Diabetic Retinopathy Analysis Challenge 2022: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Diabetic Retinopathy Analysis Challenge 2022

Challenge acronym

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

DRAC2022

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 is one of the leading causes of blindness and affects approximately 78% people, with a history of diabetes of 15 years or longer [1]. DR often causes gradual changes in vasculature structure and resulting abnormalities. DR is diagnosed by visually inspecting retinal fundus images for the presence of retinal lesions, such as microaneurysms (MAs), intraretinal microvascular abnormalities (IRMAs), nonperfusion areas and neovascularization. The detection of these lesions is critical to the diagnosis of DR. With rising popularity, OCT angiography (OCTA) has the capability of visualizing the retinal and choroidal vasculature at a microvascular level in great detail [2]. Specially, swept-source (SS)-OCTA allows additionally the individual assessment of the choroidal vasculature. There are already some works using SS-OCTA to grade for qualitative features of diabetic retinopathy [3-5]. Further, ultra-wide optical coherence tomography angiography imaging (UW-OCTA) modality showed higher burden of pathology in the retinal periphery that was not captured by typical OCTA [6]. Some works already use UW-OCTA on DR analysis [6, 7]. The traditional diagnosis of DR grading mainly relies on fundus photography and FFA, especially for PDR, which seriously endangers vision health. FA is mainly used to detect the presence or absence of new blood vessels. Fundus photography is difficult to detect early or small neovascular lesions. FA is an invasive fundus imaging that cannot be used in patients with allergies, pregnancy, or poor liver and kidney function. The ultra-wide OCTA can non-invasively detect the changes of DR neovascularization, thus it is an important imaging modality to help ophthalmologist diagnose PDR. However, there are currently no works capable of automatic DR analysis using UW-OCTA. In the process of DR analysis, the image quality of UW-OCTA needs to be assessed first, and the images with better imaging quality are selected. Then DR analysis is performed, such as lesion segmentation and PDR detection. Thus, it is crucial to build a flexible and robust model to realize automatic image quality assessment, lesion segmentation and PDR detection.
In order to promote the application of machine learning and deep learning algorithms in automatic image quality assessment, lesion segmentation and PDR detection using UW-OCTA images, and promote the application of corresponding technologies in clinical diagnosis of DR, we provide a standardized ultra-wide (swept-source) optical coherence tomography angiography (UW-OCTA) data set for testing the effectiveness of various

algorithms. With this dataset, different algorithms can test their performance and make a fair comparison with other algorithms. We believe this dataset is an important milestone in automatic image quality assessment, lesion segmentation and DR grading.

[1] Tian M , Wolf S , Munk M R , et al. Evaluation of different Swept'Source optical coherence tomography angiography (SSOCTA) slabs for the detection of features of diabetic retinopathy[J]. Acta ophthalmologica, 2019, 98(1).
[2] Spaide R F, Fujimoto J G, Waheed N K, et al. Optical coherence tomography angiography[J]. Progress in retinal and eye research, 2018, 64: 1-55.
[3] Schaal K B, Munk M R, Wyssmueller I, et al. Vascular abnormalities in diabetic retinopathy assessed with swept-source optical coherence tomography angiography widefield imaging[J]. Retina, 2019, 39(1): 79-87.
[4] Stanga P E, Papayannis A, Tsamis E, et al. New findings in diabetic maculopathy and proliferative disease by swept-source optical coherence tomography angiography[J]. OCT Angiography in Retinal and Macular Diseases, 2016, 56: 113-121.
[5] Schaal K B, Munk M R, Wyssmueller I, et al. Vascular abnormalities in diabetic retinopathy assessed with swept-source optical coherence tomography angiography widefield imaging[J]. Retina, 2019, 39(1): 79-87.
[6] Zhang Q, Rezaei K A, Saraf S S, et al. Ultra-wide optical coherence tomography angiography in diabetic retinopathy[J]. Quantitative imaging in medicine and surgery, 2018, 8(8): 743.
[7] Russell J F, Shi Y, Hinkle J W, et al. Longitudinal wide-field swept-source OCT angiography of neovascularization in proliferative diabetic retinopathy after panretinal photocoagulation[J]. Ophthalmology Retina, 2019, 3(4): 350-361.

Challenge keywords

List the primary keywords that characterize the challenge.

diabetic retinopathy, ultra-wide, optical coherence tomography angiography, deep learning

Year

The challenge will take place in ...

2022

FURTHER INFORMATION FOR MICCAI 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 are hoping for at least 40-45 participants due to the popularity of the topic.

Publication and future plans

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

We plan to summarize a publication of the results after the challenge.

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 grand-challenge.org for the submission of final model.

TASK: Analysis Towards Segmentation of Diabetic Retinopathy

SUMMARY

Keywords

List the primary keywords that characterize the task.

diabetic retinopathy, microaneurysms, capillary nonperfusion, neovascularization, segmentation

ORGANIZATION

Organizers

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

Bin Sheng, Shanghai Jiaotong University, China
Huating Li, Shanghai Sixth People's Hospital, China
Hao Chen, Hong Kong University of Science and Technology, Hong Kong, China
Yiyu Cai, Nanyang Technological University, Singapore
Qiang Wu, Shanghai Sixth People's Hospital, China
Weiping Jia, Shanghai Sixth People's Hospital, China
Xiangning Wang, Shanghai Sixth People's Hospital, China
Bo Qian, Shanghai Jiaotong University, China
Ruhan Liu, Shanghai Jiaotong University, China
Ling Dai, Shanghai Jiaotong University, China

b) Provide information on the primary contact person.

Bin Sheng, Shanghai Jiaotong University, China
E-mail: shengbin@cs.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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

grand-challenge.org (will setup once the proposal is accepted)

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

None at this moment.

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 policy defined.

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.

Any organization that is a member of the organizing committee does not exclude participation in the challenge, but must ensure that the content submitted is completely independent of the organizing committee members.

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

Every team can get a certificate award. In addition, we are actively seeking sponsorship, and we anticipate being able to provide cash rewards and / or graphics cards for top-3 teams. We are confident in being able to attain 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.

Results will be announced after the submission deadline.

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 will publish at least one challenge journal paper and potentially more. Participants are requested to publish a description of their method and results on arxiv.org together with their submission. All participants can be authors of the article. Participating teams are free to publish their own results in a separate publication.
Participants may publish papers including their official performance on the challenge data set, given proper reference of 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.

We will provide a link or email before the challenge starts, and the results will be submitted through the link or email. The format of the submission file is as follows:
Mask images indicating image id, with pixel-wise label for background (0), microaneurysms (1), intraretinal microvascular abnormalities (2) , nonperfusion areas (3) , neovascularization (4).
Fine-grained instructions will be published on the challenge website.

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.

The participating teams can use the validation set we provide to evaluate their algorithms. Then they can submit the results of the algorithm on the test set to us. We will publish the score of each team's algorithm on the test set on the website. Each team is allowed to submit up to four times, and the last submitted result will be recorded as the challenge result.

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

Training data release: 01/05/2022
Validation data release: 01/06/2022
Test images release: 01/07/2022
Submission deadline: 31/08/2022
Announcement of results at MICCAI 2022: 18/09/2022

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).

We have received approval from Shanghai Sixth People's Hospital to use the data set for the purpose of research. Also, the provided challenge data is anonymized that can be used for the challenge.

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 ND.

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 a link or email for the submission of the results. For transparency, we will release the source code used for calculating final scores after the closing date of the challenge.

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

We do not force teams to open source their code.

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.

Currently, there is no explicit sponsoring of the challenge.
Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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

Research, Diagnosis, Assistance, Screening.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

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 consists of patients who 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.

Patients with diabetes retinopathy or at risk of diabetic retinopathy. The lesion include microaneurysms, intraretinal microvascular abnormalities, neovascularization, and nonperfusion areas. All images are from untreated patients with an initial diagnosis of DR.

Imaging modality(ies)

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

Ultra-wide (swept-source) optical coherence tomography angiography (UW-OCTA) mosaic image. The ultra-wide OCTA mosaic images are 2-D plane enface images.

Context information

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

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

We do not provide additional information to the dataset.

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.

The data consists of images selected from the ultra-wide OCTA mosaic images showing features of diabetic retinopathy.

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 lesion segmentation of diabetic retinopathy using ultra-wide OCTA mosaic images, i.e., microaneurysms, intraretinal microvascular abnormalities, neovascularization, and nonperfusion areas.

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)).

Mean Dice Similarity Coefficient (DSC), mean Intersection of Union (mIOU).

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 SS-OCT/OCTA system (VG200S; SVision Imaging, Henan, China) contains an SS laser with a center wavelength of approximately 1050 nm and a scan rate of 200,000 A scans per second. The system is equipped with an eye tracking utility based on integrated confocal scanning laser ophthalmoscopes to eliminate eye movement artifacts. The axial resolution is 5 um and the lateral resolution is 13 um. The scanning depth is 3 mm.
The speed of SS-OCT used reaches hundreds of thousands of times per second, which can image faster and does not require 12 seconds. Second, we have eye tracking and dynamic image quality evaluation systems. During the scan, the patient can blink. However, we know that blinking will produce artifacts. Therefore, we subsequently used the Svision algorithm to enhance the signal of weak blood flow. At the same time, the 3D artifact removal algorithm is used to remove the tomographic blood flow artifacts in the artificial intelligence layering, so as to avoid the layering disorder in the case of lesions. Finally, we can get high-quality images.

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

The ultra-wide OCTA image is composed of 5 12*12mm SS-OCTA images, including five areas centered on the macular fovea, supratemporal, inferior temporal, supranasal, and subnasal. The slab we use is the inner retina slab. Then the 5 area images are stitched into an ultra-wide OCTA image through internal software. That is, each stitched image is stitched from 5 12mm * 12mm OCTA images. Each 12*12 mm volume consists of 1024 A scans per B scan (the distance between adjacent A scans is 24 mm) and 1024 B scan positions per volume scan (the distance between adjacent B scans is 24 mm). Acquire two repeated B-scans at each B-scan position to generate OCTA images. The time to obtain a single 12*12 mm volume is at least 12 seconds. All OCTA imaging is performed by professional ophthalmologists who repeat the image acquisition as many times as possible to ensure that the image has strong OCT signal penetration and minimal motion artifacts. The quality score of the scan is expressed as the SNR in decibels (dB) and ranges from 1 (poor quality) to 10 (excellent quality). Scans with a score >8 dB are considered high quality and included.

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.

All data is provided from the Shanghai Sixth People's Hospital Diabetes Diagnosis and Treatment Center.

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 labeled by ophthalmologists with more than 5 years of professional experience. We have developed a strict annotation process for the annotators to ensure the accuracy of annotation.

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.

The data are ultra-wide OCTA images. We divide the data into: training set (70%), validation set (10%), and a hidden testing set used for participant ranking (20%). Any areas to be segmented is annotated by ophthalmologists. Participant teams can get training set and validation for model training and validation.

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

Training: 350 labelled images
Validation: 50 labelled images
Testing: 100 labelled images

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

For DR segmentation, data labeling is time-consuming and laborious. It is crucial to propose a model that get the high accuracy with as little data as possible. We encourage researchers to use some techniques such as

unsupervised learning, semi- supervised and transfer learning. Moreover, we may provide some unlabeled pictures for the challenge.

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 division of training set, validation set and test set was completely random. We thus assume that there is no statistically significant difference between the sets.

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.

The ground truth was produced by professional ophthalmologists from Shanghai Sixth People's Hospital. The results were checked by at least two ophthalmologists.
For the label of segmentation, the corresponding classes of pixel values are: microaneurysms (1), intraretinal microvascular abnormalities (2) , nonperfusion areas (3) , neovascularization (4). The value in brackets is the pixel value. The background pixel value is 0.

The labeling process is as follows. In the first step, after an ophthalmologist annotates all the images, we randomly shuffle the images and send them to the same ophthalmologist for annotation one by one. For a lesion area, if the IOU of the first and second labeling is less than the threshold (currently 0.7), the ophthalmologist is asked to relabel the lesion area twice until the IOU labelled twice is greater than the threshold. The final marked mask of this ophthalmologist consists of the overlapping region of the two marked masks.

In the second step, another ophthalmologist will use the same process to complete the labeling. For a lesion area, if the IOU marked by the two ophthalmologists is greater than the threshold (currently 0.7), the final mask of the lesion area is composed of the overlapping area of the masks marked by the two ophthalmologists; if the IOU marked by the two ophthalmologists is less than the threshold, then the mask of the lesion area is discussed and decided by two other more experienced ophthalmologists.

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.

We have sent the instructions of the annotation process to the annotators , asking them to annotate strictly according to the process.

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.

All subjects involved with the annotation process are several professional ophthalmologists who have worked in the field for more than 5 years.

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.

Since this would reveal too much information and to prevent cheating, we do not fully disclose this information and are willing to share it with the reviewers if it would be helpful.

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.

There is no special data preprocessing method, we provide original resolution images and labels.

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.

The following errors are plausible to occur:
Missed annotation: In semantic segmentation, it is difficult to completely mark all target areas, especially for small areas, such as the lesion areas in our task. Moreover, the lesion areas in SS-OCTA are small and fuzzy, and the ophthalmologist may miss some lesion areas when marking.
Bias in the contour of lesion areas: Because the contour of the lesion areas are irregular, there will be bias when the ophthalmologist marks the contours of the lesion areas.
Therefore, we have formulated the above-mentioned instructions of the annotation process to improve the annotation accuracy and asking them to annotate strictly according to the process.

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

Besides annotation errors we do not expect other sources of error.

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)

mean Dice Similarity Coefficient (DSC), mean Intersection of Union (mIOU)

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

Our task is to segment the lesion areas from ultra-wide OCTA mosaic image and compare the similarity between the predicted segmentation area and the area marked by experts. The metrics listed above are often used to measure this similarity.

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.

We calculate the mean Dice Similarity Coefficient (DSC) of all classes on the test set as the primary metric used for

ranking.

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

All the missing results on an image will be treated as no prediction on the image. Therefore, such cases will be treated as false negatives.

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

This method provides an average performance of whether all classes of lesions are effectively segmented. We give an average dice score on the test data which is not involved in the model training. This gives an emphasis on the generalization of the method.

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.

Stability will be investigated via bootstrapping and hypothesis testing

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

Bootstrapping was identified as an appropriate approach to investigate ranking variability.

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.

In the case of the same mean DSC, we will further compare the other metrics of the algorithm, such as mIOU.

TASK: Image Quality Assessment

SUMMARY

Keywords

List the primary keywords that characterize the task.

quality assessment, ultra-wide, optical coherence tomography angiography, classification

ORGANIZATION

Organizers

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

Bin Sheng, Shanghai Jiaotong University, China
Huating Li, Shanghai Sixth People's Hospital, China
Hao Chen, Hong Kong University of Science and Technology, Hong Kong, China
Yiyu Cai, Nanyang Technological University, Singapore
Qiang Wu, Shanghai Sixth People's Hospital, China
Weiping Jia, Shanghai Sixth People's Hospital, China
Xiangning Wang, Shanghai Sixth People's Hospital, China
Bo Qian, Shanghai Jiaotong University, China
Ruhan Liu, Shanghai Jiaotong University, China
Ling Dai, Shanghai Jiaotong University, China

b) Provide information on the primary contact person.

Bin Sheng, Shanghai Jiaotong University, China
E-mail: shengbin@cs.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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

grand-challenge.org (will setup once the proposal is accepted)

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

None at this moment.

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 policy defined.

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.

Any organization that is a member of the organizing committee does not exclude participation in the challenge, but must ensure that the content submitted is completely independent of the organizing committee members.

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

Every team can get a certificate award. In addition, we are actively seeking sponsorship, and we anticipate being able to provide cash rewards and / or graphics cards for top-3 teams. We are confident in being able to attain 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.

Results will be announced after the submission deadline.

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 will publish at least one challenge journal paper and potentially more. Participants are requested to publish a description of their method and results on arxiv.org together with their submission. All participants can be authors of the article. Participating teams are free to publish their own results in a separate publication.
Participants may publish papers including their official performance on the challenge data set, given proper reference of 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.

We will provide a link or email before the challenge starts, and the results will be submitted through the link or email. The format of the submission file is as follows:
A txt file indicating image id, with level of image quality for poor quality level (0), good quality level (1), and excellent quality level (2).
Fine-grained instructions will be published on the challenge website.

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.

The participating teams can use the validation set we provide to evaluate their algorithms. Then they can submit the results of the algorithm on the test set to us. We will publish the score of each team's algorithm on the test set on the website. Each team is allowed to submit up to four times, and the last submitted result will be recorded as the challenge result.

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

Training data release: 01/05/2022
Validation data release: 01/06/2022
Test images release: 01/07/2022
Submission deadline: 31/08/2022
Announcement of results at MICCAI 2022: 18/09/2022

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).

We have received approval from Shanghai Sixth People's Hospital to use the data set for the purpose of research. Also, the provided challenge data is anonymized that can be used for the challenge.

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 ND.

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 a link or email for the submission of the results. For transparency, we will release the source code used for calculating final scores after the closing date of the challenge.

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

We do not force teams to open source their code.

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.

Currently, there is no explicit sponsoring of the challenge.
Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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

Research, Assistance, 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 consists of patients who 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.

Patients undergoing DR screening in the hospital using the scan of ultra-wide OCTA mosaic image. All images are from untreated patients with an initial diagnosis of DR

Imaging modality(ies)

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

Ultra-wide (swept-source) optical coherence tomography angiography (UW-OCTA) mosaic image. The ultra-wide OCTA mosaic images are 2-D plane enface images.

Context information

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

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

We do not provide additional information to the dataset.

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.

The data consists of images selected from the ultra-wide OCTA mosaic image with the different levels of image quality.

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 classify the quality of a given image, class 0 represents poor quality level, class 1 represents good quality level, and class 2 represents excellent quality level.

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)).

Quadratic-weighted Kappa, macro-AUC, micro-AUC, macro-precision, macro-sensitivity and macro-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 SS-OCT/OCTA system (VG200S; SVision Imaging, Henan, China) contains an SS laser with a center wavelength of approximately 1050 nm and a scan rate of 200,000 A scans per second. The system is equipped with an eye tracking utility based on integrated confocal scanning laser ophthalmoscopes to eliminate eye movement artifacts. The axial resolution is 5 um and the lateral resolution is 13 um. The scanning depth is 3 mm.
The speed of SS-OCT used reaches hundreds of thousands of times per second, which can image faster and does not require 12 seconds. Second, we have eye tracking and dynamic image quality evaluation systems. During the scan, the patient can blink. However, we know that blinking will produce artifacts. Therefore, we subsequently used the Svision algorithm to enhance the signal of weak blood flow. At the same time, the 3D artifact removal algorithm is used to remove the tomographic blood flow artifacts in the artificial intelligence layering, so as to avoid the layering disorder in the case of lesions. Finally, we can get high-quality images.

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

The ultra-wide OCTA image is composed of 5 12*12mm SS-OCTA images, including five areas centered on the macular fovea, supratemporal, inferior temporal, supranasal, and subnasal. The slab we use is the inner retina slab. Then the 5 area images are stitched into an ultra-wide OCTA image through internal software. That is, each stitched image is stitched from 5 12mm * 12mm OCTA images. Each 12*12 mm volume consists of 1024 A scans per B scan (the distance between adjacent A scans is 24 mm) and 1024 B scan positions per volume scan (the distance between adjacent B scans is 24 mm). Acquire two repeated B-scans at each B-scan position to generate OCTA images. The time to obtain a single 12*12 mm volume is at least 12 seconds. All OCTA imaging is performed by professional ophthalmologists who repeat the image acquisition as many times as possible to ensure that the image has strong OCT signal penetration and minimal motion artifacts. The quality score of the scan is expressed as the SNR in decibels (dB) and ranges from 1 (poor quality) to 10 (excellent quality). Scans with a score >8 dB are considered high quality and included.

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.

All data is provided from the Shanghai Sixth People's Hospital Diabetes Diagnosis and Treatment Center.

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 labelled by ophthalmologists with more than 5 years of professional experience. We have developed a strict annotation process for the annotators to ensure the accuracy of annotation.

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.

The data are ultra-wide OCTA images. We divide the data into: training set (70%), validation set (10%), and a hidden testing set used for participant ranking (20%). The class of each image in all data is annotated by ophthalmologists. Participant teams can get training set and validation for model training and validation.

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

Training: 700 labelled images
Validation: 100 labelled images
Testing: 200 labelled images

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

In our task, the goal is to obtain a generalized network model with fewer training images. In addition, we may provide some unlabeled pictures for the challenge. We encourage researchers to use some techniques such as

unsupervised learning, semi-supervised and transfer learning.

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 division of training set, validation set and test set was completely random. We thus assume that there is no statistically significant difference between the sets.

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.

The ground truth was produced by a professional ophthalmologist from Shanghai Sixth People's Hospital. The results were checked by at least two ophthalmologists.
For the assessment of image quality level, class 0 represents poor quality level, class 1 represents good quality level, and class 2 represents excellent quality level.

The labeling process is as follows. In the first step, when an ophthalmologist labels all the images, we randomly shuffle the images and send them to the same ophthalmologist for classification. For the same image, if the classes of the first and second classifications are different, the ophthalmologist is asked to reclassify the image.

In the second step, another ophthalmologist will use the same process to complete the classification of the images. For the same picture, if the classes given by two ophthalmologists are different, the class of the image is discussed and decided by two other more experienced ophthalmologists.

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.

We have sent the instructions of the annotation process to the annotators , asking them to annotate strictly according to the process. In particular, we will evaluate the image quality from three aspects: image quality, macular clarity and retinal vascular clarity.
Poor quality level: image quality (insufficient), macular clarity (blurring), retinal vascular clarity (blurring).

Good quality level: image quality (moderate blurring or stripe noise), macular clarity (clear), retinal vascular clarity (moderate blurring).

Excellent quality level: image quality (fully visible without blurring or with slight stripe noise), macular clarity (clear), retinal vascular clarity (clear or slight blurring).

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.

All subjects involved with the annotation process are several professional ophthalmologists who have worked in the field for more than 5 years.

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.

Since this would reveal too much information and to prevent cheating, we do not fully disclose this information and are willing to share it with the reviewers if it would be helpful.

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.

There is no special data preprocessing method, we provide original resolution images and labels.

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.

For the same image with unclear quality, different doctors may mark differently. Therefore, we have formulated the above-mentioned instructions of the annotation process to improve the annotation accuracy and asking them to annotate strictly according to the process.

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

Besides annotation errors we do not expect other sources of error.

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)

Quadratic-weighted Kappa is used as the ranking metric, and the auxiliary ranking metric are macro-AUC, micro-AUC, macro-precision, macro-sensitivity and macro-specificity, respectively.
"macro" means calculate metrics for each label, and find their unweighted mean. “micro” means calculate metrics globally by counting the total true positives, true negatives, false negatives and false positives.

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

The metrics listed above are commonly used in the classification task. In particular, quadratic-weighted Kappa is often used for the multi-classification.

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.

We calculate the quadratic-weighted Kappa of all classes on the test set as the primary metric that is used for

ranking.

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

All the missing results on an image will be treated as no prediction on the image. Therefore, such cases will be treated as false negatives.

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

This method provides an performance of whether all classes are effectively classified. We give an quadratic-weighted Kappa on the test data which is not involved in the model training. This gives an emphasis on the generalization of the method.

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.

Stability will be investigated via bootstrapping and hypothesis testing

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

Bootstrapping was identified as an appropriate approach to investigate ranking variability.

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.

In the case of the same quadratic-weighted Kappa, we will further compare the other metrics of the algorithm, the order is macro-AUC, micro-AUC, macro-precision, macro-sensitivity and macro-specificity.

TASK: Detection of Proliferative Diabetic Retinopathy

SUMMARY

Keywords

List the primary keywords that characterize the task.

diabetic retinopathy, ultra-wide, optical coherence tomography angiography, grading

ORGANIZATION

Organizers

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

Bin Sheng, Shanghai Jiaotong University, China
Huating Li, Shanghai Sixth People's Hospital, China
Hao Chen, Hong Kong University of Science and Technology, Hong Kong, China
Yiyu Cai, Nanyang Technological University, Singapore
Qiang Wu, Shanghai Sixth People's Hospital, China
Weiping Jia, Shanghai Sixth People's Hospital, China
Xiangning Wang, Shanghai Sixth People's Hospital, China
Bo Qian, Shanghai Jiaotong University, China
Ruhan Liu, Shanghai Jiaotong University, China
Ling Dai, Shanghai Jiaotong University, China

b) Provide information on the primary contact person.

Bin Sheng, Shanghai Jiaotong University, China
E-mail: shengbin@cs.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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

grand-challenge.org (will setup once the proposal is accepted)

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

None at this moment.

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 policy defined.

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.

Any organization that is a member of the organizing committee does not exclude participation in the challenge, but must ensure that the content submitted is completely independent of the organizing committee members.

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

Every team can get a certificate award. In addition, we are actively seeking sponsorship, and we anticipate being able to provide cash rewards and / or graphics cards for top-3 teams. We are confident in being able to attain 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.

Results will be announced after the submission deadline.

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 will publish at least one challenge journal paper and potentially more. Participants are requested to publish a description of their method and results on arxiv.org together with their submission. All participants can be authors of the article. Participating teams are free to publish their own results in a separate publication.
Participants may publish papers including their official performance on the challenge data set, given proper reference of 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.

We will provide a link or email before the challenge starts, and the results will be submitted through the link or email. The format of the submission file is as follows:
A txt file indicating image id, with class of DR for non-PDR (0) and PDR (1).
Fine-grained instructions will be published on the challenge website.

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.

The participating teams can use the validation set we provide to evaluate their algorithms. Then they can submit the results of the algorithm on the test set to us. We will publish the score of each team's algorithm on the test set on the website. Each team is allowed to submit up to four times, and the last submitted result will be recorded as the challenge result.

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

Training data release: 01/05/2022
Validation data release: 01/06/2022
Test images release: 01/07/2022
Submission deadline: 31/08/2022
Announcement of results at MICCAI 2022: 18/09/2022

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).

We have received approval from Shanghai Sixth People's Hospital to use the data set for the purpose of research. Also, the provided challenge data is anonymized that can be used for the challenge.

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 ND.

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 a link or email for the submission of the results. For transparency, we will release the source code used for calculating final scores after the closing date of the challenge.

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

We do not force teams to open source their code.

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.

Currently, there is no explicit sponsoring of the challenge.
Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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

Research, Diagnosis, Assistance, 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 consists of patients who 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.

Patients undergoing DR screening in the hospital using the scan of ultra-wide OCTA mosaic image. All images are from untreated patients with an initial diagnosis of DR

Imaging modality(ies)

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

Ultra-wide (swept-source) optical coherence tomography angiography (UW-OCTA) mosaic image. The ultra-wide OCTA mosaic images are 2-D plane enface images.

Context information

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

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

We do not provide additional information to the dataset.

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.

The data consists of images selected from the ultra-wide OCTA mosaic image with the classes of non-PDR and PDR.

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 determine whether the given image is PDR, class 0 represents non-PDR, class 1 represents PDR.

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)).

Quadratic-weighted Kappa, macro-AUC, micro-AUC, macro-precision, macro-sensitivity and macro-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 SS-OCT/OCTA system (VG200S; SVision Imaging, Henan, China) contains an SS laser with a center wavelength of approximately 1050 nm and a scan rate of 200,000 A scans per second. The system is equipped with an eye tracking utility based on integrated confocal scanning laser ophthalmoscopes to eliminate eye movement artifacts. The axial resolution is 5 um and the lateral resolution is 13 um. The scanning depth is 3 mm.
The speed of SS-OCT used reaches hundreds of thousands of times per second, which can image faster and does not require 12 seconds. Second, we have eye tracking and dynamic image quality evaluation systems. During the scan, the patient can blink. However, we know that blinking will produce artifacts. Therefore, we subsequently used the Svision algorithm to enhance the signal of weak blood flow. At the same time, the 3D artifact removal algorithm is used to remove the tomographic blood flow artifacts in the artificial intelligence layering, so as to avoid the layering disorder in the case of lesions. Finally, we can get high-quality images.

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

The ultra-wide OCTA image is composed of 5 12*12mm SS-OCTA images, including five areas centered on the macular fovea, supratemporal, inferior temporal, supranasal, and subnasal. The slab we use is the inner retina slab. Then the 5 area images are stitched into an ultra-wide OCTA image through internal software. That is, each stitched image is stitched from 5 12mm * 12mm OCTA images. Each 12*12 mm volume consists of 1024 A scans per B scan (the distance between adjacent A scans is 24 mm) and 1024 B scan positions per volume scan (the distance between adjacent B scans is 24 mm). Acquire two repeated B-scans at each B-scan position to generate OCTA images. The time to obtain a single 12*12 mm volume is at least 12 seconds. All OCTA imaging is performed by professional ophthalmologists who repeat the image acquisition as many times as possible to ensure that the image has strong OCT signal penetration and minimal motion artifacts. The quality score of the scan is expressed as the SNR in decibels (dB) and ranges from 1 (poor quality) to 10 (excellent quality). Scans with a score >8 dB are considered high quality and included.

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.

All data is provided from the Shanghai Sixth People's Hospital Diabetes Diagnosis and Treatment Center.

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 labelled by ophthalmologists with more than 5 years of professional experience. We have developed a strict annotation process for the annotators to ensure the accuracy of annotation.

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.

The data are ultra-wide OCTA images. We divide the data into: training set (70%), validation set (10%), and a hidden testing set used for participant ranking (20%). The class of each image in all data is annotated by ophthalmologists. Participant teams can get training set and validation for model training and validation.

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

Training: 700 labelled images
Validation: 100 labelled images
Testing: 200 labelled images

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

In our task, the goal is to obtain a generalized network model with fewer training images. In addition, we may provide some unlabeled pictures for the challenge. We encourage researchers to use some techniques such as

unsupervised learning, semi-supervised and transfer learning.

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 division of training set, validation set and test set was completely random. We thus assume that there is no statistically significant difference between the sets.

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.

The ground truth was produced by a professional ophthalmologist from Shanghai Sixth People's Hospital. The results were checked by at least two ophthalmologists.
For PDR detection, class 0 represents non-PDR, class 1 represents PDR.

The labeling process is as follows. In the first step, when an ophthalmologist labels all the images, we randomly shuffle the images and send them to the same ophthalmologist for classification. For the same image, if the classes of the first and second classifications are different, the ophthalmologist is asked to reclassify the image.

In the second step, another ophthalmologist will use the same process to complete the classification of the images. For the same picture, if the classes given by two ophthalmologists are different, the class of the image is discussed and decided by two other more experienced ophthalmologists.

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.

We have sent the instructions of the annotation process to the annotators , asking them to annotate strictly according to the process. Because the criteria for DR grading are based on fundus photography and FA, all our ultra-wide OCTA images have corresponding fundus photography. In patients with suspected PDR, FA was also collected. So we will send the fundus photography and FA to the annotators to help them determine if the image is PDR.

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.

All subjects involved with the annotation process are several professional ophthalmologists who have worked in the field for more than 5 years.

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.

Since this would reveal too much information and to prevent cheating, we do not fully disclose this information and are willing to share it with the reviewers if it would be helpful.

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.

There is no special data preprocessing method, we provide original resolution images and labels.

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.

For the same image with unclear DR category, different doctors may mark differently. Therefore, we have formulated the above-mentioned instructions of the annotation process to improve the annotation accuracy and asking them to annotate strictly according to the process.

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

Besides annotation errors we do not expect other sources of error.

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)

Quadratic-weighted Kappa is used as the ranking metric, and the auxiliary ranking metric are macro-AUC, micro-AUC, macro-precision, macro-sensitivity and macro-specificity, respectively.
"macro" means calculate metrics for each label, and find their unweighted mean. “micro” means calculate metrics globally by counting the total true positives, true negatives, false negatives and false positives.

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

The metrics listed above are commonly used in the classification task. Although this is a binary classification task, we still use Kappa in task 2 as one of the evaluation metrics.

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.

We calculate the quadratic-weighted Kappa of all classes on the test set as the primary metric that is used for ranking.

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

All the missing results on an image will be treated as no prediction on the image. Therefore, such cases will be treated as false negatives.

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

This method provides an average performance of whether all classes are effectively classified. We give an quadratic-weighted Kappa on the test data which is not involved in the model training. This gives an emphasis on the generalization of the method.

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.

Stability will be investigated via bootstrapping and hypothesis testing

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

Bootstrapping was identified as an appropriate approach to investigate ranking variability.

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.

In the case of the same quadratic-weighted Kappa, we will further compare the other metrics of the algorithm, the order is macro-AUC, micro-AUC, macro-precision, macro-sensitivity and macro-specificity.

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.