Justified Referral in AI Glaucoma Screening: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Justified Referral in AI Glaucoma Screening

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

JustRAIGS

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Glaucoma is a leading cause of irreversible blindness and impaired vision. In its early stages, the disease is typically asymptomatic. With more advanced glaucoma, the visual field is affected; as a result, patients stumble more often, bump into objects and other people, and may be more often involved in traffic accidents and falls. Only in the late stages of the disease, are patients more aware of their visual impairment. They may experience trouble reading, suffer from night-blindness, or suffer from other symptoms of impaired vision. Once detected, glaucoma can be treated so that any disease progression be effectively stopped or slowed down, but the damage cannot be repaired. Early detection and timely treatment of this disease can therefore avoid visual impairment; early detection could be facilitated through population-based glaucoma screening. Glaucoma affects the optic nerve, i.e., the connection between the eye and the brain; this disease is also known as glaucomatous optic neuropathy (GON). It is typically identified based on the appearance of the optic nerve head and its surroundings, for instance on color fundus photographs (CFPs) or other imaging modalities. In clinical practice, one other imaging technique is optical coherence tomography (OCT), which plays an ever-growing role in the diagnosis and follow-up of GON. For screening purposes, however, CFPs are relatively inexpensive. These photographs provide crucial information for assessing various features of glaucomatous damage. These features include neuroretinal rim thinning and notching, increased cupping and optic disc hemorrhages. In addition, glaucomatous thinning of the retinal nerve fiber layer (RNFL) may be readily visible on CFPs. Furthermore, CFPs have an additional benefit in that they provide a record of the eye's baseline condition, serving as a reference for future follow-up.
Manual identification of these features can provide higher accuracy when performed by experienced specialists. However, manual segmentation is subjective and can vary among different observers. Automated detection algorithms, on the other hand, can provide consistent and reproducible results and subsequently reduce inter-observer and intra-observer variability. Manual segmentation can be also time-consuming and labor-intensive, especially for large datasets or complex cases. On the other hand, automated algorithms can process images more rapidly, thus may provide efficient solutions for large scale screening. Artificial intelligence (AI) approaches for detecting glaucoma based on CFPs have been extensively investigated previously and have

provided promising results. In the context of screening, for low prevalence diseases such as glaucoma, specificity is of primary importance and should be very high in order to prevent referring many false positive cases to the health care system. Therefore, the model should be highly dependable and provide clinically relevant outcomes. However, current AI methods merely indicate whether an individual requires to be referred to an ophthalmologist or not, but do not provide any justification for the underlying pathology. Understanding the typically glaucomatous features that the algorithm suggests for referring an individual improves trust as well as enables human experts to identify errors in the decision process due to physiological or pathological deviations.
To initiate the development of such AI algorithms for glaucoma screening and to evaluate their performance, we propose the Justified Referral in AI Glaucoma Screening (JustRAIGS) challenge, for which we have provided a unique large dataset with over 110k carefully annotated fundus photographs collected from about 60,000 screenees. We have generated a training subset with 101,442 gradable fundus images (from 'referable glaucoma 'eyes and 'no referable glaucoma' eyes) and a test subset with 9,741 fundus images. Each fundus photograph thus has been labeled as either 'referable glaucoma' or 'no referable glaucoma'. In addition, all fundus images of referable glaucoma eyes have been further annotated with up to ten additional labels associated with different glaucomatous features. In this challenge, participants will be tasked with analyzing the fundus images and assigning each image to one of two classes: 'referable glaucoma' or 'no referable glaucoma'. 'Referable glaucoma' refers to eyes where the fundus image exhibits signs or features indicative of glaucoma that require further examination or referral to a specialist. In this case, visual field damage is expected. On the other hand, 'no referable glaucoma' refers to cases where the fundus image does not show significant indications of glaucoma and does not require immediate referral. Very early disease, in which visual field damage is not yet expected, would also be classified as 'no referable glaucoma_. In addition to the referable glaucoma classification, participants will be further instructed to perform multi-label classification for ten additional features related to glaucoma. These features are specific characteristics or abnormalities that may be present in the fundus images of glaucoma patients. The multi-label classification task involves assigning relevant labels to each fundus image based on the presence or absence of these specific features. These additional features provide more detailed information about the specific characteristics observed in the fundus images of 'referable glaucoma' cases. By combining both the binary classification task (referable vs. no referable glaucoma) and the multi-label classification task (for the ten additional features), we aim to evaluate the participants' ability to accurately identify and classify fundus images associated with referable glaucoma. The results of this classification task can provide insights into the development of automated systems or algorithms for glaucoma detection, ultimately assisting in the early identification and treatment of glaucoma patients, thereby reducing avoidable visual impairment and blindness from glaucoma.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Glaucoma, screening, fundus images, referral, justification, ophthalmology, Artificial Intelligence (AI) model

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

N/A

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

/

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

/

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

N/A

TASK 1: JustRAIGS

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Glaucoma is a leading cause of irreversible blindness and impaired vision. In its early stages, the disease is typically asymptomatic. With more advanced glaucoma, the visual field is affected; as a result, patients stumble more often, bump into objects and other people, and may be more often involved in traffic accidents and falls. Only in the late stages of the disease, are patients more aware of their visual impairment. They may experience trouble reading, suffer from night-blindness, or suffer from other symptoms of impaired vision. Once detected, glaucoma can be treated so that any disease progression be effectively stopped or slowed down, but the damage cannot be repaired. Early detection and timely treatment of this disease can therefore avoid visual impairment; early detection could be facilitated through population-based glaucoma screening. Glaucoma affects the optic nerve, i.e., the connection between the eye and the brain; this disease is also known as glaucomatous optic neuropathy (GON). It is typically identified based on the appearance of the optic nerve head and its surroundings, for instance on color fundus photographs (CFPs) or other imaging modalities. In clinical practice, one other imaging technique is optical coherence tomography (OCT), which plays an ever-growing role in the diagnosis and follow-up of GON. For screening purposes, however, CFPs are relatively inexpensive. These photographs provide crucial information for assessing various features of glaucomatous damage. These features include neuroretinal rim thinning and notching, increased cupping and optic disc hemorrhages. In addition, glaucomatous thinning of the retinal nerve fiber layer (RNFL) may be readily visible on CFPs. Furthermore, CFPs have an additional benefit in that they provide a record of the eye's baseline condition, serving as a reference for future follow-up.
Manual identification of these features can provide higher accuracy when performed by experienced specialists. However, manual segmentation is subjective and can vary among different observers. Automated detection algorithms, on the other hand, can provide consistent and reproducible results and subsequently reduce inter-observer and intra-observer variability. Manual segmentation can be also time-consuming and labor-intensive, especially for large datasets or complex cases. On the other hand, automated algorithms can process images more rapidly, thus may provide efficient solutions for large scale screening. Artificial intelligence (AI) approaches for detecting glaucoma based on CFPs have been extensively investigated previously and have provided promising results. In the context of screening, for low prevalence diseases such as glaucoma, specificity is of primary importance and should be very high in order to prevent referring many false positive cases to the health care system. Therefore, the model should be highly dependable and provide clinically relevant outcomes. However, current AI methods merely indicate whether an individual requires to be referred to an ophthalmologist or not, but do not provide any justification for the underlying pathology. Understanding the typically glaucomatous features that the algorithm suggests for referring an individual improves trust as well as enables human experts to identify errors in the decision process due to physiological or pathological deviations.
To initiate the development of such AI algorithms for glaucoma screening and to evaluate their performance, we propose the Justified Referral in AI Glaucoma Screening (JustRAIGS) challenge, for which we have provided a unique large dataset with over 110k carefully annotated fundus photographs collected from about 60,000 screenees. We have generated a training subset with 101,442 gradable fundus images (from 'referable glaucoma

'eyes and 'no referable glaucoma' eyes) and a test subset with 9,741 fundus images. Each fundus photograph thus has been labeled as either 'referable glaucoma' or 'no referable glaucoma'. In addition, all fundus images of referable glaucoma eyes have been further annotated with up to ten additional labels associated with different glaucomatous features. In this challenge, participants will be tasked with analyzing the fundus images and assigning each image to one of two classes: 'referable glaucoma' or 'no referable glaucoma'. 'Referable glaucoma' refers to eyes where the fundus image exhibits signs or features indicative of glaucoma that require further examination or referral to a specialist. In this case, visual field damage is expected. On the other hand, 'no referable glaucoma' refers to cases where the fundus image does not show significant indications of glaucoma and does not require immediate referral. Very early disease, in which visual field damage is not yet expected, would also be classified as 'no referable glaucoma_. In addition to the referable glaucoma classification, participants will be further instructed to perform multi-label classification for ten additional features related to glaucoma. These features are specific characteristics or abnormalities that may be present in the fundus images of glaucoma patients. The multi-label classification task involves assigning relevant labels to each fundus image based on the presence or absence of these specific features. These additional features provide more detailed information about the specific characteristics observed in the fundus images of 'referable glaucoma' cases. By combining both the binary classification task (referable vs. no referable glaucoma) and the multi-label classification task (for the ten additional features), we aim to evaluate the participants' ability to accurately identify and classify fundus images associated with referable glaucoma. The results of this classification task can provide insights into the development of automated systems or algorithms for glaucoma detection, ultimately assisting in the early identification and treatment of glaucoma patients, thereby reducing avoidable visual impairment and blindness from glaucoma.

Keywords

List the primary keywords that characterize the task.

Glaucoma, screening, fundus images, referral, justification, ophthalmology, Artificial Intelligence (AI) model

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

1. Koenraad A. Vermeer: Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Rotterdam, The Netherlands
2. Hans G. Lemij: Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Rotterdam, The Netherlands
3. Siamak Yousefi, Department of Ophthalmology, Department of Genetics, Genomics, and Informatics, Director of the Data Mining and Machine Learning (DM2L) Laboratory, University of Tennessee Health Science Center, Memphis, USA
4. Yeganeh Madadi, Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, USA
5. Hina Raja, Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, USA

b) Provide information on the primary contact person.

ymadadi@uthsc.edu
hraja@uthsc.edu
Siamak.Yousefi@uthsc.edu

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Open call challenge.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

IEEE - ISBI 2024

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

https://justraigs.grand-challenge.org/

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

No additional data allowed.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Cash prizes:
1st rank: € 3000,
2nd rank: € 2000,
3rd rank: € 1000.
Participants are only eligible for the challenge prizes if they:
- Make their code open source.
- Are willing to write a (short) paper describing the key methods of their approach
- Participate in the ISBI 2024 meeting and will be available in person to present during the JustRAIGS challenge

session

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The best result of the submissions of each contributor will be made public. The ranking of the top 3 performing methods will be announced publicly during the conference, subject to the eligibility criteria listed above.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All participating teams may publish their own results. Participants who produced the top performing algorithms or who's approach is of specific interest will also be invited to co-author on a publication about the results of this challenge.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants will upload their Docker container through the challenge website https://justraigs.grand-challenge.org/.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Phase 1: Participants may submit up to 10 Docker containers with different, modified or updated algorithms during the phase 1. They will all be listed in the results, but only the best one is used in the ranking.
Phase 2: Participants may submit only 1 Docker container during the entire phase.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)

- associated workshop days (if any)
- the release date(s) of the results

November 5, 2023: Challenge website launched.
November 30, 2023: Pre-evaluation system ready
January 8,2024: Registration
January 8, 2024: Phase 1: Development Phase opens: The training dataset is released. Evaluation will be done on only 10% of test dataset.
February 29 2024: Phase 2:Test Phase opens: Evaluation will be done on 100% of test dataset.
April 6, 2024: Submission of paper to ISBI 2024 closes.
April 20, 2024: Participant will receive the reviews for the paper submission. Participants will have only one week to address those comments.
April 27, 2024: Submission of camera-ready paper to CMT
May 6, 2024: Notification of oral versus poster presenters - Note that Oral papers are top-performers & interesting methods
May 27 ISBI 2024: JustRAIGS session with presentations and reward ceremony.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

No ethics approval is needed. The data is provided by a third party and they have obtained the required consent for the proposed use of the data.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY-NC-SA

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The evaluation software is not (yet) available.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Participants must release their source code on an open-source platform, such as GitHub (https://www.github.com) in order to eligible for prize money.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

There is no conflict of interest, because any affiliates of the organizers are not eligible for the challenge. Access to the test case labels is limited to the organizing team (Siamak Yousefi, Hans Lemij, Koen Vermeer, Yeganeh Madadi, Hina Raja).

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Prevention, Decision support, Screening, Diagnosis.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling

- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification, Detection.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort is the general public. Most likely, screening would be targeted at those with certain risk factors, such as age over 50 years, family history of glaucoma, ethnicity and certain medical conditions such as diabetes and myopia (short-sightedness).

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The data was acquired from a screening population for diabetic retinopathy. This population closely resembles the target population.
Diabetes is generally considered to be a risk factor for glaucoma. As a result, the prevalence of glaucoma is likely to be somewhat higher in the challenge cohort than in the final cohort. This may introduce some bias with an average relative risk of approximately 1.4 (Zhou M et al. PLoS One. 2014 Aug, 19;9(8):e102972. doi: 10.1371/journal.pone.0102972.). However, the evaluation metrics are robust to the prevalence. The glaucomatous features in fundus photographs, however, are the same, at least clinically, so we think that any bias for this particular study is acceptably small.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Color fundus photography

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Classification:
- Referable glaucoma or no referable glaucoma

Additional labels:
- Appearance neuroretinal rim superiorly
- Appearance neuroretinal rim inferiorly
- Retinal nerve fiber layer defect superiorly
- Retinal nerve fiber layer defect inferiorly
- Baring circumlinear vessel superiorly
- Baring circumlinear vessel inferiorly
- Nasalisation of vessel trunk
- Disc haemorrhage
- Laminar dots
- Large cup

b) ... to the patient in general (e.g. sex, medical history).

Age in years

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Color fundus photograph of the retina, covering the optic nerve head, its surroundings and the macula.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Glaucomatous features in the color fundus image that are indicative of visual field loss.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Sensitivity, Specificity, Area Under the Receiver Operating Characteristic Curve (AUROC), Modified Hamming loss

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g.

tracking system used in a surgical setting).

Various color fundus cameras were used to acquire the data. The data were acquired in approximately 500 screening centers for diabetic retinopathy. Although specific information on the manufacturer and model of the color fundus cameras is not available, these cameras are representative of the type of devices employed in screening centers.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

All images were acquired in a real-world screening setting for diabetic retinopathy, in approximately 500 centers in the USA. The data acquisition was not controlled for this challenge. Instead, these images represent the images that are acquired in these screening settings.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

The image data was provided by EyePACS LLC, Santa Cruz, CA, USA; labels were provided by the Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Rotterdam, The Netherlands. The color fundus photographs were acquired in approximately 500 screening centers across the USA on a large variety of cameras. Per eye, 3 images were taken, to reduce the risk that an eye could not be properly judged because the image was poorly aligned, over- or underexposed, or otherwise ungradable because of a closed eyelid or media opacities.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

The images were acquired by trained color fundus camera operators. These operators were not trained specifically for generating the data set for this challenge. The images therefore represent the real-world range of images that are acquired in such a screening setting. The dataset was multi-ethnic and included people of African descent (6%), Caucasians (8%), Asians (4%), Latin Americans (52%), native Americans (1%), people from the Indian subcontinent (3%) and people of mixed ethnicity (1%) as well as people of unspecified ethnicity (25%). Such information on ethnicity is, however, not available per image or subject. The participants' mean age (SD) was 57.1 (10.4) years. The aim of each image was that the field of view contained both the optic nerve head and the macula. Any signs of coexisting eye disease, e.g., diabetic retinopathy, were not specifically labeled.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case consists of one eye of one person and the corresponding labels (classification (referable glaucoma/no referable glaucoma) and additional labels justifying the referral) provided by the graders. In addition, the age of the person is provided. In the training data, a pseudonymized person ID is also available to identify eyes of the same person.

b) State the total number of training, validation and test cases.

101,442 training cases and 9,741 test cases. Participants may split the provided training set into training and validation cases as they see fit.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The full data set contains 111,183 images. 9,741 test cases are selected to be able to assess the specificity with a 95% CI of 0.5% and the sensitivity with a 95% CI of 2%.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

The class distribution in the training set is representative for the general population (prevalence of 3.2% in the training set). In the test set, the class distribution is somewhat more balanced (16%); this does not affect the evaluation results since sensitivity and specificity are reported.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Each image was annotated by 2 graders, randomly selected from a pool of qualified graders. If they agreed in their main classification (referable glaucoma (RG), no referable glaucoma (NRG) or ungradable (UG)), this classification was considered as final; in case of disagreement, however, the image was subsequently graded by 1 of 2 glaucoma specialists; their classification was considered as final. Note that the ungradable images are not considered in this challenge.
When graders selected 'referable glaucoma' as their main classification, they were asked to check boxes corresponding to the reasons why they felt the patient should be referred. There were 10 options available; graders could select as many as they thought to be applicable to the image shown. Disagreement on these additional labels was not resolved. If for any of the additional labels there was disagreement, that label will be ignored during evaluation of the algorithms.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Prospective graders were first asked to do an entry exam. If they passed this exam, they received instructions on how to access the grading website. They could first get acquainted with the website in a practice environment. Once they felt comfortable with the system and the grading interface, they could access the working environment. All graders received a manual (not available online) with instructions on how to use the grading interface. The main task for the graders was to provide a main classification per image (RG, NRG or UG); in case they selected RG, they were required to select up to 10 additional labels why they considered the eye as referable glaucoma.

These labels reflected typical features of glaucomatous damage.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Optic disc assessment training courses (targeted at detecting glaucoma) had been given for several years by one of the organizers (HL), an experienced glaucoma specialist. The audience of these courses consisted of general ophthalmologists, glaucoma specialists, residents in ophthalmology and optometrists. Upon a call, 90 of these wanted to become a grader. All of these had to pass an exam; those that passed (n = 32) all showed a minimum specificity of 92% and a minimum sensitivity of 85% for detecting glaucoma on fundus photographs. During the grading process, the performance of all graders was monitored; those that showed a sensitivity lower than 80% and/or a specificity lower than 95% were removed from the graders pool. All the images they had graded were redistributed amongst the remaining graders. In the end, the pool consisted of 22 graders of which 2 were experienced glaucoma specialists that only graded the images of which the main classification was inconclusive (i.e., there was disagreement between the first two graders).

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

If the initial two graders agreed in their main classification (referable glaucoma (RG), no referable glaucoma (NRG) or ungradable (UG), this classification was considered as final; in case of disagreement, however, the image was subsequently graded by 1 of 2 glaucoma specialists; their classification was considered as final. Note that the ungradable images are not considered in this challenge.
Disagreement on the additional labels was not resolved. Instead, the ground truth of the additional labels is based on both the initial graders (if they agreed on the main classification), on one grader and the glaucoma specialist (if one of the graders agreed on the main classification provided by the specialist) or on the glaucoma specialist (if none of the graders agreed on the main classification provided by the specialist). If for any of the additional labels there was disagreement, that label will be ignored for that image during the evaluation of the algorithms.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

All images were initially graded by two graders from a larger pool; if both graders agreed on the classification, that label became final. If both selected the wrong label, this implies that the final label is also wrong. To reduce this as much as possible, we monitored the performance of each grader (against the final label). If graders performed below 80% sensitivity or 95% specificity, they were eventually removed from the pool of graders and their gradings were redone. Note that graders were not operating in fixed pairs; instead, a new pair of graders was randomly selected for each block of 200 images to be graded.

For the additional labels, no resolution for possible disagreement was sought.

b) In an analogous manner, describe and quantify other relevant sources of error.

Not applicable.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Two metrics will be used: for the referral performance and for the justification performance.
Referral performance (P_ref) will be assessed based on the sensitivity at 95% specificity. Based on the algorithm's output, we will determine the operating point for 95% specificity and assess the corresponding sensitivity.
Justification performance (P_just): For the referable glaucoma cases, the 10 additional labels will be compared against the additional labels produced by the algorithm. We use a modified Hamming distance: If the manual graders did not agree on one or more of the additional labels, the algorithm's result will not be evaluated on those labels. Normalization of the Hamming distance will be done based on the number of labels that both graders agreed on.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

A specificity of (at least) 95% is needed to ensure that the post predictive value (PPV) of the screening test is large enough, given the low prevalence of glaucoma in the general population.
The modified Hamming distance is proposed to be able to handle the possible disagreement between manual graders, which was not resolved in the annotation process. Without normalization, the maximum distance for images with disagreement between graders would be smaller than that for images without disagreement.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Submissions will be ranked according to each metric, resulting in a ranking R_1 based on referral performance P_ref and a ranking R_2 based on justification performance P_just.
Aggregation will be done by calculating a final, combined ranking:
Each submission will get a score of R_1+R_2 and will be ranked again according to this score.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Any missing result is considered to be an error and will be scored accordingly.

c) Justify why the described ranking scheme(s) was/were used.

The final ranking is done based on the ranking of both metrics, since both metrics relate to different relevant performance measures of the algorithms.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We do not need to handle any missing data - any missing data is considered to be an error and will be penalized.
The size of the test set was chosen in order to be able to evaluate the sensitivity with a 95% CI of at most 0.5% (assuming that the sensitivity is 80%) and the specificity with a 95% CI of at most 2% (assuming that the specificity is 95%), as calculated by the Wilson score interval.
For the initial analysis of the challenge data, we do not make any assumptions to perform any of the statistical approaches, which only included counting the number of patients and eyes (in total and separately for the referable and no referable groups).
For data analysis, we used Python and NumPy.

b) Justify why the described statistical method(s) was/were used.

Since there was no missing data, we also did not need to handle this.
The Wilson score interval is a commonly used method to determine a binomial confidence interval. It results in asymmetrical CIs, which is especially relevant for proportions close to 0 or 1.
Python and NumPy are commonly used and open-source tools.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

We will analyze which (type of) images were often classified incorrectly by the algorithms. This may identify specific features of these color fundus photographs that are difficult for AI algorithms to handle. This will be done both on the level of the main classification (referral) and for the additional labels.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

Further comments

Further comments from the organizers.

N/A