DIAMOND: Device-Independent diAbetic Macular edema ONset preDiction: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

DIAMOND: Device-Independent diAbetic Macular edema ONset preDiction

Challenge acronym

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

DIAMOND

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 macular edema (DME) is a major complication of diabetes. Characterized by retinal thickening in the macula and often accompanied by hard exudate deposition, DME is a prevalent cause of vision loss among diabetic patients. The challenge's focus is on center-involved diabetic macular edema (ci-DME), a critical form of DME responsible for significant vision impairment. The presence of DME is generally assessed through 3D optical coherence tomography (OCT) imaging. A recent study has shown that the presence of ci-DME can also be assessed using 2D color fundus photography (CFP) [1].
The DIAMOND challenge seeks to revolutionize the approach to diagnosing and treating ci-DME by integrating AI and deep learning with ultra-wide-field color fundus photography (UWF-CFP), a recent evolution of CFP. More challenging than assessing the presence of ci-DME, the goal is to develop and evaluate models that can predict if a patient will develop ci-DME within a year, using UWF-CFP images alone. Because it offers a much wider field of view, we hypothesize UWF-CFP is more likely to capture early signs of DME than standard CFP. Success in this challenge could significantly improve early detection and treatment planning, reduce vision loss incidents, and exemplify AI's efficacy in healthcare.
For training, the DIAMOND Challenge uses data collected in 14 French Hospitals in the framework of the EVIRED project (https://evired.org/). EVIRED not only focuses on predicting the development of ci-DME but also broadly aims at forecasting the onset of diabetic retinopathy (DR) complications in general. Initial experiments with a simple baseline algorithm (ResNet-50) on EVIRED data suggest the feasibility of ci-DME onset prediction using UWF-CFP (area under the ROC curve = 0.73). For performance evaluation, DIAMOND will also use independent data from the LAZOUNI Ophthalmology Clinic in Tlemcen, Algeria.
By leveraging diverse datasets, including data from French and Algerian populations and images captured using different UWF-CFP devices, the challenge underscores its commitment to developing solutions that are universally applicable, transcending geographic and demographic boundaries. This generality is critical in ensuring that the

predictive models developed are robust and effective across different population groups, enhancing their clinical utility on a global scale.
Moreover, the DIAMOND Challenge introduces a significant methodological challenge that sets it apart from typical predictive modeling competitions. Participants in this challenge will not have access to the training, validation, or test datasets and, consequently, will not have the opportunity to train their models directly. Instead, they are required to submit their code, in a Docker container, which the organizing committee will then run on a specialized cloud-based cluster. This unique approach simulates a real-world scenario where data accessibility is often restricted due to privacy regulations, ethical considerations, or logistical issues.
[1] Varadarajan AV, Bavishi P, Ruamviboonsuk P, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. Nat Commun. 2020;11(1):130. doi:10.1038/s41467-019-13922-8

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Diabetic macular edema, Ultra-wide-field Color Fundus Photography, Deep Learning, Prediction, Device-independence, Population-independence

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.

Based on the participation trends observed in comparable past challenges, we anticipate around 60 teams to participate in our DIAMOND Challenge. This estimate is grounded in the numbers from similar challenges in the field. For instance, the GAMMA challenge at MICCAI 2021 saw a total of 70 teams. Similarly, the DRAC challenge at MICCAI 2022 attracted participation from 91 teams, out of which 17 submitted valid methods. The RIADD challenge at ISBI 2021, which we co-organized, had 74 submissions by various individuals and teams.
In addition to these encouraging figures from past challenges, we already have confirmed interest from five international teams, one each from China, France, Belgium, Tunisia, and Algeria.

Publication and future plans

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

We plan to coordinate a research article summarizing the challenge results. We also plan to publish a data description paper about the Algerian dataset upon completion of 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.

The challenge is planned for a 2-hour slot. During the challenge event, we will simply need regular equipment for presentations (monitor, microphones, etc.).
To guarantee equitable access to resources, every participant will be given identical computing environments, both for training and performance evaluation. The algorithms will be executed on dedicated servers provided by OVHcloud (https://www.ovhcloud.com) and funded by the EVIRED project. The data sets for training, validation, and testing are also hosted on OVHcloud, and the computation will be performed on machines equipped with NVIDIA Tesla V100S GPUs. This setup ensures that all participants have equal computational resources at their disposal.

TASK 1: Predicting the onset of ci-DME using UWF-CFP images

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.

Participants are tasked with developing models to predict the likelihood of ci-DME development within a year using UWF-CFP images. This is the sole task of this challenge.

Keywords

List the primary keywords that characterize the task.

Diabetic macular edema, Ultra-wide-field Color Fundus Photography, Deep Learning, Prediction, Device-independence, Population-independence

ORGANIZATION

Organizers

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

Mostafa El Habib Daho, University of West Brittany, Brest, France & University of Tlemcen, Tlemcen, Algeria
Mohammed El Amine Lazouni, University of Tlemcen, Tlemcen, Algeria & LAZOUNI Ophthalmology Clinic, Tlemcen, Algeria
Sarah Matta, University of West Brittany, Brest, France
Zineb Aziza Elaouaber, LAZOUNI Ophthalmology Clinic, Tlemcen, Algeria
Leila Ryma Lazouni, University of Tlemcen, Tlemcen, Algeria
Mohammed Youcef Bouayad Agha, LAZOUNI Ophthalmology Clinic, Tlemcen, Algeria
Rachid Zeghlache, University of West Brittany, Brest, France
Alireza Rezaei, University of West Brittany, Brest, France
Mathieu Lamard, University of West Brittany, Brest, France
Pierre-Henri Conze, IMT Atlantique, Brest, France
Béatrice Cochener, University of West Brittany and Brest University Hospital, Brest, France
Aude Couturier, AP-HP, Paris, France
Sophie Bonnin, Fondation Rothschild, Paris, France
Ramin Tadayoni, AP-HP, Paris, France
Gwenolé Quellec, Inserm, Brest, France

b) Provide information on the primary contact person.

Gwenolé Quellec, Research Director, Inserm, France
Email: gwenole.quellec@inserm.fr

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

Repeated annual event with different aspects of the challenge to address. New imaging modalities (OCT, OCT angiography), new targets (e.g., onset of proliferative diabetic retinopathy) and new international datasets are expected to be added.

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.

CodaLab (https://codalab.lisn.upsaclay.fr/).

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

The challenge platform will be set up once the proposal is accepted.

Participation policies

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

Only automatic methods, with blind training, are 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 for models' pretraining.

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 organizations/companies affiliated with members of the organizing committee may participate but are not eligible for awards and are not listed in the leaderboard.

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

Certificates will be provided for the TOP-3 performing teams. We are actively seeking sponsorship, and we anticipate being able to provide prizes. We are currently discussing this with Zeiss, who is a partner of the EVIRED project, as well as with OVHCloud.

e) Define the policy for result announcement.

Examples:

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

The TOP-10 performing teams based on validation data (namely the finalists) will be invited to the challenge event to present their work. The TOP-3 performing methods among the finalists, based on independent test data, will be publicly announced during the challenge event.

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 research manuscript will be prepared post-challenge. TOP-10 participating teams (finalists) may include up to two members as co-authors (up to four members for the TOP-5). Teams may publish their results separately after the challenge. Finalists must submit an abstract describing their approach and results, which will be incorporated into the challenge publication. All policies will be detailed on the challenge website, and participants must agree to these terms.
Additionally, the African dataset (from LAZOUNI Ophthalmology Clinic) will be publicly released upon completion of the challenge and a data description paper will be written.

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.

DIAMOND is a blind challenge. Participants must submit their codes for training and evaluation through a Docker submission. The organizing committee will provide a starting PyTorch Lightning-based code and Docker container with the full pipeline, which will also serve as the baseline for performance evaluation.
Upon training/validation completion, participants will receive an email with the performances of their algorithm. In detail, they will receive the training and validation scores for each user-defined checkpoint (e.g., a YAML dictionary mapping checkpoint identifiers to training/validation losses and validation metrics). Next, the validation data leaderboard will be updated with the best validation score.
For illustration purposes, participants will have access to images of one patient (sample data). To facilitate model development, an artificial dataset will be generated through random data augmentations of this sample data: it will help participants understand how the hidden challenge dataset is structured and test run their code.

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.

We will authorize each team to have 3 training/validation runs, and each training/validation run will be limited to 7 days.
Each container will be run through Slurm on one computer node with Ubuntu Server 20.04 LTS installed. Each computer node has the following resources: 4× Nvidia Tesla V100s GPUs (32Gb), 2x Intel Xeon Gold 6226R CPUs

(16 cores/32 threads - 2.9 GHz/3.9 GHz) and 384 Gb of RAM (ECC 2933 MHz). A quarter of these resources (1 GPU, 16 CPU threads, 96 Gb RAM) will be allocated to each container for 7 days.
During one training/validation run, participants are encouraged to evaluate multiple neural architectures, hyperparameters, etc. A code example will be provided to participants for that purpose. Should the run complete or crash before the end of the 7-day period (say after n days), participants will be informed and the run will be resumed once, upon request, for the remaining 7 - n days.
Progress information will be provided to participants during a training/validation run: after 1 day and at mid-run (after 3 days and a half), they will receive an email with the current checkpoint YAML dictionary.
No checkpoint will be erased until challenge completion, so checkpoints from previous runs can be loaded in subsequent training/validation runs, as well as in the final test run.
This approach ensures that teams have adequate opportunities to refine their algorithms while maintaining a structured and fair timeline for all participants.

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 period: April 1 - June 20, 2024
- Release date of sample data (one patient) and artificial dataset: April 1, 2024
- Training Docker container submission period: April 1 - June 30, 2024
- Final Docker container submission period (evaluation only): July 1 - July 5, 2024
- Finalists' announcement date: July 15, 2024
- Report submission period (for finalists): July 15 - September 15, 2024
- Release date of the results: challenge day (October 6 or 10, 2024)

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).

All examinations were conducted with patients' informed consent. The Declaration of Helsinki was followed during all procedures. The EVIRED study protocol was approved by the French South-West and Overseas Ethics Committee 4 on 28 August 2020 (Clinical Trial NCT04624737).

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)

Participants will have to sign a data usage agreement when registering to the challenge. The license for the sample data will be restrictive: 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.

In the interest of transparency, the source code utilized for computing the final scores will be disclosed in a GitHub repository, prior to the Docker container submission period.

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

Similarly, the source code of the finalist teams will be disclosed by the organizers in the same GitHub repository, upon the finalists announcement date.

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.

Several challenge authors are consultants for Zeiss, a prospective sponsor. Only the organizers will have access to the test cases and 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
-

Diagnosis, Intervention planning, Research, Prognosis, Screening

Task category(ies)

State the task category(ies)

Examples:

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

Prediction

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 includes any individual with either type 1 or type 2 diabetes, in particular those with known diabetic retinopathy. The WHO recommends that these patients undergo yearly eye examinations (https://www.who.int/europe/publications/i/item/9789289055321).

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

The challenge cohort consists of two datasets of diabetic patients, with both type 1 and type 2 diabetes, who have already been diagnosed with ci-DME a year later, as well as a control group without ci-DME diagnosed a year later. This dataset was collected in the frame of the EVIRED project (https://evired.org/) and in the LAZOUNI

Ophthalmology Clinic (Tlemcen, Algeria).

Imaging modality(ies)

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

Ultra-wide field color fundus photography (UWF-CFP).

Context information

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

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

Photograph (or photograph montage depending on the device) centered on the posterior pole of the retina.

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

Inclusion criteria: Patient with type 1 or type 2 diabetes, aged 18 years or more, diabetes duration greater than 10 years for type 1 diabetes, no previous pan-retinal photocoagulation, intravitreal injection, or previous vitrectomy. Exclusion criteria: pregnancy at baseline, ungradable UWF-CFP or OCT imaging.

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.

Retinal images of diabetic patients photographed using UWF-CFP device.

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.

Center-involved Diabetic Macular Edema (ci-DME) in a retinal image.

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)).

1) Predict with high sensitivity, specificity and precision if a patient will develop ci-DME within a year using UWF-CFP from in-domain data.
2) Predict with high sensitivity, specificity and precision if a patient will develop ci-DME within a year using UWF-CFP from out-of-domain data (data originating from LAZOUNI Ophthalmology Clinic).

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).

1) CLARUS (CLARUS 500, Carl Zeiss Meditec Inc.) and OPTOS (P200Dx, Optos plc) devices in the European dataset.
2) OPTOS (P200Dx, Optos plc) and EIDON (EIDON, iCare, Revenio Group) devices in the African dataset.

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

Images were acquired by experienced orthoptists in ophthalmology clinics.

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 European dataset utilized for training, validation, and testing purposes was collected across 14 hospital partners in France, all of which are participants in the EVIRED project. This dataset encompasses a diverse range of data points essential for our analysis.
Regarding the African dataset used for testing, these were gathered from the LAZOUNI Ophthalmology Clinic, located in Tlemcen, Algeria (https://clinique-lazouni.business.site/).

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).

Images were acquired by experienced orthoptists.

Training and test case characteristics

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

Examples:

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

A case refers to one UWF-CFP image. All training, validation, and test cases are labeled as onset of ci-DME or absence of ci-DME one year after the UWF-CFP image was taken.

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

The dataset includes:
- 7,065 cases from the EVIRED cohort. These cases are from 2,672 yearly visits of 1,917 patients (from 1 to 3 visits per patient). During one visit, both eyes are generally imaged, using one or two devices (CLARUS and/or OPTOS).
- 300 cases from the private clinic in Algeria. These cases are from 150 patients (1 visit per patient, 1 image per

eye). During the visit, each eye is imaged either by the OPTOS or by the EIDON device.
The training set contains 5,652 cases from the EVIRED cohort. The validation set contains 707 cases from the EVIRED cohort. The test set contains 706 cases from the EVIRED cohort and all 300 cases from the private clinic in Algeria. In the EVIRED cohort, all images from the same patient are assigned to the same subset (training, validation or test).
There are 7.5% of positive cases in the EVIRED cohort. The percentage of positive cases in the Algerian dataset will be disclosed upon challenge completion to ensure unbiased blind evaluation on an unseen population.

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

We have chosen standard proportions of 80:10:10 for training, validation and test sets to ensure a balanced approach in model development. The total number of samples is consistent with medical image classification studies in the literature.

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 training set includes images from two different devices, OPTOS and CLARUS, from the same cohort, promoting variability in the data. For the test set, we use images from the same cohort to maintain consistency. Additionally, we include images from a different country and continent using the OPTOS device and introduce a new device, EIDON, not used in training. This strategy is designed, according to real-world scenarios, to encourage participants to develop device-agnostic models that can accurately predict ci-DME onset independently of the population and imaging device used.

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 one experienced ophthalmologist based on data acquired one year later: one Optical Coherence Tomography (OCT) acquired from the same eye, as well as the visual acuity of this eye.

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.

Ophthalmologists were instructed to consider center-involved DME only.

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 annotation of the datasets was meticulously performed by ophthalmologists who possess a minimum of 10 years of professional experience. These medical experts conducted the annotations during clinical visits. This process was consistent across the training, validation, and test cases to ensure uniformity and reliability in the annotations. The annotations in the African center follow the same protocols.

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.

N/A

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.

Images were converted from DICOM to PNG for deidentification purposes and to ensure proper RGB encoding.

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.

Ungradable OCT images (used for ground truth acquisition) is an exclusion criterion. Inter-grader and intra-grader agreement for the presence of DME in OCT is rather high (respectively 82%-95% and 91%-98% using a single 2D scan, see https://iovs.arvojournals.org/article.aspx?articleid=2399698, probably more using the full 3D volume).

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

N/A

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)

Area under the receiver operating characteristic curve (AUC), F1-score (F1) and Expected Calibration Error (ECE). All three metrics are used for ranking: S = AUC + 0.5 x F1 + 0.5 x ECE.

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

The use of AUC in our study aligns with the goal of achieving high sensitivity and specificity in predictions. It is a commonly used metric to evaluate medical classification/prediction challenges and competitions on Grand Challenge, Kaggle, etc., as well as in the medical literature.
Two secondary metrics are considered (with a lower weight) to further assess the relevance of the computed probabilities.
- The use of F1 ensures that binary predictions (when using a probability cutoff of 0.5) are relevant. It also assesses precision (not assessed in AUC).
- The use of ECE ensures the calibration properties for a model, i.e., ensures that its predicted probabilities match the actual probabilities of the ground truth distribution.

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.

The final ranking of the submitted algorithms will be determined by a composite score that reflects their performances across four distinct test subsets (D1: CLARUS / European population, D2: OPTOS / European population, D3: OPTOS / African population and D4: EIDON / African population). This approach is specifically designed to assess the device-agnostic and population-agnostic capabilities of the proposed methods.
Each algorithm will be evaluated separately on the four test subsets. The performance on each subset D will be measured using the chosen metric, namely S(D) (with S = AUC + 0.5 x F1 + 0.5 x ECE). To arrive at the final score for each submission, we will aggregate the individual performance scores through averaging: final score = S(D1) + S(D2) + S(D3) + S(D4).
Note that, even though equal weights are given to each test subset, more emphasis is put on out-of-domain test cases, since the corresponding test subsets (D3 and D4) are smaller.

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

The tests will be run by the organizing committee. Each algorithm is expected to give one prediction for each case. It should be noted that if a Docker container fails at training or evaluation, the submission will be ignored and an error message will be reported to the participating team to fix it. In particular, a failure in the submission does not count against the 3 training submission credits allotted to each team.
A Docker container failing at evaluation means that predictions are missing for at least one test case. However, if participants fail to correct that issue, incomplete submissions will be evaluated as follows: if a prediction is missing for a test case, it will be given a default probability of 0 (meaning "absence of ci-DME one year after the UWF-CFP image was taken").

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

The rationale behind this methodology is to ensure that algorithms that perform consistently well across various devices and populations receive a higher ranking, thus highlighting their robustness and generalizability.

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.

For statistical analysis, we will primarily use the Delong test to compare receiver operating characteristic (ROC) curves.

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

The Delong test is particularly well-suited for our purposes for several reasons. Firstly, it is widely used for comparing the performance of two correlated ROC curves, making it an excellent choice for evaluating and comparing two different methods within our dataset. It allows us to statistically determine whether there is a significant difference in the performance of these methods.
Secondly, the Delong test is also useful for constructing confidence intervals for the difference between the areas under two ROC curves. This is critical in our analysis as it provides a measure of the reliability and precision of the

performance estimates of the algorithms being tested.

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.

Further analyses will be performed for the challenge summary paper in particular: combining algorithms via ensembling, ranking variability, inference times, etc. The challengeR package will be used for performance evaluation and visualization (https://github.com/wiesenfa/challengeR - https://doi.org/10.1038/s41598-021-82017-6).

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.

A study showing that ci-DME can be diagnosed using fundus photography:
- Varadarajan AV, Bavishi P, Ruamviboonsuk P, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. Nat Commun. 2020;11(1):130. doi:10.1038/s41467-019-13922-8

Publications on the EVIRED data:
- El Habib Daho M, Li Y, Zeghlache R, et al. Improved automatic diabetic retinopathy severity classification using deep multimodal fusion of UWF-CFP and OCTA images. In: Proc MICCAI OMIA Works. Springer-Verlag; 2023:11-20. doi:10.1007/978-3-031-44013-7_2
- El Habib Daho M, Li Y, Zeghlache R, et al. DISCOVER: 2-D multiview summarization of optical coherence tomography angiography for automatic diabetic retinopathy diagnosis. Artificial Intelligence in Medicine. 2024 Mar;149:102803. doi:10.1016/j.artmed.2024.102803
- Li Y, El Habib Daho M, Conze PH, et al. Hybrid Fusion of High-Resolution and Ultra-Widefield OCTA Acquisitions for the Automatic Diagnosis of Diabetic Retinopathy. Diagnostics (Basel). 2023;13(17):2770. doi:10.3390/diagnostics13172770

Further comments

Further comments from the organizers.

The organizing team has a strong track record in diabetes research, computer vision, and machine learning. The LaTIM laboratory, in particular, displays 17 years of experience in designing machine and deep learning solutions in ophthalmology. We have already organized or co-organized multiple challenges in ophthalmology:
- The RIADD challenge in 2021 (https://riadd.grand-challenge.org/) - a challenge paper is under review at Medical Image Analysis - data was released in IEEE Dataport

(https://ieee-dataport.org/open-access/retinal-fundus-multi-disease-image-dataset-rfmid)
- The IDRID challenge in 2018 (https://idrid.grand-challenge.org/) - a challenge paper was published in Medical Image Analysis (https://doi.org/10.1016/j.media.2019.101561) - data was released in IEEE Dataport (https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid)
- The CATARACTS challenge in 2017 (https://cataracts.grand-challenge.org/) - a challenge paper was published in Medical Image Analysis (https://doi.org/10.1016/j.media.2018.11.008) - data was released in IEEE Dataport (https://ieee-dataport.org/open-access/cataracts).
We also have experience in challenge participation. We won the first international challenge on DR screening (Retinopathy Online Challenge 2009) (Niemeijer M, 2010). Over the past four years, we have demonstrated remarkable success in various international medical imaging challenges, consistently achieving high rankings across different competitions. In the Automatic Detection challenge on Age-related Macular degeneration (ISBI ADAM 2020), we reached second place in AMD versus non-AMD classification and in fovea localization. In the Diabetic Retinopathy Analysis Challenge (MICCAI DRAC2022), we reached fifth place in lesion segmentation, fourth in image quality detection, and third in diabetic retinopathy classification. In the Myopic Maculopathy Analysis Challenge 2023 (MICCAI MMAC2023), we achieved eighth place in myopic maculopathy classification, second in lesion segmentation, and first in spherical equivalent prediction. Our performance in the MICCAI2023 STAGE Challenge was notable, achieving first place in the preliminary round. In the ischemic stroke segmentation challenge (ISBI2023 APIS), we ranked fifth. During the MICCAI2022 GOALS Challenge, we finished in 10th place out of 100 teams in tasks related to glaucoma diagnosis using OCT images.
Gwenolé Quellec has been a challenge reviewer for 3 MICCAI editions (2020, 2021, 2022). He received the Best Challenge Reviewer Award in 2020.