Ultra-Widefield Fundus Imaging for Diabetic Retinopathy: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Ultra-Widefield Fundus Imaging for Diabetic Retinopathy

Challenge acronym

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

UWF4DR

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.

Diabetic retinopathy (DR), a common and specific complication of diabetes mellitus, is one of the leading causes of preventable blindness among working-aged people [1]. An estimated 103 million adults worldwide were affected by DR in 2020, and the number of people with DR is projected to rise to 130 million in 2030 and 161 million in 2045 [2].

DR can be classified into five categories: no apparent retinopathy, mild nonproliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, and proliferative diabetic retinopathy (PDR) according to the International Clinical Diabetic Retinopathy (ICDR) Severity Scale [3]. In addition, referable DR (RDR) was defined as moderate NPDR or worse, including diabetic macular edema (DME) [4]. Prompt screening, timely referral, and early treatment are widely accepted as consensus for preventing visual loss [5].

Standard colour fundus photography (CFP) that captures macular and optic nerve with a field view of 30 to 50-degree is the gold standard photograph method for the classification of DR [6], and deep learning methods for classification of DR using CFP have gradually matured [7, 8]. However, ultra-widefield (UWF) fundus images, which have a wide range up to 200-degree view of the retina, have emerged as an alternative photograph method for DR management, which not only has agreement with the standard 7-field of the Early Treatment Diabetic Retinopathy Study photos but decreases the rate of ungradable images compared to CFP [9, 10]. Furthermore, UWF fundus images allow the identification of predominantly peripheral lesions (PPL), presenting in 30%-40% of eyes with DR and suggesting a more severe DR level in 11% of eyes [11]. However, the classification of UWF fundus images is time-consuming and labor-intensive, requiring significant effort from human graders, and studies of using computer-aided system for effective analysis of UWF fundus images are limited.

Aiming to advance the state-of-the-art in automatic DR analysis from UWF fundus images, we organize the ultra-widefield fundus imaging for diabetic retinopathy challenge. The challenge encourages researchers to develop algorithms for different tasks in DR analysis using UWF fundus images regarding UWF fundus images quality assessment, classification of DRD, and the classification of DME. On the one hand, the images quality assessment task ensures that the images used for classification are of sufficient quality. On the other hand, the classification of DR and DME provide the foundation for automatic analysis that can help the management of DR patients. This challenge serves as an important milestone in the analysis of DR using ultra-widefield images. We hope that the challenge will drive innovation in automatic medical image analysis, propelling advancements in the field.

References
1. Cheung, N., P. Mitchell, and T.Y. Wong, Diabetic retinopathy. Lancet, 2010. 376(9735): p. 124-36.
2. Teo, Z.L., et al., Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis. Ophthalmology, 2021. 128(11): p. 1580-1591.
3. Wilkinson, C.P., et al., Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology, 2003. 110(9): p. 1677-82.
4. Bellemo, V., et al., Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. Lancet Digit Health, 2019. 1(1): p. e35-e44.
5. Ting, D.S., G.C. Cheung, and T.Y. Wong, Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. Clin Exp Ophthalmol, 2016. 44(4): p. 260-77.
6. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. Ophthalmology, 1991. 98(5 Suppl): p. 786-806.
7. Ting, D.S.W., et al., Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. Jama, 2017. 318(22): p. 2211-2223.
8. Lin, D., et al., Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. Lancet Digit Health, 2021. 3(8): p. e486-e495.
9. Sun, J.K. and L.P. Aiello, The Future of Ultrawide Field Imaging for Diabetic Retinopathy: Pondering the Retinal Periphery. JAMA Ophthalmology, 2016. 134(3): p. 247-248.
10. Silva, P.S., et al., Identification of Diabetic Retinopathy and Ungradable Image Rate with Ultrawide Field Imaging in a National Teleophthalmology Program. Ophthalmology, 2016. 123(6): p. 1360-7.
11. Aiello, L.P., et al., Comparison of Early Treatment Diabetic Retinopathy Study Standard 7-Field Imaging With Ultrawide-Field Imaging for Determining Severity of Diabetic Retinopathy. JAMA Ophthalmology, 2019. 137(1): p. 65-73.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Diabetic retinopathy, Diabetic macular edema, Ultra-widefield fundus, Deep learning

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.

none

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.

We have successfully hosted two challenges at MICCAI2022 and MICCAI2023, each with more than 100 participants. For this challenge at MICCAI2024, we expect more than 100 participants.

Publication and future plans

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

After the challenge, we plan to publish a challenge paper summarizing the main results and conclusions of the challenge. In previous challenges, we have published the challenge summary paper in Patterns. In addition, we are willing to coordinate a publication of the challenge results in MICCAI proceedings.

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.

We plan to use CodaLab (codalab.lisn.upsaclay.fr) for the online submission and ranking of challenge results. If possible, we plan to hold our challenge virtually, so no technical equipment is needed.

TASK 1: Image quality assessment for ultra-widefield fundus

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.

Sufficient ultra-widefield (UWF) fundus image quality is essential to achieve better performance of the subsequent classification performance [1]. However, the quality of UWF images can vary significantly due to factors such as the varying level of experience among operating personnel, and different types of cameras used. Aiming to advance the state-of-the-art in automatic quality assessment of UWF images, we organize the UWF imaging quality assessment task. Participants can evaluate their algorithm's performance and make fair comparisons with other algorithms. We hope this task will serve dual purposes: first, it will help to automatically filter out poor quality data, thus preventing it from influencing the accuracy of subsequent classification tasks; second, it can assist operating personnel in quickly identifying low-quality images, facilitating immediate recapture of the image.

1. Li, Z., et al., Deep learning from "passive feeding" to "selective eating" of real-world data. NPJ Digit Med, 2020. 3: p. 143.

Keywords

List the primary keywords that characterize the task.

Image quality, Image classification, Deep learning

ORGANIZATION

Organizers

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

Bo Qian, Shanghai Jiao Tong University, China
Xiaoyan Hu, The Chinese University of Hong Kong, Hong Kong, China
Xiangning Wang, Shanghai Sixth People's Hospital, China
Junlin Hou, The Hong Kong University of Science and Technology, Hong Kong, China
Dawei Yang, The Chinese University of Hong Kong, Hong Kong, China
Zhouyu Guan, Shanghai Sixth People's Hospital, China
An Ran Ran, The Chinese University of Hong Kong, Hong Kong, China
Tingyao Li, Shanghai Jiao Tong University, China
Timothy Lai, The Chinese University of Hong Kong, Hong Kong, China
Yixiao Jin, Tsinghua University, China
Carmen Chan, The Chinese University of Hong Kong, Hong Kong, China
Simon Szeto, The Chinese University of Hong Kong, Hong Kong, China
Mary Ho, The Chinese University of Hong Kong, Hong Kong, China
Hao Chen, The Hong Kong University of Science and Technology, Hong Kong, China

Carol Cheung, The Chinese University of Hong Kong, Hong Kong, China
Bin Sheng, Shanghai Jiao Tong University, China

b) Provide information on the primary contact person.

Bin Sheng, Shanghai Jiao Tong University, Shanghai, China
E-mail: shengbin@sjtu.edu.cn

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).

N/A

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

codalab.lisn.upsaclay.fr

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

The challenge website will be made available once the challenge has been accepted.

Participation policies

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

Only fully automatic methods allowed.

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.

Publicly available data is 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.

Members of the organizers' institutes that are not associated with the challenge may participate and are eligible for awards.

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

The top three teams will receive certificates and will be invited to be part of the challenge manuscript. In addition, we are actively seeking sponsorship.

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.

All participating teams will receive the model validation results after the challenge. But only the team that submits the method description paper will be eligible for the final ranking.

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).

The challenge organizers aim to publish a summary paper in a relevant journal. Teams that get high performance with novel algorithms will be invited to contribute to this publication. The first and last authors of the submitted papers from these teams will qualify as co-authors in the summary paper. For these teams, they need to submit the corresponding method description papers and source code to ensure fairness. Participating teams are free to publish their results in a separate publication, given proper reference to the challenge.
There is no embargo time.

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.

Submission instructions will be described on the challenge website, where results will be automatically assessed.

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.

Participants will be allowed to validate their results multiple times on validation set before submitting final results. This can also help participant ensure the submission is the correct format. During the final testing phase, the participating teams are allowed to submit multiple runs with a limit of 4 submissions to evaluate their algorithms on test set. Only the best submission will be counted for the official ranking. The participants will not receive feedback before the submission deadline. Allowing a maximum of 4 submissions per team is a reasonable trade-off between overfitting and exploration.

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

Registration date: As soon as possible after the acceptance of the challenge.
Release of training data: No later than Jun 1st 2024.
Release of validation data: No later than Jun 15th 2024.
Submission deadline: August 15th 2024 for test set predictions, September 15th 2024 for method description paper.
Winner and invitation speakers: No later than September 30th 2024
Associated workshop day: Virtually, during the MICCAI 2024 workshop.
(Subject to change depending on the MICCAI 2024 deadlines)

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).

This study conformed to the tenets of the Helsinki Declaration and was approved by the Ethics Committee of the Shanghai Sixth People's Hospital (No. 2021-YS-271).

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

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.

We will provide the code of evaluation metrics on the challenge website.

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

We strongly encourage participating teams to release their codes on GitHub, but it is not mandatory. However, to ensure the reproducibility and credibility of the algorithms, and to make the best-performing methods available for the research community, the top three teams that will receive a certificate are required to release their code with a permissive open source license.

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 are no conflicts of interest. Currently, there is no explicit sponsoring of the challenge, but we aim to contact technology companies with an interest in this challenge. Only organizing members will have access to the test case labels.

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

Assistance, Research, Diagnosis, Screening.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection

- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification

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 patients who are with or at risk of diabetic retinopathy.

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

The challenge cohort is the same as target cohort.

Imaging modality(ies)

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

The imaging modality in this challenge is ultra-widefield fundus.

Context information

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

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

For each image, an anonymous case id will be provided, along with the classification label of this image.

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

None

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.

Eye shown in ultra-widefield fundus image.

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.

The algorithm target is the category of image quality for ultra-widefield fundus image. Category 0 represents ungradable images, category 1 represents gradable images

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)).

AUROC (Area Under the Receiver Operating Characteristic Curve), AUPRC (Area Under the Precision-Recall Curve), sensitivity, specificity.
To perform well in this task, the algorithms to be optimized must give an accurate prediction for the image quality levels. We will consider appropriate metrics, i.e. AUROC, AUPRC, sensitivity, specificity.

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).

The device used to acquire the ultra-widefield fundus images is Optos California (Optos plc, Dunfermline, Scotland, UK).

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

UWF fundus imaging was performed in a single-shot without mydriasis and independently of the clinical examination. All the images were centered on the macula and exported for analysis.

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 training and test cases are provided by Shanghai Sixth People's Hospital.

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).

Data are acquired by well-trained experts in ophthalmology. Also, the clinical evaluators of the challenge have more than 5 years of professional experience in ophthalmology.

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.

One case refers to a ultra-widefield fundus image from one eye of a patient. The analysis will be performed on a per image basis. Data will be separated at a patient level between training and test sets. Training and test cases have the same annotation form.

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

The total number of images is approximately 500. The total number of training images is approximately 250. The total number of validation images is approximately 100. The total number of test images is approximately 150.

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

We would like to provide as many cases as possible, but the availability of clinical data acquisition and the workload of image annotation restrict the total number of training and testing cases. We divide the training set, validation set and test set according to the commonly used ratio of 5:2:3, and the test set is sufficiently powered to distinguish different methods.

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 data are from a diabetic retinopathy screening, so the class distribution is representative of the real-world distribution.

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.

This is a binary image classification task: category 0 represents ungradable, category 1 represents gradable. The annotations are generated and checked manually by professional ophthalmologists. Firstly, original images are assigned separately to two ophthalmologists respectively, and each image is annotated by these two authorized ophthalmologists independently. If the two ophthalmologists give different labeling results for the same image, then the third ophthalmologist who serves as the senior supervisor will help confirm and correct the diagnostic label. The annotation for training and test data will follow the same annotation procedure, with the same annotators and reviewers.

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.

It must be noted that the annotated category for each image should strictly refer to the definition provided.

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.

The two initial annotators are ophthalmologists with more than 5 years of professional experience. The ophthalmologist who serves as the senior supervisor has more than 10 years of professional experience.

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.

A single annotation was performed, and no merging was needed.

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.

No additional pre-processing will be performed on the raw training and test data.

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.

Human annotators could misclassify the corresponding image category, so we aim to alleviate this by using at least two annotators. However, two ophthalmologists may give different labeling results for the same image. Under this condition, a third senior ophthalmologist will serve as the senior supervisor and help confirm the diagnostic label.

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

None

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)

AUROC is used as the primary metric for ranking algorithms. The auxiliary metrics, including AUPRC, sensitivity, and specificity, will also be presented and used as tie-breakers.
We will release the computing codes for all these metrics.

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

AUROC measures how well a model can differentiate between positive and negative samples across various thresholds, and has been widely used in many papers to evaluate the performance of classification tasks in biomedical applications. In order to be consistent with the currently published work and conveniently compare the performance between different studies and algorithms, we decide to use AUROC as the ranking metric in our challenge. However, AUROC may not reflect performance accurately where there is a significant class imbalance. To address this, we will ensure that the number of positive and negative samples in the test set is relatively balanced, which makes the performance evaluation of each algorithm reasonable and fair. In addition, we will also show the metrics including AUPRC, sensitivity, and specificity, which will be used as tie-breakers.

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.

First, separate metric scores are computed for AUROC, AUPRC, sensitivity, and specificity on all test cases. Second, the ranking score is obtained by AUROC. The achieved AUPRC, sensitivity, and specificity will be used as tie-breakers.

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

Each case will be machine-readable, and we expect each trained model to be able to produce a prediction for each case. Otherwise, our submission system will print an error and will not output the calculation results of the metric.

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

AUROC measures how well a model can differentiate between positive and negative samples across various thresholds, and has been widely used in many papers to evaluate the performance of classification tasks in biomedical applications. In order to be consistent with the currently published work and conveniently compare the performance between different studies and algorithms, we decide to use AUROC as the ranking metric in our challenge. However, AUROC may not reflect performance accurately where there is a significant class imbalance. To address this, we will ensure that the number of positive and negative samples in the test set is relatively balanced, which makes the performance evaluation of each algorithm reasonable and fair. In addition, we will also show the metrics including AUPRC, sensitivity, and specificity, which will be used as tie-breakers.

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.

The hypothesis testing (Wilcoxon signed-rank test) will be used to analyze the statistical differences between teams. The python SciPy library and challengeR toolkit will be used for the statistical analyses.

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

The Wilcoxon test is non-parametric and allows us to perform the analysis with minimal hypotheses.

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.

The challenge results will be further analyzed after the challenge. The techniques, such as an ensemble of multiple algorithms and inter-observer agreement, will be conducted. The report and analysis of the challenge results will be published in a relevant journal paper with some of the participating team members as co-authors.

TASK 2: Identification of referable diabetic retinopathy

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.

Diabetic retinopathy (DR), which remains one of the leading causes of preventable visual loss globally, is a common and specific complication of diabetes mellitus [1]. This condition has been estimated to affect 103 million adults in 2020 and has been projected to affect 161 million adults in 2035 [2]. Therefore, an automatic computer-aided system that can detect DR effectively and accurately is of great help to the management of the patient with DR to prevent further visual loss. Ultra-widefield (UWF) fundus images that have a wide range up to 200-degree view of the retina including the central retinal area and the peripheral zones have emerged as an alternative photograph method for DR management. Aiming to advance the state-of-the-art in the identification of referable DR using UWF fundus images, we organize the identification of referable DR from UWF fundus images task. Participants can evaluate their algorithm's performance and make fair comparisons with other algorithms. We hope this task will serve as a milestone to drive innovation in the identification of referable DR using UWF fundus images to achieve better clinical workflow and patient outcomes.

1. Cheung, N., P. Mitchell, and T.Y. Wong, Diabetic retinopathy. Lancet, 2010. 376(9735): p. 124-36.
2. Teo, Z.L., et al., Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis. Ophthalmology, 2021. 128(11): p. 1580-1591.

Keywords

List the primary keywords that characterize the task.

Referable diabetic retinopathy, Ultra-widefield fundus, Deep learning

ORGANIZATION

Organizers

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

Bo Qian, Shanghai Jiao Tong University, China
Xiaoyan Hu, The Chinese University of Hong Kong, Hong Kong, China
Xiangning Wang, Shanghai Sixth People's Hospital, China
Junlin Hou, The Hong Kong University of Science and Technology, Hong Kong, China
Dawei Yang, The Chinese University of Hong Kong, Hong Kong, China
Zhouyu Guan, Shanghai Sixth People's Hospital, China
An Ran Ran, The Chinese University of Hong Kong, Hong Kong, China
Tingyao Li, Shanghai Jiao Tong University, China
Timothy Lai, The Chinese University of Hong Kong, Hong Kong, China
Yixiao Jin, Tsinghua University, China

Carmen Chan, The Chinese University of Hong Kong, Hong Kong, China
Simon Szeto, The Chinese University of Hong Kong, Hong Kong, China
Mary Ho, The Chinese University of Hong Kong, Hong Kong, China
Hao Chen, The Hong Kong University of Science and Technology, Hong Kong, China
Carol Cheung, The Chinese University of Hong Kong, Hong Kong, China
Bin Sheng, Shanghai Jiao Tong University, China

b) Provide information on the primary contact person.

Bin Sheng, Shanghai Jiao Tong University, Shanghai, China
E-mail: shengbin@sjtu.edu.cn

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).

N/A

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

codalab.lisn.upsaclay.fr

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

The challenge website will be made available once the challenge has been accepted.

Participation policies

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

Only fully automatic methods allowed.

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.

Publicly available data is 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.

Members of the organizers' institutes that are not associated with the challenge may participate and are eligible for awards.

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

The top three teams will receive certificates and will be invited to be part of the challenge manuscript. In addition, we are actively seeking sponsorship.

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.

All participating teams will receive the model validation results after the challenge. But only the team that submits the method description paper will be eligible for the final ranking.

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).

The challenge organizers aim to publish a summary paper in a relevant journal. Teams that get high performance with novel algorithms will be invited to contribute to this publication. The first and last authors of the submitted papers from these teams will qualify as co-authors in the summary paper. For these teams, they need to submit the corresponding method description papers and source code to ensure fairness. Participating teams are free to publish their results in a separate publication, given proper reference to the challenge.
There is no embargo time.

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.

Submission instructions will be described on the challenge website, where results will be automatically assessed.

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.

Participants will be allowed to validate their results multiple times on validation set before submitting final results. This can also help participant ensure the submission is the correct format. During the final testing phase, the participating teams are allowed to submit multiple runs with a limit of 4 submissions to evaluate their algorithms on test set. Only the best submission will be counted for the official ranking. The participants will not receive feedback before the submission deadline. Allowing a maximum of 4 submissions per team is a reasonable trade-off between overfitting and exploration.

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

Registration date: As soon as possible after the acceptance of the challenge.
Release of training data: No later than Jun 1st 2024.
Release of validation data: No later than Jun 15th 2024.
Submission deadline: August 15th 2024 for test set predictions, September 15th 2024 for method description paper.
Winner and invitation speakers: No later than September 30th 2024
Associated workshop day: Virtually, during the MICCAI 2024 workshop.
(Subject to change depending on the MICCAI 2024 deadlines)

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).

This study conformed to the tenets of the Helsinki Declaration and was approved by the Ethics Committee of the Shanghai Sixth People's Hospital (No. 2021-YS-271).

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

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.

We will provide the code of evaluation metrics on the challenge website.

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

We strongly encourage participating teams to release their codes on GitHub, but it is not mandatory. However, to ensure the reproducibility and credibility of the algorithms, and to make the best-performing methods available for the research community, the top three teams that will receive a certificate are required to release their code with a permissive open source license.

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 are no conflicts of interest. Currently, there is no explicit sponsoring of the challenge, but we aim to contact technology companies with an interest in this challenge. Only organizing members will have access to the test case labels.

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

Assistance, Research, Diagnosis, Screening.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification

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 patients who are with or at risk of diabetic retinopathy.

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

The challenge cohort is the same as target cohort.

Imaging modality(ies)

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

The imaging modality in this challenge is ultra-widefield fundus.

Context information

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

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

For each image, an anonymous case id will be provided, along with the classification label of this image.

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

None

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.

Eye shown in ultra-widefield fundus image.

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.

The algorithm target is to detect whether a given ultra-widefield fundus image is a referable DR. Category 0 represents non-referable DR, category 1 represents referable DR.

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)).

AUROC (Area Under the Receiver Operating Characteristic Curve), AUPRC (Area Under the Precision-Recall Curve), sensitivity, specificity.
To perform well in this task, the algorithms to be optimized must give an accurate prediction for the presence of referable DR. We will consider appropriate metrics, i.e. AUROC, AUPRC, sensitivity, specificity.

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).

The device used to acquire the ultra-widefield fundus images is Optos California (Optos plc, Dunfermline, Scotland, UK).

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

UWF fundus imaging was performed in a single-shot without mydriasis and independently of the clinical examination. All the images were centered on the macula and exported for analysis.

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 training and test cases are provided by Shanghai Sixth People's Hospital.

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).

Data are acquired by well-trained experts in ophthalmology. Also, the clinical evaluators of the challenge have more than 5 years of professional experience in ophthalmology.

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.

One case refers to a ultra-widefield fundus image from one eye of a patient. The analysis will be performed on a per image basis. Data will be separated at a patient level between training and test sets. Training and test cases have the same annotation form.

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

The total number of images is approximately 400. The total number of training images is approximately 200. The total number of validation images is approximately 80. The total number of test images is approximately 120.

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

We would like to provide as many cases as possible, but the availability of clinical data acquisition and the workload of image annotation restrict the total number of training and testing cases. We divide the training set, validation set and test set according to the commonly used ratio of 5:2:3, and the test set is sufficiently powered to distinguish different methods.

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 data are from a diabetic retinopathy screening, so the class distribution is representative of the real-world distribution.

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.

This is a binary image classification task: category 0 represents non-referable DR, category 1 represents referable DR. The annotations are generated and checked manually by professional ophthalmologists. Firstly, original images are assigned separately to two ophthalmologists respectively, and each image is annotated by these two authorized ophthalmologists independently. If the two ophthalmologists give different labeling results for the same image, then the third ophthalmologist who serves as the senior supervisor will help confirm and correct the diagnostic label. The annotation for training and test data will follow the same annotation procedure, with the same annotators and reviewers.

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.

It must be noted that the annotated category for each image should strictly refer to the definition provided.

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.

The two initial annotators are ophthalmologists with more than 5 years of professional experience. The ophthalmologist who serves as the senior supervisor has more than 10 years of professional experience.

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.

A single annotation was performed, and no merging was needed.

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.

No additional pre-processing will be performed on the raw training and test data.

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.

Human annotators could misclassify the corresponding image category, so we aim to alleviate this by using at least two annotators. However, two ophthalmologists may give different labeling results for the same image. Under this condition, a third senior ophthalmologist will serve as the senior supervisor and help confirm the diagnostic label.

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

None

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)

AUROC is used as the primary metric for ranking algorithms. The auxiliary metrics, including AUPRC, sensitivity, and specificity, will also be presented and used as tie-breakers.
We will release the computing codes for all these metrics.

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

AUROC measures how well a model can differentiate between positive and negative samples across various thresholds, and has been widely used in many papers to evaluate the performance of classification tasks in biomedical applications. In order to be consistent with the currently published work and conveniently compare the performance between different studies and algorithms, we decide to use AUROC as the ranking metric in our challenge. However, AUROC may not reflect performance accurately where there is a significant class imbalance. To address this, we will ensure that the number of positive and negative samples in the test set is relatively balanced, which makes the performance evaluation of each algorithm reasonable and fair. In addition, we will also show the metrics including AUPRC, sensitivity, and specificity, which will be used as tie-breakers.

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.

First, separate metric scores are computed for AUROC, AUPRC, sensitivity, and specificity on all test cases. Second, the ranking score is obtained by AUROC. The achieved AUPRC, sensitivity, and specificity will be used as tie-breakers.

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

Each case will be machine-readable, and we expect each trained model to be able to produce a prediction for each case. Otherwise, our submission system will print an error and will not output the calculation results of the metric.

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

AUROC measures how well a model can differentiate between positive and negative samples across various thresholds, and has been widely used in many papers to evaluate the performance of classification tasks in biomedical applications. In order to be consistent with the currently published work and conveniently compare the performance between different studies and algorithms, we decide to use AUROC as the ranking metric in our challenge. However, AUROC may not reflect performance accurately where there is a significant class imbalance. To address this, we will ensure that the number of positive and negative samples in the test set is relatively balanced, which makes the performance evaluation of each algorithm reasonable and fair. In addition, we will also show the metrics including AUPRC, sensitivity, and specificity, which will be used as tie-breakers.

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.

The hypothesis testing (Wilcoxon signed-rank test) will be used to analyze the statistical differences between teams. The python SciPy library and challengeR toolkit will be used for the statistical analyses.

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

The Wilcoxon test is non-parametric and allows us to perform the analysis with minimal hypotheses.

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.

The challenge results will be further analyzed after the challenge. The techniques, such as an ensemble of multiple algorithms and inter-observer agreement, will be conducted. The report and analysis of the challenge results will be published in a relevant journal paper with some of the participating team members as co-authors.

TASK 3: Identification of diabetic macular edema

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.

Diabetic Macular Edema (DME), one of the causes of vision loss in individuals with diabetes, results from fluid accumulation in the macula due to damaged retinal blood vessels [1, 2]. The integration of automatic computer-aided systems has improved the identification of DME, significantly improving both accuracy and efficiency through state-of-the-art image analysis and machine learning technologies. These systems are instrumental in facilitating prompt and precise diagnosis, and they aid healthcare professionals in formulating more accurate treatment strategies, thereby improving patient care and clinical outcomes. Currently, Optical Coherence Tomography (OCT) is recognized as the gold standard for DME identification, offering detailed cross-sectional views of the retina. Additionally, ultra-widefield (UWF) fundus imaging provides comprehensive 200-degree views of the retina, encompassing both the central and peripheral areas, thereby delivering a more complete understanding than traditional fundus photography. Exploring DME identification using a combination of UWF fundus images and OCT-based labeling is a promising area of research. Therefore, we design a task of detecting DME from UWF fundus images. This initiative aims to drive innovation in the identification of DME using UWF imaging.

1. Vidal, P. L., et al. Diabetic macular edema characterization and visualization using optical coherence tomography images. Applied Sciences, 2020. 10(21), 7718.
2. Davidson, J. A., et al. How the diabetic eye loses vision. Endocrine, 2007. 32: p. 107-116.

Keywords

List the primary keywords that characterize the task.

Diabetic macular edema, Ultra-widefield fundus, Computer-aided systems

ORGANIZATION

Organizers

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

Bo Qian, Shanghai Jiao Tong University, China
Xiaoyan Hu, The Chinese University of Hong Kong, Hong Kong, China
Xiangning Wang, Shanghai Sixth People's Hospital, China
Junlin Hou, The Hong Kong University of Science and Technology, Hong Kong, China
Dawei Yang, The Chinese University of Hong Kong, Hong Kong, China
Zhouyu Guan, Shanghai Sixth People's Hospital, China
An Ran Ran, The Chinese University of Hong Kong, Hong Kong, China
Tingyao Li, Shanghai Jiao Tong University, China

Timothy Lai, The Chinese University of Hong Kong, Hong Kong, China
Yixiao Jin, Tsinghua University, China
Carmen Chan, The Chinese University of Hong Kong, Hong Kong, China
Simon Szeto, The Chinese University of Hong Kong, Hong Kong, China
Mary Ho, The Chinese University of Hong Kong, Hong Kong, China
Hao Chen, The Hong Kong University of Science and Technology, Hong Kong, China
Carol Cheung, The Chinese University of Hong Kong, Hong Kong, China
Bin Sheng, Shanghai Jiao Tong University, China

b) Provide information on the primary contact person.

Bin Sheng, Shanghai Jiao Tong University, Shanghai, China
E-mail: shengbin@sjtu.edu.cn

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).

N/A

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

codalab.lisn.upsaclay.fr

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

The challenge website will be made available once the challenge has been accepted.

Participation policies

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

Only fully automatic methods allowed.

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.

Publicly available data is 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.

Members of the organizers' institutes that are not associated with the challenge may participate and are eligible for awards.

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

The top three teams will receive certificates and will be invited to be part of the challenge manuscript. In addition, we are actively seeking sponsorship.

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.

All participating teams will receive the model validation results after the challenge. But only the team that submits the method description paper will be eligible for the final ranking.

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).

The challenge organizers aim to publish a summary paper in a relevant journal. Teams that get high performance with novel algorithms will be invited to contribute to this publication. The first and last authors of the submitted papers from these teams will qualify as co-authors in the summary paper. For these teams, they need to submit the corresponding method description papers and source code to ensure fairness. Participating teams are free to publish their results in a separate publication, given proper reference to the challenge.
There is no embargo time.

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.

Submission instructions will be described on the challenge website, where results will be automatically assessed.

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.

Participants will be allowed to validate their results multiple times on validation set before submitting final results. This can also help participant ensure the submission is the correct format. During the final testing phase, the participating teams are allowed to submit multiple runs with a limit of 4 submissions to evaluate their algorithms on test set. Only the best submission will be counted for the official ranking. The participants will not receive feedback before the submission deadline. Allowing a maximum of 4 submissions per team is a reasonable trade-off between overfitting and exploration.

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

Registration date: As soon as possible after the acceptance of the challenge.
Release of training data: No later than Jun 1st 2024.
Release of validation data: No later than Jun 15th 2024.
Submission deadline: August 15th 2024 for test set predictions, September 15th 2024 for method description paper.
Winner and invitation speakers: No later than September 30th 2024
Associated workshop day: Virtually, during the MICCAI 2024 workshop.
(Subject to change depending on the MICCAI 2024 deadlines)

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).

This study conformed to the tenets of the Helsinki Declaration and was approved by the Ethics Committee of the Shanghai Sixth People's Hospital (No. 2021-YS-271).

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

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.

We will provide the code of evaluation metrics on the challenge website.

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

We strongly encourage participating teams to release their codes on GitHub, but it is not mandatory. However, to ensure the reproducibility and credibility of the algorithms, and to make the best-performing methods available for the research community, the top three teams that will receive a certificate are required to release their code with a permissive open source license.

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 are no conflicts of interest. Currently, there is no explicit sponsoring of the challenge, but we aim to contact technology companies with an interest in this challenge. Only organizing members will have access to the test case labels.

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

Assistance, Research, Diagnosis, Screening.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification

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 patients who are with or at risk of diabetic retinopathy.

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

The challenge cohort is the same as target cohort.

Imaging modality(ies)

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

The imaging modality in this challenge is ultra-widefield fundus.

Context information

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

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

For each image, an anonymous case id will be provided, along with the classification label of this image.

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

None

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.

Eye shown in ultra-widefield fundus image.

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.

The algorithm target is to detect the presence of DME given ultra-widefield fundus image. Category 0 represents non-DME, category 1 represents DME.

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)).

AUROC (Area Under the Receiver Operating Characteristic Curve), AUPRC (Area Under the Precision-Recall Curve), sensitivity, specificity.
To perform well in this task, the algorithms to be optimized must give an accurate prediction for the presence of DME. We will consider appropriate metrics, i.e. AUROC, AUPRC, sensitivity, specificity.

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).

The device used to acquire the ultra-widefield fundus images is Optos California (Optos plc, Dunfermline, Scotland, UK).

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

UWF fundus imaging was performed in a single-shot without mydriasis and independently of the clinical examination. All the images were centered on the macula and exported for analysis.

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 training and test cases are provided by Shanghai Sixth People's Hospital.

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).

Data are acquired by well-trained experts in ophthalmology. Also, the clinical evaluators of the challenge have more than 5 years of professional experience in ophthalmology.

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.

One case refers to a ultra-widefield fundus image from one eye of a patient. The analysis will be performed on a per image basis. Data will be separated at a patient level between training and test sets. Training and test cases have the same annotation form.

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

The total number of images is approximately 400. The total number of training images is approximately 200. The total number of validation images is approximately 80. The total number of test images is approximately 120.

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

We would like to provide as many cases as possible, but the availability of clinical data acquisition and the workload of image annotation restrict the total number of training and testing cases. We divide the training set, validation set and test set according to the commonly used ratio of 5:2:3, and the test set is sufficiently powered to distinguish different methods.

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 data are from a diabetic retinopathy screening, so the class distribution is representative of the real-world distribution.

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.

This is a binary image classification task: category 0 represents non-DME, category 1 represents the presence of DME. The annotations are generated and checked manually by professional ophthalmologists. Firstly, original images are assigned separately to two ophthalmologists respectively, and each image is annotated by these two authorized ophthalmologists independently. If the two ophthalmologists give different labeling results for the same image, then the third ophthalmologist who serves as the senior supervisor will help confirm and correct the diagnostic label. The annotation for training and test data will follow the same annotation procedure, with the same annotators and reviewers.

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.

It must be noted that the annotated category for each image should strictly refer to the definition provided.

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.

The two initial annotators are ophthalmologists with more than 5 years of professional experience. The ophthalmologist who serves as the senior supervisor has more than 10 years of professional experience.

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.

A single annotation was performed, and no merging was needed.

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.

No additional pre-processing will be performed on the raw training and test data.

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.

Human annotators could misclassify the corresponding image category, so we aim to alleviate this by using at least two annotators. However, two ophthalmologists may give different labeling results for the same image. Under this condition, a third senior ophthalmologist will serve as the senior supervisor and help confirm the diagnostic label.

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

None

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)

AUROC is used as the primary metric for ranking algorithms. The auxiliary metrics, including AUPRC, sensitivity, and specificity, will also be presented and used as tie-breakers.
We will release the computing codes for all these metrics.

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

AUROC measures how well a model can differentiate between positive and negative samples across various thresholds, and has been widely used in many papers to evaluate the performance of classification tasks in biomedical applications. In order to be consistent with the currently published work and conveniently compare the performance between different studies and algorithms, we decide to use AUROC as the ranking metric in our challenge. However, AUROC may not reflect performance accurately where there is a significant class imbalance. To address this, we will ensure that the number of positive and negative samples in the test set is relatively balanced, which makes the performance evaluation of each algorithm reasonable and fair. In addition, we will also show the metrics including AUPRC, sensitivity, and specificity, which will be used as tie-breakers.

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.

First, separate metric scores are computed for AUROC, AUPRC, sensitivity, and specificity on all test cases. Second, the ranking score is obtained by AUROC. The achieved AUPRC, sensitivity, and specificity will be used as tie-breakers.

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

Each case will be machine-readable, and we expect each trained model to be able to produce a prediction for each case. Otherwise, our submission system will print an error and will not output the calculation results of the metric.

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

AUROC measures how well a model can differentiate between positive and negative samples across various thresholds, and has been widely used in many papers to evaluate the performance of classification tasks in biomedical applications. In order to be consistent with the currently published work and conveniently compare the performance between different studies and algorithms, we decide to use AUROC as the ranking metric in our challenge. However, AUROC may not reflect performance accurately where there is a significant class imbalance. To address this, we will ensure that the number of positive and negative samples in the test set is relatively balanced, which makes the performance evaluation of each algorithm reasonable and fair. In addition, we will also show the metrics including AUPRC, sensitivity, and specificity, which will be used as tie-breakers.

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.

The hypothesis testing (Wilcoxon signed-rank test) will be used to analyze the statistical differences between teams. The python SciPy library and challengeR toolkit will be used for the statistical analyses.

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

The Wilcoxon test is non-parametric and allows us to perform the analysis with minimal hypotheses.

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.

The challenge results will be further analyzed after the challenge. The techniques, such as an ensemble of multiple algorithms and inter-observer agreement, will be conducted. The report and analysis of the challenge results will be published in a relevant journal paper with some of the participating team members as co-authors.

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.

We confirm to consider a virtual setting for our challenge.