4] D. Peters, B. Bengtsson, and A. Heijl, "Lifetime risk of blindness in open-angle glaucoma," *American Journal of Ophthalmology*, vol. 156, no. 4, pp. 724–730, 2013.

[5] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: a review," *JAMA*, vol. 311, no. 18, pp. 1901–1911, 2014.

[6] X. Chen, Y. Xu, F. Yin, Z. Zhang, D. W. K. Wong, T. Y. Wong, and J. Liu, "Multiple ocular diseases detection based on joint sparse multi-task learning," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 5260–5263.

[7] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee *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*, vol. 318, no. 22, pp. 2211–2223, 2017.

[8] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, 2018.

[9] S. Phene, R. C. Dunn, N. Hammel, Y. Liu, J. Krause, N. Kitade, M. Schaekermann, R. Sayres, D. J. Wu, A. Bora *et al.*, "Deep learning and glaucoma specialists: the relative importance of optic disc features to predict glaucoma referral in fundus photographs," *Ophthalmology*, vol. 126, no. 12, pp. 1627–1639, 2019.

[10] T. W. Rogers, N. Jaccard, F. Carbonaro, H. G. Lemij, K. A. Vermeer, N. J. Reus, and S. Trikha, "Evaluation of an ai system for the automated detection of glaucoma from stereoscopic optic disc photographs: the european optic disc assessment study," *Eye*, vol. 33, no. 11, pp. 1791–1797, 2019.

[11] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[12] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "Origa-light: An online retinal fundus image database for glaucoma analysis and research," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 3065–3068.

[13] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "Rim-one: An open retinal image database for optic nerve evaluation," in *24th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2011, pp. 1–6.

[14] J. Odstrcilik, R. Kolar, A. Budai, J. Hornegger, J. Jan, J. Gazarek, T. Kubena, P. Cernosek, O. Svoboda, and E. Angelopoulou, "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database," *IET Image Processing*, vol. 7, no. 4, pp. 373–383, 2013.

[15] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[16] S. Holm, G. Russell, V. Nourrit, and N. McLoughlin, "Dr hagis—a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients," *Journal of Medical Imaging*, vol. 4, no. 1, p. 014503, 2017.

[17] J. I. Orlando, J. Barbosa Breda, K. v. Keer, M. B. Blaschko, P. J. Blanco, and C. A. Bulant, "Towards a glaucoma risk index based on simulated hemodynamics from fundus images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 65–73.

[18] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical image analysis*, vol. 59, p. 101570, 2020.

[19] H. Fang, F. Li, H. Fu, X. Sun, X. Cao, J. Son, S. Yu, M. Zhang, C. Yuan, C. Bian *et al.*, "Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment," *arXiv preprint arXiv:2202.08994*, 2022.

[20] J. Wu, H. Fang, F. Li, H. Fu, F. Lin, J. Li, L. Huang, Q. Yu, S. Song, X. Xu *et al.*, "Gamma challenge: glaucoma grading from multi-modality images," *arXiv preprint arXiv:2202.06511*, 2022.

[21] J. Cuadros and G. Bresnick, "Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening," *Journal of diabetes science and technology*, vol. 3, no. 3, pp. 509–516, 2009.

[22] N. J. Reus, H. G. Lemij, D. F. Garway-Heath, P. J. Airaksinen, A. Anton, A. M. Bron, C. Faschinger, G. Holló, M. Iester, J. B. Jonas *et al.*, "Clinical assessment of stereoscopic optic disc photographs for glaucoma: the european optic disc assessment trial," *Ophthalmology*, vol. 117, no. 4, pp. 717–723, 2010.

[23] H. G. Lemij, C. de Vente, C. I. Sánchez, and K. A. Vermeer, "Characteristics of a large, labeled dataset for the training of artificial intelligence for glaucoma screening with fundus photographs," *Ophthalmology Science*, 2023.

[24] U. Sevik, C. Kose, T. Berber, and H. Erdol, "Identification of suitable fundus images using automated quality assessment methods," *Journal of biomedical optics*, vol. 19, no. 4, p. 046006, 2014.

[25] D. K. McClish, "Analyzing a portion of the roc curve," *Medical decision making*, vol. 9, no. 3, pp. 190–195, 1989.

[26] H. Vaahtoranta-Lehtonen, A. Tuulonen, P. Aronen, H. Sintonen, L. Suoranta, N. Kovanen, M. Linna, E. Läärä, and A. Malmivaara, "Cost effectiveness and cost utility of an organized screening programme for glaucoma," *Acta Ophthalmologica Scandinavica*, vol. 85, no. 5, pp. 508–518, 2007.

[27] M. M. de Vries, R. Stoutenbeek, R. P. Müskens, and N. M. Jansonius, "Glaucoma screening during regular optician visits: the feasibility and specificity of screening in real life," *Acta ophthalmologica*, vol. 90, no. 2, pp. 115–121, 2012.

[28] C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data," *Academic radiology*, vol. 7, no. 6, pp. 413–419, 2000.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[31] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[33] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

[34] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[35] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[36] S. G. Müller and F. Hutter, "Trivialaugment: Tuning-free yet state-of-the-art data augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 774–782.

[37] H. Wang, H. Sun, Y. Fang, S. Li, M. Feng, and R. Wang, "A workflow for computer-aided diagnosis of glaucoma," in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE, 2022, pp. 1–4.

[38] F. Khader, C. Haarburger, J.-C. Kirr, M. Menke, J. N. Kather, J. Stegmaier, C. Kuhl, S. Nebelung, and D. Truhn, "Elevating fundoscopic evaluation to expert level-automatic glaucoma detection using data from the airogs challenge," in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE, 2022, pp. 1–4.

[39] T. Araújo, G. Aresta, and H. Bogunović, "Deep dirichlet uncertainty for unsupervised out-of-distribution detection of eye fundus photographs in glaucoma screening," in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE, 2022, pp. 1–5.

[40] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[41] G. Sharir, A. Noy, and L. Zelnik-Manor, "An image is worth 16x16 words, what is a video worth?" *arXiv preprint arXiv:2103.13915*, 2021.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

[42] W. Liu, H. Yang, T. Tian, Z. Cao, X. Pan, W. Xu, Y. Jin, and F. Gao, "Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4623–4634, 2022.

[43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[44] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[45] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[46] ——, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[47] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, F. Ingham, J. Poznanski, J. Fang, L. Yu *et al.*, "ultralytics/yolov5," *Zenodo*, 2020.

[48] T. Aimyshev and K. Eleusiz. (2022) Glaucoma detection algorithm for the artificial intelligence for robust glaucoma screening challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/34b337f9-551a-4c82-834a-32a9f99ba690/AIROGS.pdf

[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[50] T.-D. Le. (2022) Combination of supervised learning and unsupervised learning to detect ungradable images in the airogs challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/3e9f9bbb-9d22-430f-9448-6a4e21a29f2f/Airogs_Combination_Algorithm.pdf

[51] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[52] A. Galdran, G. Carneiro, and M. A. G. Ballester. (2022) Open-set glaucoma screening from eye fundus images: Domain knowledge to the rescue. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/e01b1325-be85-4d1e-9322-c0c3726ab9ae/airogs_osr.pdf

[53] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[54] D. Puthussery, P. S. Hrishikesh, R. G. Devika, and C. V. Jiji. (2022) A self-supervised approach for glaucoma screening. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/fd5a5254-8faf-42df-97ec-b500c35c7c2a/A_Self-Supervised_appr_GsyqCdA.pdf

[55] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[56] P. Oza and V. M. Patel, "One-class convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 277–281, 2018.

[57] Z. Yang, H. Liu, and Z. Shang. (2022) Deep learning for referable glaucoma screening and out-of-distribution detection. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/d259c884-5eda-4bb9-b9fb-73a29e9eaf41/airogs.pdf

[58] S. Kondo, S. Kasai, and K. Hirasawa, "Computer aided diagnosis and out-of-distribution detection in glaucoma screening using color fundus photography," *arXiv preprint arXiv:2202.11944*, 2022.

[59] I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph, "Revisiting resnets: Improved training and scaling strategies," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 614–22 627, 2021.

[60] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 464–21 475, 2020.

[61] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.

[62] E. Wang, A. Durvasula, D. Deng, A. Sivajohan, E. Ho, and K. Lane. (2022) Ensemble network for glaucoma screening in airogs challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/ea161fad-ca7f-4e1a-9133-bed2128107c3/AIROGS_Manuscript.pdf

[63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[65] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[66] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[67] J. Heras, D. Royo, and M. A. Zapata. (2022) A good closed-set classifier is all you need for the airogs challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/d79d3e55-505a-416a-b389-0a51170b1271/AIROGS.pdf

[68] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," *arXiv preprint arXiv:2110.06207*, 2021.

[69] A. P. Dempster, "A generalization of bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 205–232, 1968.

[70] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[71] M. Arikan. (2022) Multi-model ensemble for robust glaucoma screening. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/c9b4a350-62f4-4f80-aa16-cfc71b6640c3/Multi_model_ensemble_f_dsy4aS0.pdf

[72] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

[73] Y. C. Lee, H. B. Cho, and Y. H. Choi. (2022) Classification for referable glaucoma with fundus photographs using multimodal deep learning. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/7b0a4c21-e8b4-4fee-9e01-f48e25b2b1b4/Classification_for_ref_sFDsTWh.pdf

[74] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[75] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[76] A. Qayyum, M. Mazher, and I. Razzak. (2022) ConvNeXts and vision transformer based framework for glaucoma screening. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/33dea053-4ae5-4ac4-8772-f4599c6af590/ISIB_AIROGS__Imran.pdf

[77] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[78] K. Bibas, M. Feder, and T. Hassner, "Single layer predictive normalized maximum likelihood for out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1179–1191, 2021.

[79] T. Lei, R. Wang, Y. Wan, X. Du, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey." 2020.

[80] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Medical image analysis*, p. 102762, 2023.

[81] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 730–20 740.

[82] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 2023, pp. 205–218.

[83] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313786

segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[84] H. G. Lemij, C. de Vente, C. Sánchez, J. Cuadros, N. Jaccard, and K. Vermeer, "Glaucomatous features in fundus photographs of eyes with 'referable glaucoma' of a large population based labeled data set for training an artificial intelligence (AI) algorithm for glaucoma screening," in *Association for Research in Vision and Ophthalmology*, vol. 63, no. 7. The Association for Research in Vision and Ophthalmology, 2022, pp. 2041–A0482.

[85] C. González-Gonzalo, E. F. Thee, C. C. Klaver, A. Y. Lee, R. O. Schlingemann, A. Tufail, F. Verbraak, and C. I. Sánchez, "Trustworthy ai: Closing the gap between development and integration of ai systems in ophthalmic practice," *Progress in retinal and eye research*, vol. 90, p. 101034, 2022.

ACKNOWLEDGEMENT

Coen de Vente and Clara I. Sánchez are with the Quantitative Healthcare Analysis (QurAI) Group, Informatics Institute, Universiteit van Amsterdam, Amsterdam, Noord-Holland, Netherlands and the Department of Biomedical Engineering and Physics, Amsterdam UMC Locatie AMC, Amsterdam, Noord-Holland, Netherlands (e-mail: research@coendevente.com).

Coen de Vente and Bram van Ginneken are with the Diagnostic Image Analysis Group (DIAG), Department of Radiology and Nuclear Medicine, Radboudumc, Nijmegen, Gelderland, Netherlands.

Koenraad A. Vermeer and Hans G. Lemij are with the Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Rotterdam, Netherlands.

Nicolas Jaccard is with Project Orbis International Inc., New York, United States.

He Wang is with the Peking Union Medical College Hospital, Beijing, China and with the Xuanwu Hospital Capital Medical University, Beijing, 100053, China.

Hongyi Sun is with the Tsinghua University, Beijing, China.

Firas Khader and Daniel Truhn are with the Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany.

Temirgali Aimyshev and Yerkebulan Zhanibekuly are with CMC Technologies LLP, Nur-Sultan, Kazakhstan.

Tien-Dung Le is with KBC, Belgium.

Adrian Galdran and Miguel Ángel González-Ballester are with Universitat Pompeu Fabra, Barcelona, Spain.

Adrian Galdran and Gustavo Carneiro are with the Australian Institute for Machine Learning AIML, University of Adelaide, Australia.

Miguel Ángel González-Ballester is also with ICREA, Barcelona, Spain.

Gustavo Carneiro is also with the Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom.

Devika R G is with the College of Engineering, Trivandrum, India.

Hrishikesh P S and Densen Puthussery are with Founding Minds Software, India.

Hong Liu and Zekang Yang are with the Institute of Computing Technology, Chinese Academy of Sciences.

Satoshi Kondo is with the Muroran Institute of Technology, Japan.

Satoshi Kasai is with the Niigata University of Health and Welfare, Japan.

Edward Wang and Ashritha Durvasula are with the Schulich School of Medicine and Dentistry, University of Western Ontario, London, Canada.

Jónathan Heras is with the Department of Mathematics and Computer Science, University of La Rioja, Spain.

Miguel Ángel Zapata is with Hospital Vall Hebron, Passeig Roser 126, Sant Cugat del Vallès, 08195 Barcelona, Spain and UPRetina, Barcelona, Spain.

Teresa Araújo, Guilherme Aresta and Hrvoje Bogunović are with the Christian Doppler Laboratory for Artificial Intelligence in Retina, Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria.

Mustafa Arikan is with the Institute of Ophthalmology, University College London, London, United Kingdom.

Yeong Chan Lee is with the Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea.

Hyun Bin Cho and Yoon Ho Choi are with the Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Samsung Medical Center, Seoul, Republic of Korea.

Yoon Ho Choi is also with the Department of Artificial Intelligence and Informatics, Mayo Clinic, United States of America, Florida, Jacksonville.

Abdul Qayyum is with the Department of Biomedical Engineering, King's College London, UK.

Imran Razzak is with the University of New South Wales, Sydney, Australia.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/