Monitoring Age-related Macular Degeneration Progression In Optical Coherence Tomography: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Monitoring Age-related Macular Degeneration Progression In Optical Coherence Tomography

Challenge acronym

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

MARIO

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.

Age-related macular degeneration (AMD) is a progressive deterioration of the macula, the central portion of the retina, affecting approximately 196 million individuals worldwide [1]. It typically manifests from the age of 50, becoming more prevalent after 65, leading to a significant decline in visual acuity without completely impairing vision. AMD is a complex and multifactorial disease, influenced by a combination of genetic and environmental risk factors. Advanced stages of the disease (atrophy and neovascularization) impact nearly 20% of patients, serving as the primary cause of severe vision loss and blindness in developed countries.

Since their introduction in 2007, anti-VEGF treatments have demonstrated remarkable efficacy in mitigating disease progression and even enhancing visual function in neovascular forms of AMD [2]. This effectiveness is optimized by minimizing the time between diagnosis and initiating treatment, along with regular follow-up examinations and retreatment as needed [3]. The indication for anti-VEGF therapy is now widely acknowledged as the presence of exudative signs (subretinal and intraretinal fluid, intraretinal hyperreflective spots, etc.) evident on optical coherence tomography (OCT) [4], a 3D imaging modality. The utilization of AI for AMD prediction [5] primarily focuses on the initial onset of early/intermediate (iAMD), atrophic (GA), and neovascular (nAMD) stages. Notably, there is a lack of research on forecasting AMD progression in patients undergoing anti-VEGF treatment monitoring.

Therefore, reliably detecting changes in neovascularization activity [6] by monitoring exudative signs is crucial for the precise implementation of individualized anti-VEGF treatment strategies. The primary objective of this challenge is to assess existing and novel algorithms [7] for recognizing the evolution of neovascularization activity in OCT scans of patients with exudative AMD, ultimately aiming to improve the planning of anti-VEGF treatments.

The challenge will address two main tasks:

Task 1: This task focuses on pairs of 2D slices (B-scans) from two consecutive OCT acquisitions. The objective is to classify the evolution between these two slices (before and after), which are typically examined side-by-side on clinicians' screens.

Task 2: This task focuses on the level of individual 2D slices. The goal is to predict the future evolution within three months for patients undergoing close monitoring as part of an anti-VEGF treatment plan.

In essence, Task 1 aims to automate the initial step of the analysis (useful for decision support), while Task 2 aims to automate the complete analysis process (valuable for autonomous AI).

Preliminary experiments utilizing a ResNet-50 Siamese [8] baseline algorithm on the dataset indicate the feasibility to predict AMD change using OCT images (Kappa score of 0.55 for task 1). To evaluate performance, the MARIO challenge will leverage independent data from the LAZOUNI Ophthalmology Clinic in Tlemcen, Algeria.

The challenge's utilization of a comprehensive dataset comprising data from both French and Algerian populations and images captured using diverse OCT devices underscores its dedication to creating universally applicable solutions that overcome geographic and demographic limitations. This cross-population applicability is paramount to ensuring that the predictive models developed are robust and effective across a wide range of patient groups. This commitment to universality further fosters the adoption of pre-training strategies and promotes the development of population-agnostic approaches, aligning seamlessly with the objectives of this challenge and the MICCAI conference.

1 - Jonas, J. B., Cheung, C. M. G., & Panda-Jonas, S. (2017). Updates on the epidemiology of age-related macular degeneration. Asia Pacific Journal of Ophthalmology (Phila), 6(6), 493-497.

2 - Rosenfeld, P. J., Brown, D. M., Heier, J. S., Boyer, D. S., Kaiser, P. K., Chung, C. Y., ... & Macular Photocoagulation Study Group. (2006). Ranibizumab for neovascular age-related macular degeneration. New England Journal of Medicine, 355(14), 1419-1431.

3 - Rasmussen, A., & Sander, B. (2014). Long-term longitudinal study of patients treated with ranibizumab for neovascular age-related macular degeneration. Current Opinion in Ophthalmology, 25(3), 158-163.

4 - Freund, K. B., Korobelnik, J.-F., Devenyi, R., Framme, C., Galic, J., Herbert, E., ... & European Society for Intravitreal Implant and Surgery, Research Committee. (2015). Treat-and-extend regimens with anti-VEGF agents in retinal diseases: A literature review and consensus recommendations. Retina (Philadelphia, Pa), 35(8), 1489-1506.

5 - Bhuiyan A, Wong TY, Ting DSW, Govindaiah A, Souied EH, Smith RT. Artificial Intelligence to Stratify Severity of Age-Related Macular Degeneration (AMD) and Predict Risk of Progression to Late AMD. Transl Vis Sci Technol. 2020 Apr 24;9(2):25. doi: 10.1167/tvst.9.2.25. PMID: 32818086; PMCID: PMC7396183.

6 - Li E, Donati S, Lindsley KB, Krzystolik MG, Virgili G. Treatment regimens for administration of anti-vascular endothelial growth factor agents for neovascular age-related macular degeneration. Cochrane Database Syst Rev. 2020 May 5;5(5):CD012208. doi: 10.1002/14651858.CD012208.pub2. PMID: 32374423; PMCID: PMC7202375.

7 - Emre, T., Chakravarty, A., Rivail, A., Riedl, S., Schmidt-Erfurth, U., Bogunovic, H. (2022). TINC: Temporally Informed Non-contrastive Learning for Disease Progression Modeling in Retinal OCT Volumes. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13432. Springer, Cham. https://doi.org/10.1007/978-3-031-16434-7 60

8 - Antoine Rivail, Ursula Schmidt-Erfurth, Wolf DieterVogel, Sebastian M. Waldstein, Sophie Riedl, Christoph Grechenig, Zhichao Wu, and Hrvoje Bogunovic, Modeling disease progression in retinal octs with longitudinal self-supervised learning, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11843 LNCS, pp. 44-52, 2019.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Age-related macular degeneration (AMD), Optical Coherence Tomography (OCT), Progression measurement, Change detection, Treament plan.

Year

2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

In case of acceptance the challenge will be held during the International Workshop on PRedictive Intelligence In MEdicine (PRIME 7 edition) workshop.

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 anticipate that this challenge will attract a substantial number of participants. We anticipate at least 20 teams and 50 attendees. We have already received interest from AI labs worldwide, including Austria, the United States, France, Algeria, and the United Kingdom. Additionally, the challenge's association with the PRIME workshop, which focuses on the analysis of longitudinal and predictive tasks in medical imaging, will undoubtedly pique the

interest of PRIME participants. Moreover, our previous challenge, RIADD (https://riadd.grand-challenge.org/evaluation/test-set-submission/leaderboard/), attracted 17 teams to the final round (and 74 to the semi-final).

Publication and future plans

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

We plan to publish at least one paper in a top medical image analysis journals. The top ten performing teams will be invited to participate in a challenge paper. The top-5 teams can have up to 5 co-authors each, while the 6th to 10th place teams can have up to 3 co-authors each. In the paper, we will publish a substantial summary of the challenge results. The participating teams may publish their own results separately after a 6 month embargo or once the challenge paper is published on Arxiv (whichever occurs first).

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.

During the PRIME workshop we will require a projector, a desktop PC, loud speakers and two/three microphones in order to host the presentations from the top performing team of the challenge along with talks from a number of invited speakers.

TASK 1: Classify evolution between two pairs of 2-D slices from two consecutive 2D OCT acquisitions.

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.

The first task focuses on pairs of 2-D slices (B-scans) from two consecutive OCT acquisitions. The goal is to classify evolution between these two slices (before and after), which clinicians typically look side by side on their screen.

The development of algorithms [7] that can help plan anti-VEGF treatment has the potential to significantly improve the outcomes of patients with age-related macular degeneration (AMD). By automating the analysis of OCT scans and predicting the future evolution of neovascular activity, these algorithms can help clinicians make more informed decisions about the timing and frequency of treatment. This, in turn, can help to slow the progression of the disease and preserve vision.

Earlier detection and treatment of neovascularization: By accurately detecting subtle changes in OCT scans, algorithms can help to identify patients who are at risk of developing neovascularization and initiate treatment sooner rather than later. This can help to prevent irreversible damage to the macula. Reduced number of treatment injections: By predicting the future evolution of neovascular activity, algorithms can help to optimize the timing of treatment injections. This can reduce the number of injections that patients need and improve their quality of life.

In summary, the first task aims to automate the initial step of the analysis (useful for decision support)

Keywords

List the primary keywords that characterize the task.

Age-related macular degeneration (AMD), Optical coherence tomography (OCT), Progression measurement, Change detection

ORGANIZATION

Organizers

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

Rachid Zeghlache (1, 2), Mathieu Lamard (1, 2), Pierre-Henri Conze (3, 2), Mostafa El Habib Daho (1,2) , Mohammed El Amine Lazouni (5) , Thomas Monfort (4), Anas-Alexis Benyoussef (4), Béatrice Cochener (1, 2, 4),Alireza Rezaei (1,2), Zineb Aziza Elaouaber (5,6), Leila Ryma Lazouni (5), Sofiane Zehar (6), Karim Boukli Hacene (6), Gwenolé Quellec (2)

1: Univ Bretagne Occidentale, Brest, F-29200 France
2: Inserm, UMR 1101, Brest, F-29200 France
3: IMT Atlantique, Brest, F-29200 France
4: Service d'Ophtalmologie, CHRU Brest, Brest, F-29200 France
5: University of Tlemcen, Algeria
6: LAZOUNI Ophthalmology Clinic, Tlemcen, Algeria

b) Provide information on the primary contact person.

Rachid Zeghlache (rachid.zeghlache@univ-brest.fr), University of West Brittany, Brest, France.

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 (Type 1 for pre-evaluation and Type 2 for the final phase).

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

mario.grand-challenge.org

Participation policies

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

Fully automatic.

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

Participants may use publicly available data with correct reference however external use of private and public data is not allowed.

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

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

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

The top ten teams will receive certificates and will be invited to be part of the challenge manuscript. No challenge prize has been secured yet, but sponsors will be looked for and still in negotiation. For instance, a total of 3600 € was awarded for our latest challenge (RIADD, sponsor: OphtAI).

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 of all methods will be announced publicly. The top-5 teams on board will be invited to give a 5 min presentation for during the PRIME workshop challenge session.

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

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

The challenge organizers aim to publish a summary paper in a relevant journal. The top ten performing teams will be invited to participate in a challenge paper and ask to submit a method description paper. The participating teams may publish their own results separately after a 6 month embargo or once the challenge paper is published on Arxiv (whichever occurs first).

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.

During the off-site evaluation period, algorithm outputs will be submitted through grand-challenge. Providing a Docker container to the organizers will be a requirement for participating in the final evaluation. This approach ensures the reproducibility and fairness of proposed solutions, enabling continued exploration and utilization by the research community.

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.

During the off-site evaluation period, algorithm outputs will be submitted through grand-challenge.

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

The following information outlines a proposed timeline for the challenge, subject to adjustments based on the MICCAI 2024 schedule and the organizers' assessment to guarantee the quality of the challenge.

Registration Period:
April 1, 2024 - July 10, 2024

Release of Training and Off-site Validation Data (Without Labels):
May 2, 2024

Off-site Result Submission Period:
May 2, 2024 - July 10, 2024

Announcement of Finalists:
July 11, 2024

Report (2-page Summary) and Docker Container Submission Period (For Finalists):
July 11, 2024 - 15 september, 2024

Release of Challenge Results:
Challenge Day (October 6 or 10, 2024) during the PRIME workshop

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).

Concerning the Data from Brest, France:

This study follows the French MR-004 methodology (CNIL), designed for research involving the re-use of anonymized data. The research must demonstrably be in the public interest. A bilingual article explaining this procedure in more detail can be found here: Link to ScienceDirect article: https://www.sciencedirect.com/science/article/abs/pii/S2352580023001399. All incoming patients sign a written informed consent form, authorizing the use of their anonymized data for research purposes. Additionally, the

study has received approval by the Institutional Review Board (IRB) of the French Society of Ophthalmology (IRB 00008855 Société Française d'Ophtalmologie IRB1).

Concerning the Data from Tlemcen, Algeria:

Similar to the data collection process in Brest, all patients in Tlemcen provided written informed consent for the use of their anonymized data for research purposes. The research adheres to the ethical principles outlined in the Declaration of Helsinki, an international standard promoting participant well-being and ethical research practices. Approval for this research was granted by the Ethics Committee Board of the LAZOUNI Ophthalmology Clinic for the use of anonymized OCT images of patients with wet AMD for research purposes as described above.

For further inquiries or verification of this ethical approval, please contact:

LAZOUNI Ophthalmology Clinic, Imama, Mansourah, 13000, Tlemcen, Algeria
Email: cliniquelazouni@gmail.com
Phone: +213-500-568-090
Data Use Agreement (DUA) for All Participating Teams:

Each participating team is required to submit a signed Data Use Agreement (DUA) verifying their institutional affiliation and agreeing to all conditions specified in the Terms of Use.

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)

The data associated with this challenge is currently available for restricted use. This means it can only be used for the specific purpose of participating in the challenge and for the duration of the challenge itself. Distribution of the data to third parties is prohibited. Following the successful completion of the challenge and acceptance of the associated paper, the data will be released publicly under the terms of the Creative Commons Attribution (CC-BY) license, as was done with data from previous challenges (https://ieee-dataport.org/open-access/cataracts). We are currently actively negotiating a transition to the CC-BY 4.0 license, which offers broader usage rights. This includes allowing commercial entities like industry partners to utilize the data for purposes such as prototyping new technologies or developing robust models for clinical deployment. This broader license will ultimately maximize the impact of the provided data, benefiting not only the research community but also patients through the

development of improved clinical solutions.

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.

A cross-platform Python evaluation script, employed to generate the ranking, will be made available on the challenge's official website. Additionally, comprehensive instructions for submission and docker image creation will be accessible on the challenge's associated GitHub repository. To further support participants, we will provide a baseline algorithm code and weights.

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

Participating teams will be encouraged to release their code on a public code repository, but this will not be a requirement. Naturally, providing a Docker container to the organizers will be a requirement for participating in the final evaluation.

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.

Sponsoring information is not known yet. All funding from sponsors will be for participant prizes. Challenge organization was funded exclusively by the organizer's institutions. Only the organizing team will have access to the test (and validation) case labels.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Intervention planning, Intervention follow up.

Task category(ies)

State the task category(ies)

Examples:

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

Classification, Change detection.

Cohorts

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

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

Patients with wet AMD followed-up for intravitreal injection planning.

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

Patients with wet AMD followed-up for intravitreal injection planning.

Imaging modality(ies)

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

Imaging data from Brest University Hospital (France) was acquired with a Spectralis OCT device (Heidelberg Engineering).
Imaging data from LAZOUNI Ophthalmology Clinic (Algeria) was acquired with an Avanti XR OCT device (Optovue).

Context information

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

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

None (in particular, the acquisition dates are masked).

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

Age, gender, medical history, and date of dose injection information will be included in the metadata. Additionally, an indicator will be incorporated to identify unregistered image pairs.

Target entity(ies)

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

Eye (posterior segment). The Data was collected in two countries. The French dataset was collected at the Ophthalmology department of Brest University Hospital (France) and the Algerian dataset et LAZOUNI Ophthalmology Clinic in Tlemcen. All patients are followed-up for monitoring and treatment of neovascular AMD.

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.

Find change detection algorithm for classifying evolution between two 2-D OCT slices (B-scans) from two consecutive OCT acquisitions. Image-level labels were reordered, to make weighted Kappa scoring possible.
Each task will be using 2D slices + (localizers and clinical data it's up to the participant choice)

Task 1 (Detection of change between two consecutives visits, three classes are defined + 1 class for event that are not of interest)

Were we have the following class

<REDUCED:0>: <ELIMINATED: 1> or <PERSISTENT_REDUCED: 2>

<STABLE:1>: <INACTIVE: 0> or <PERSISTENT_STABLE: 3>

<WORSENED:2>: <PERSISTENT_WORSENED: 4> or <APPEARED: 6>

<OTHER:-1> class is defined for <ININTERPRETABLE: -1> and <APPEARED_AND_ELIMINATED: 5>.

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)).

Task 1: Change Detection in OCT Scans

This task focuses on identifying changes in the retinal tissue between two consecutive OCT acquisitions using a 2-dimensional image processing algorithm. The algorithm aims to classify whether "change" or "no change" has occurred by analyzing the corresponding B-scans (2D OCT slices).

Key Points:

- Change Detection Focus: This task prioritizes algorithms that can accurately distinguish between changes and no changes in the retinal tissue.
- Emphasis on Classification Metrics: While standard evaluation metrics like specificity, sensitivity, and accuracy will be considered, robustness is also crucial for ensuring the algorithm performs well under various conditions.
- Prioritization of QWK: The quadratic weighted kappa (QWK) metric is particularly important as it accounts for chance agreements, ensuring the algorithm performs better than random guessing.

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).

Imaging data was acquired with a Spectralis OCT device (Heidelberg Engineering).

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

French dataset:

- The B-scan definition was 496x1024 pixels, with a variable number of B-scans per C-scan (typically 49 B-scans).
- Consecutive examinations were registered using the Spectralis follow-up option.
- The data was exported in XML+bitmap format using the Spectralis software version 6.9.4.0.

Algerian dataset:

- The B-scan definition was 981x1622 pixels, with a variable number of B-scans per C-scan (typically 49 B-scans).
- Consecutive examinations were registered using the Spectralis follow-up option.
- The data was exported in JPEG format using the RTVue XR software version 2017.0.0.16.

Image format:

All images will be provided as RGB images.

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.

Data was collected at the Ophthalmology department of Brest University Hospital (France). All patients are followed-up for monitoring and treatment of neovascular AMD. Regarding the African datasets 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).

Imaging data for both centers was collected by ophthalmologists (retina specialists) with at least two years of experience.

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 represents the complete dataset for one patient. It typically includes 10 3D volumes (C-scans) of OCT acquisitions per eye (two eyes per patient) plus a corresponding 2D infrared image (localizer) for each 3D volume and relevant clinical information (e.g., age, gender, medical history, date of dose injection).

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

Data Distribution

The dataset originates from two centers: Brest, France, and Tlemcen, Algeria.

Brest, France:

- Patients: 136 (both eyes included)
- Training: 68 patients (14,496 B-scan pairs)
- Validation (off-site evaluation): 34 patients (7,010 B-scan pairs)
- Test (on-site evaluation): 34 patients (8,341 B-scan pairs) + 30 additional patients (details to be confirmed) from Tlemcen

Tlemcen, Algeria:

- Patients: 100 (both eyes included)
- Test: 30 patients (selected from the 100)
- Training: The remaining 70 patients will be withheld, unlabeled, to encourage pre-training strategies and the development of population-agnostic methods in alignment with the MICCAI conference's goals.

Class Distribution (French Dataset):

- Reduced (Class 0): Includes labels: eliminated (1) or persistent reduced (2) (Total: 4280 pairs)
- Stable (Class 1): Includes labels: inactive (0) or persistent stable (3) (Total: 20141 pairs)
- Worsened (Class 2): Includes labels: persistent worsened (4) or appeared (6) (Total: 3114 pairs)
- Other (Class -1): Includes labels: uninterpretable (-1) and appeared and eliminated (5) (Total: 2312 pairs)

Note: The exact number of data points and patient involvement for the Algerian test set will be confirmed and disclosed later.

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

Data export was manual and therefore time-consuming. Manual annotation of thousands of image pairs by two retinal experts was also a time-consuming task. This was the reason for not collecting a larger dataset. However, this is larger than a similar study published by one of the organizers (Quellec, G., Kowal, J., Hasler, P.W., Scholl, H.P.N., Zweifel, S., Konstantinos, B., de Carvalho, J.E.R., Heeren, T., Egan, C., Tufail, A. and Maloca, P.M. (2019), Feasibility of support vector machine learning in age-related macular degeneration using small sample yielding sparse optical coherence tomography data. Acta Ophthalmol, 97: e719-e728. https://doi.org/10.1111/aos.14055), which involved 70 patients, as opposed to (136 + 100) in our case. A ratio of 2:1:1 was used for training:validation:testing.

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.

Patient Selection and Class Distribution:

For this challenge, patients were randomly selected from a population diagnosed with neovascular age-related macular degeneration (AMD). This sampling method ensures the class distribution within the selected set reflects the real-world distribution of AMD progression, increasing the generalizability of the trained model. While random sampling is widely used for training data selection, it carries a slight risk of underrepresenting certain classes in the chosen subset. To mitigate this risk, stratified sampling was imposed where the sample proportions mirror the class distribution of the entire population. This was performed by randomly sampling multiple time distribution a large number of time and keeping the distribution that best reflect the class distribution in each split

Data Splitting:

The selected patients were then randomly assigned to separate training, validation, and test sets. Importantly, data separation occurred at the patient level, meaning entire patients were assigned to a specific set, ensuring consistency in annotation format across all sets (training, validation, and test). This approach also prevents data

leakage, where information from the test set might influence the training process and lead to overly optimistic performance evaluations.

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.

In both centers for the test cases, two ophthalmologists performed the annotation independently. For training and validation cases, one ophthalmologist performed the annotation.

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.

An online annotation tool was designed for the study.
Consecutive C-scans were viewed jointly on the same screen (older examination at the top, newer examination at the bottom). With a slider, the annotator could browse through pairs of matched B-scans from the two consecutive C-scan.

For each pair of consecutive C-scan, the annotator had to assign one of the following 3 labels: improvement (increase delay before next examination) / stability (maintain delay) / worsening (reduce delay).
5 patients (from the training set) were used to train the annotators. After annotating these 5 patients: 1) annotations were compared by the challenge organizers and presented to the annotators, 2) the organizers discussed their annotation strategy with each other and 3) they were given the opportunity to revise their annotations. Next, they annotated the remaining patients independently.

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 annotators are ophthalmologists (retina specialists) with at least two years of experience in neovascular AMD patient monitoring. 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.

Annotations from both human graders will not be merged. Agreement between all annotations will be used as a baseline to assess the agreement between an algorithm and a human grader.

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.

For the French dataset:
- The XML metadata files were de-identified and all examination dates were masked (as they may be used to predict treatment decisions).
- Bitmap images (BMP) are converted to PNG to save space.

- Two consecutives 3-D volumes are registered.
For the Algerian dataset:
The pipeline is the same except the fact that image are directly save a JPG

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 main sources of inter-annotator variability are 1) the distinction between an absence of disease activity and a stable activity (two types of non-evolution), and 2) the distinction between an eliminated activity and a reduced (but not fully eliminated) activity.

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

Another source of error arised during image acquisition: the operator sometimes forgot to activate the follow-up mode, resulting in non-registered OCT volumes. This occured in about 10% of consecutive acquisitions.

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 (QWK), F1 score, Specificity, Rk-correlation.

Additionally, let a2 denote the automatic annotations. Let h2a and h2b denote the human annotations for the task. Agreement scores for task 2 (k2) is defined as follows:
Average quadratic-weighted Kappa with a human grader (k2 = 0.5 weighted_kappa(a2, h2a) + 0.5 weighted_kappa(a2, h2b)).

Note that there is one label per pair of B-scans. We will not compute a k2 score per case and then average it. Instead, we will directly compute a single k2 score across all pairs. This is because some label categories in some patients.

We will release the computing for all metrics in challenge website mario.grand-challenge.org.

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

The labels are ordered categorical variables (ordinal variables), so quadratic-weighted Kappa is used. F1 score, and specificity are commonly employed metrics to assess the performance of multi-class classification methods, particularly in biomedical applications where classes represent distinct severity. F1 score harmonizes precision and recall, while specificity quantifies the proportion of truly negative samples accurately classified. These metrics

provide valuable insights into the effectiveness of classification models in biomedical contexts. Rk-correlation offers a compelling approach for evaluating model performance in medical data challenges, especially when dealing with class imbalance, multi-class classification, and the need for interpretable metrics. It provides a valuable complement to traditional metrics and helps researchers gain deeper insights into their models' effectiveness in real-world medical applications.

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 simultaneous utilization of quadratic-weighted Kappa, F1 score, and specificity, provides a more comprehensive assessment of the classification model, reducing the risk of bias.

Rk-correlation coefficient (RkC) provides a valuable alternative to traditional accuracy by focusing on agreement across all classes and offering a clear interpretation of performance.

Additionally, in the final ranking, task 2 will be assigned a higher weight due to its greater challenge and clinical significance. Finally, given that the labels are ordered categorical variables, quadratic-weighted Kappa will be given greater emphasis.

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

Submissions with missing results will not be evaluated. Only participants with comprehensive results for the task will appear in the final ranking. Our submission system will provide feedback error to the participant and will not output the calculation results of the metric.

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

Separate metric scores are computed for quadratic-weighted Kappa, F1 score and specificity and RkC on all test cases.
Second, the ranking score is obtained by averaging the scores of all metrics. The achieved quadratic-weighted Kappa will be used as a tie-break. algorithms will be ranked by decreasing order of the average (Cohen or quadratic-weighted) Kappa (i.e. k1 or k2). Additionally, a final ranking will be produced to determine the challenge winner. Only participants with a result for both tasks will be ranked. The aggregate score (k) will be defined as k = k1 + 2 k2. Algorithms will be sorted by
decreasing order of k.

Statistical analyses

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

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

The variability of ranking (for the final ranking, which implies an aggregation of two metrics) will be assessed through Kendall's tau analysis. The statistical analysis will be performed in Python and R (scikit-learn for Kappa computation, challengeR for Kendall's tau analysis).

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

Kendall's tau is non-parametric, which means that it does not make any assumptions about the distribution of the data. It is robust to outliers. It is relatively easy to interpret and was recommended from this study for ranking scheme evaluation: Maier-Hein, L., Eisenmann, M., Reinke, A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 9, 5217 (2018).

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.

Utilizing the ChallengeR toolkit (Wiesenfarth, M., Reinke, A., Landman, B.A., et al. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci Rep 11, 2369 (2021). https://doi.org/10.1038/s41598-021-82017-6), we will conduct thorough analyses to investigate the ranking scheme evaluation (Kendall's tau), algorithm ensemble strategies, and inter-algorithm, inter-human, and human-algorithm variability. The developed code for this competition analysis will be made publicly accessible to foster transparency and reproducibility.

TASK 2: Prediction of evolution within 3 mounts of AMD on OCT 2D slices for planning treatment anti-VEGF.

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.

The second task delves into the level of individual 2D slices, aiming to predict the future evolution within a three-month timeframe for patients undergoing close monitoring as part of an anti-VEGF treatment plan.

Personalized treatment plans: Algorithms can be utilized to create personalized treatment regimens for each patient based on their individual risk factors and OCT scan findings [8]. This can help ensure that patients receive the most effective treatment for their condition.

In conclusion, this challenge holds the potential to make a substantial contribution to the field of AMD treatment by developing AI algorithms that can automatically detect and classify the evolution of neovascular AMD activity in OCT scans. The second task aims to automate the entire analysis process (useful for autonomous AI). These algorithms could aid physicians in enhancing patient care, reducing costs, and improving quality of life.

Keywords

List the primary keywords that characterize the task.

Age-related macular degeneration (AMD), Optical coherence tomography (OCT), Progression measurement, Treatment plan

ORGANIZATION

Organizers

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

Rachid Zeghlache (1, 2), Mathieu Lamard (1, 2), Pierre-Henri Conze (3, 2), Gwenolé Quellec (2), Thomas Monfort (4), Anas-Alexis Benyoussef (4), Béatrice Cochener (1, 2, 4) Mostafa El Habib Daho (1,2) + (Mohammed El Amine Lazouni (5) , Zineb Aziza Elaouaber (5,6), Leila Ryma Lazouni (5), Sofiane Zehar (6), Karim Boukli Hacene (6)

1: Univ Bretagne Occidentale, Brest, F-29200 France
2: Inserm, UMR 1101, Brest, F-29200 France
3: IMT Atlantique, Brest, F-29200 France
4: Service d'Ophtalmologie, CHRU Brest, Brest, F-29200 France
5: University of Tlemcen, Algeria
6: LAZOUNI Ophthalmology Clinic, Tlemcen, Algeria

b) Provide information on the primary contact person.

Rachid Zeghlache (rachid.zeghlache@univ-brest.fr), University of West Brittany, Brest, France.

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 (Type 1 for pre-evaluation and Type 2 for the final phase).

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

mario.grand-challenge.org

Participation policies

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

Fully automatic.

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

Participants may use publicly available data with correct reference however external use of private and public data is not allowed.

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

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

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

The top ten teams will receive certificates and will be invited to be part of the challenge manuscript. No challenge prize has been secured yet, but sponsors will be looked for and still in negotiation. For instance, a total of 3600 € was awarded for our latest challenge (RIADD, sponsor: OphtAI).

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 of all methods will be announced publicly. The top-5 teams on board will be invited to give a 5 min presentation for during the PRIME workshop challenge session.

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

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

The challenge organizers aim to publish a summary paper in a relevant journal. The top ten performing teams will be invited to participate in a challenge paper and ask to submit a method description paper. The participating teams may publish their own results separately after a 6 month embargo or once the challenge paper is published on Arxiv (whichever occurs first).

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.

During the off-site evaluation period, algorithm outputs will be submitted through grand-challenge. Providing a Docker container to the organizers will be a requirement for participating in the final evaluation. This approach ensures the reproducibility and fairness of proposed solutions, enabling continued exploration and utilization by the research community.

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.

During the off-site evaluation period, algorithm outputs will be submitted through grand-challenge.

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

The following information outlines a proposed timeline for the challenge, subject to adjustments based on the MICCAI 2024 schedule and the organizers' assessment to guarantee the quality of the challenge.

Registration Period:
April 1, 2024 - July 10, 2024

Release of Training and Off-site Validation Data (Without Labels):
May 2, 2024

Off-site Result Submission Period:
May 2, 2024 - July 10, 2024

Announcement of Finalists:
July 11, 2024

Report (2-page Summary) and Docker Container Submission Period (For Finalists):
July 11, 2024 - 15 september, 2024

Release of Challenge Results:
Challenge Day (October 6 or 10, 2024) during the PRIME workshop

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).

Concerning the Data from Brest, France:

This study follows the French MR-004 methodology (CNIL), designed for research involving the re-use of anonymized data. The research must demonstrably be in the public interest. A bilingual article explaining this procedure in more detail can be found here: Link to ScienceDirect article: https://www.sciencedirect.com/science/article/abs/pii/S2352580023001399. All incoming patients sign a written informed consent form, authorizing the use of their anonymized data for research purposes. Additionally, the study has received approval by the Institutional Review Board (IRB) of the French Society of Ophthalmology (IRB 00008855 Société Française d'Ophtalmologie IRB1).

Concerning the Data from Tlemcen, Algeria:

Similar to the data collection process in Brest, all patients in Tlemcen provided written informed consent for the use of their anonymized data for research purposes. The research adheres to the ethical principles outlined in the Declaration of Helsinki, an international standard promoting participant well-being and ethical research practices. Approval for this research was granted by the Ethics Committee Board of the LAZOUNI Ophthalmology Clinic for

the use of anonymized OCT images of patients with wet AMD for research purposes as described above.

For further inquiries or verification of this ethical approval, please contact:

LAZOUNI Ophthalmology Clinic, Imama, Mansourah, 13000, Tlemcen, Algeria
Email: cliniquelazouni@gmail.com
Phone: +213-500-568-090
Data Use Agreement (DUA) for All Participating Teams:

Each participating team is required to submit a signed Data Use Agreement (DUA) verifying their institutional affiliation and agreeing to all conditions specified in the Terms of Use.

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)

The data associated with this challenge is currently available for restricted use. This means it can only be used for the specific purpose of participating in the challenge and for the duration of the challenge itself. Distribution of the data to third parties is prohibited. Following the successful completion of the challenge and acceptance of the associated paper, the data will be released publicly under the terms of the Creative Commons Attribution (CC-BY) license, as was done with data from previous challenges (https://ieee-dataport.org/open-access/cataracts). We are currently actively negotiating a transition to the CC-BY 4.0 license, which offers broader usage rights. This includes allowing commercial entities like industry partners to utilize the data for purposes such as prototyping new technologies or developing robust models for clinical deployment. This broader license will ultimately maximize the impact of the provided data, benefiting not only the research community but also patients through the development of improved clinical solutions.

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.

A cross-platform Python evaluation script, employed to generate the ranking, will be made available on the challenge's official website. Additionally, comprehensive instructions for submission and docker image creation will be accessible on the challenge's associated GitHub repository. To further support participants, we will provide a baseline algorithm code and weights.

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

Participating teams will be encouraged to release their code on a public code repository, but this will not be a requirement. Naturally, providing a Docker container to the organizers will be a requirement for participating in the final evaluation.

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.

Sponsoring information is not known yet. All funding from sponsors will be for participant prizes. Challenge organization was funded exclusively by the organizer's institutions. Only the organizing team will have access to the test (and validation) case labels.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Intervention planning, Intervention follow up.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling

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

Classification, 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.

Patients with wet AMD followed-up for intravitreal injection planning.

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

Patients with wet AMD followed-up for intravitreal injection planning.

Imaging modality(ies)

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

Imaging data from Brest University Hospital (France) was acquired with a Spectralis OCT device (Heidelberg Engineering).
Imaging data from LAZOUNI Ophthalmology Clinic (Algeria) was acquired with an Avanti XR OCT device (Optovue).

Context information

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

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

None (in particular, the acquisition dates are masked).

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

Age, gender, medical history, date of dose injection information.

Target entity(ies)

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

Eye (posterior segment). The Data was collected in two countries. The French dataset was collected at the Ophthalmology department of Brest University Hospital (France) and the Algerian dataset et LAZOUNI Ophthalmology Clinic in Tlemcen. All patients are followed-up for monitoring and treatment of neovascular AMD.

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.

Predict the progression of the AMD within 3 months with an algorithm using 2-D OCT slices (B-scans).
Image-level labels were reordered, to make weighted Kappa scoring possible.
Each task will be using 2D slices + (localizers and clinical data it's up to the participant choice)

Task 2 (prediction of the future development of the DMLA within 3 month)

The following label will be used for the prediction of evolution

<REDUCED:0>: <ELIMINATED: 1> or <PERSISTENT_REDUCED: 2>

<STABLE:1>: <INACTIVE: 0> or <PERSISTENT_STABLE: 3>

<WORSENED:2>: <PERSISTENT_WORSENED: 4> or <APPEARED: 6>

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)).

Task 2: Predicting the Progression of Age-related Macular Degeneration (AMD)

This task aims to forecast the course of AMD within a 3-month timeframe based on a single 2-dimensional OCT image (B-scan) using an image processing algorithm. While standard evaluation metrics like specificity, sensitivity, accuracy, and robustness will be considered, the quadratic weighted kappa (QWK) will be prioritized for evaluating accurate predictions, particularly for the algorithm's target categories (reduced, stable, worsened).

Key Points:

- Focus on Future Progression: Unlike Task 1, this task focuses on predicting future disease progression, not just change detection.
- Shared Metrics: Similar to Task 1, standard evaluation metrics are used.
- Emphasis on QWK: The QWK metric is particularly crucial for assessing the model's performance in predicting specific categories.

Clinical Significance of Task 2:

Compared to Task 1, Task 2 holds greater clinical significance due to its ability to predict future disease progression. This allows for:

- Proactive Intervention: Early detection enables timely adjustments to treatment plans, potentially delaying or preventing vision loss.
- Improved Patient Management: Clinicians can tailor treatment plans and resource allocation based on the predicted disease course.
- Enhanced Research: Reliable prediction models can facilitate research by identifying high-risk individuals for clinical trials and accelerating the development of new therapies.

Therefore, while both tasks contribute to ophthalmic diagnosis and care, Task 2's ability to predict future progression elevates its importance for improving patient outcomes and driving advancements in AMD research and treatment.

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).

Imaging data was acquired with a Spectralis OCT device (Heidelberg Engineering).

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

French dataset:

- The B-scan definition was 496x1024 pixels, with a variable number of B-scans per C-scan (typically 49 B-scans).
- Consecutive examinations were registered using the Spectralis follow-up option.
- The data was exported in XML+bitmap format using the Spectralis software version 6.9.4.0.

Algerian dataset:

- The B-scan definition was 981x1622 pixels, with a variable number of B-scans per C-scan (typically 49 B-scans).
- Consecutive examinations were registered using the Spectralis follow-up option.
- The data was exported in JPEG format using the RTVue XR software version 2017.0.0.16.

Image format:

All images will be provided as RGB images.

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.

Data was collected at the Ophthalmology department of Brest University Hospital (France). All patients are followed-up for monitoring and treatment of neovascular AMD. Regarding the African datasets 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).

Imaging data for both centers was collected by ophthalmologists (retina specialists) with at least two years of experience.

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 represents the complete dataset for one patient. It typically includes 10 3D volumes (C-scans) of OCT acquisitions per eye (two eyes per patient) plus a corresponding 2D infrared image (localizer) for each 3D volume and relevant clinical information (e.g., age, gender, medical history, date of dose injection).

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

Data Description:

The dataset originates from two clinical centers: Brest, France, and Tlemcen, Algeria.

Data Acquisition:

Brest, France:

- Training set: 68 patients (5257 B-scans)
- Validation set (off-site evaluation): 34 patients (2486 B-scans)
- Test set (on-site evaluation): 64 patients (7291 B-scans)
- 34 patients from Brest
- 30 patients (to be specified later) from Tlemcen

Tlemcen, Algeria:

- Raw data: 100 patients with both eyes (number of B-scan pairs not specified)
- Test set: 30 patients (from the 100 collected) will be released for testing purposes.
- Training data: The remaining 70 patients will be withheld from participants without annotations to encourage pre-training strategies and the development of population-agnostic methods, aligning with the MICCAI conference's goals.

Class Distribution (French Dataset):

- Reduced (Class 0): Includes labels "eliminated (1)" or "persistent reduced (2)" (Total: 1274)
- Stable (Class 1): Includes labels "inactive (0)" or "persistent stable (3)" (Total: 8238)
- Worsened (Class 2): Includes labels "persistent worsened (4)" or "appeared (6)" (Total: 761)

Note: The class distribution for the Algerian test set will be disclosed later.

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

Data export was manual and therefore time-consuming. Manual annotation of thousands of image pairs by two retinal experts was also a time-consuming task. This was the reason for not collecting a larger dataset. However, this is larger than a similar study published by one of the organizers (Quellec, G., Kowal, J., Hasler, P.W., Scholl, H.P.N., Zweifel, S., Konstantinos, B., de Carvalho, J.E.R., Heeren, T., Egan, C., Tufail, A. and Maloca, P.M. (2019), Feasibility of support vector machine learning in age-related macular degeneration using small sample yielding sparse optical coherence tomography data. Acta Ophthalmol, 97: e719-e728. https://doi.org/10.1111/aos.14055), which involved 70 patients, as opposed to (136 + 100) here.
A ratio of 2:1:1 was used for training:validation:testing.

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.

Patient Selection and Class Distribution:

For this challenge, patients were randomly selected from a population diagnosed with neovascular age-related macular degeneration (AMD). This sampling method ensures the class distribution within the selected set reflects the real-world distribution of AMD progression, increasing the generalizability of the trained model. While random sampling is widely used for training data selection, it carries a slight risk of underrepresenting certain classes in the chosen subset. To mitigate this risk, stratified sampling was imposed where the sample proportions mirror the class distribution of the entire population. This was performed by randomly sampling multiple time distribution a large number of time and keeping the distribution that best reflect the class distribution in each split

Data Splitting:

The selected patients were then randomly assigned to separate training, validation, and test sets. Importantly, data separation occurred at the patient level, meaning entire patients were assigned to a specific set, ensuring

consistency in annotation format across all sets (training, validation, and test). This approach also prevents data leakage, where information from the test set might influence the training process and lead to overly optimistic performance evaluations.

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.

In both centers for the test cases, two ophthalmologists performed the annotation independently. For training and validation cases, one ophthalmologist performed the annotation.

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.

An online annotation tool was designed for the study.
Consecutive B-scans were viewed jointly on the same screen (older examination at the top, newer examination at the bottom).

For each pair of matched B-scans, the annotator had to assign one of the following 7 labels: uninterpretable / no activity / activity appearance / persistent activity (worsened) / persistent activity (stable) / persistent activity (improved) / activity disappearance.
5 patients (from the training set) were used to train the annotators. After annotating these 5 patients: 1) annotations were compared by the challenge organizers and presented to the annotators, 2) the organizers discussed their annotation strategy with each other and 3) they were given the opportunity to revise their annotations. Next, they annotated the remaining 131 patients independently.

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 annotators are ophthalmologists (retina specialists) with at least two years of experience in neovascular AMD patient monitoring. 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.

Annotations from both human graders will not be merged. Agreement between all annotations will be used as a baseline to assess the agreement between an algorithm and a human grader.

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.

- The XML metadata files were de-identified and all examination dates were masked (as they may be used to predict treatment decisions).
- Bitmap images (BMP) are converted to PNG to save space.

- Two consecutives 3-D volumes are registered.
For the Algerian dataset:
The pipeline is the same except the fact that image are directly save a JPG

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 main sources of inter-annotator variability are 1) the distinction between an absence of disease activity and a stable activity (two types of non-evolution), and 2) the distinction between an eliminated activity and a reduced (but not fully eliminated) activity.

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

Another source of error arised during image acquisition: the operator sometimes forgot to activate the follow-up mode, resulting in non-registered OCT volumes. This occured in about 10% of consecutive acquisitions.

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 (QWK), F1 score, Specificity, Rk-correlation

Additionally, let a2 denote the automatic annotations. Let h2a and h2b denote the human annotations for the task. Agreement
scores for task 2 (k2) is defined as follows:
Average quadratic-weighted Kappa with a human grader (k2 = 0.5 weighted_kappa(a2, h2a) + 0.5 weighted_kappa(a2, h2b)).

Note that there is one label for each B-scans. We will not compute a k2 score per case and then average it. Instead, we will directly compute a single k2 score across all pairs. This is because some label categories in some patients.

We will release the computing for all metrics in challenge website mario.grand-challenge.org.

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

The labels are ordered categorical variables (ordinal variables), so quadratic-weighted Kappa is used. F1 score, and specificity are commonly employed metrics to assess the performance of multi-class classification methods, particularly in biomedical applications where classes represent distinct severities. F1 score harmonizes precision

and recall, while specificity quantifies the proportion of truly negative samples accurately classified. These metrics provide valuable insights into the effectiveness of classification models in biomedical contexts. Rk-correlation offers a compelling approach for evaluating model performance in medical data challenges, especially when dealing with class imbalance, multi-class classification, and the need for interpretable metrics. It provides a valuable complement to traditional metrics and helps researchers gain deeper insights into their models' effectiveness in real-world medical applications.

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 simultaneous utilization of quadratic-weighted Kappa, F1 score, and specificity, provides a more comprehensive assessment of the classification model, reducing the risk of bias.

Rk-correlation coefficient (RkC) provides a valuable alternative to traditional accuracy by focusing on agreement across all classes and offering a clear interpretation of performance.

Additionally, in the final ranking, task 2 will be assigned a higher weight due to its greater challenge and clinical significance. Finally, given that the labels are ordered categorical variables, quadratic-weighted Kappa will be given greater emphasis.

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

Submissions with missing results will not be evaluated. Only participants with comprehensive results for the task will appear in the final ranking. Our submission system will provide feedback error to the participant and will not output the calculation results of the metric.

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

Separate metric scores are computed for quadratic-weighted Kappa, F1 score and specificity and RkC on all test cases.
Second, the ranking score is obtained by averaging the scores of all metrics. The achieved quadratic-weighted Kappa will be used as a tie-break. algorithms will be ranked by decreasing order of the average (Cohen or quadratic-weighted) Kappa (i.e. k1 or k2). Additionally, a final ranking will be produced to determine the challenge winner. Only participants with a result for both tasks will be ranked. The aggregate score (k) will be defined as k = k1 + 2 k2. Algorithms will be sorted by
decreasing order of k.

Statistical analyses

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

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

The variability of ranking (for the final ranking, which implies an aggregation of two metrics) will be assessed through Kendall's tau analysis. The statistical analysis will be performed in Python and R (scikit-learn for Kappa computation, challengeR for Kendall's tau analysis).

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

Kendall's tau is non-parametric, which means that it does not make any assumptions about the distribution of the data. It is robust to outliers. It is relatively easy to interpret and was recommended from this study for ranking scheme evaluation: Maier-Hein, L., Eisenmann, M., Reinke, A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 9, 5217 (2018).

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.

Utilizing the ChallengeR toolkit (Wiesenfarth, M., Reinke, A., Landman, B.A., et al. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci Rep 11, 2369 (2021). https://doi.org/10.1038/s41598-021-82017-6), we will conduct thorough analyses to investigate the ranking scheme evaluation (Kendall's tau), algorithm ensemble strategies, and inter-algorithm, inter-human, and human-algorithm variability. The developed code for this competition analysis will be made publicly accessible to foster transparency and reproducibility.

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.

1 - Jonas, J. B., Cheung, C. M. G., & Panda-Jonas, S. (2017). Updates on the epidemiology of age-related macular degeneration. Asia Pacific Journal of Ophthalmology (Phila), 6(6), 493-497.
Rosenfeld, P. J., Brown, D. M., Heier, J. S., Boyer, D. S., Kaiser, P. K., Chung, C. Y., ... & Macular Photocoagulation Study Group. (2006). Ranibizumab for neovascular age-related macular degeneration. New England Journal of Medicine, 355(14), 1419-1431.

2 - Rasmussen, A., & Sander, B. (2014). Long-term longitudinal study of patients treated with ranibizumab for neovascular age-related macular degeneration. Current Opinion in Ophthalmology, 25(3), 158-163.

3 - Freund, K. B., Korobelnik, J.-F., Devenyi, R., Framme, C., Galic, J., Herbert, E., ... & European Society for Intravitreal Implant and Surgery, Research Committee. (2015). Treat-and-extend regimens with anti-VEGF agents in retinal diseases: A literature review and consensus recommendations. Retina (Philadelphia, Pa), 35(8), 1489-1506.

4 - Bhuiyan A, Wong TY, Ting DSW, Govindaiah A, Souied EH, Smith RT. Artificial Intelligence to Stratify Severity of Age-Related Macular Degeneration (AMD) and Predict Risk of Progression to Late AMD. Transl Vis Sci Technol. 2020 Apr 24;9(2):25. doi: 10.1167/tvst.9.2.25. PMID: 32818086; PMCID: PMC7396183.

5 - Li E, Donati S, Lindsley KB, Krzystolik MG, Virgili G. Treatment regimens for administration of anti-vascular endothelial growth factor agents for neovascular age-related macular degeneration. Cochrane Database Syst Rev. 2020 May 5;5(5):CD012208. doi: 10.1002/14651858.CD012208.pub2. PMID: 32374423; PMCID: PMC7202375.

6 - Wiesenfarth, M., Reinke, A., Landman, B.A. et al. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci Rep 11, 2369 (2021). https://doi.org/10.1038/s41598-021-82017-6

7 - Antoine Rivail, Ursula Schmidt-Erfurth, Wolf DieterVogel, Sebastian M. Waldstein, Sophie Riedl, Christoph Grechenig, Zhichao Wu, and Hrvoje Bogunovic, Modeling disease progression in retinal octs with longitudinal self-supervised learning, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11843 LNCS, pp. 44-52, 2019.

8 - Emre, T., Chakravarty, A., Rivail, A., Riedl, S., Schmidt-Erfurth, U., Bogunovic, H. (2022). TINC: Temporally Informed Non-contrastive Learning for Disease Progression Modeling in Retinal OCT Volumes. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13432. Springer, Cham. https://doi.org/10.1007/978-3-031-16434-7 60

9 - Quellec, G., Kowal, J., Hasler, P.W., Scholl, H.P.N., Zweifel, S., Konstantinos, B., de Carvalho, J.E.R., Heeren, T., Egan, C., Tufail, A. and Maloca, P.M. (2019), Feasibility of support vector machine learning in age-related macular degeneration using small sample yielding sparse optical coherence tomography data. Acta Ophthalmol, 97: e719-e728. https://doi.org/10.1111/aos.14055

Further comments

Further comments from the organizers.

The organizers of the MICCAI MARIO Challenge possess extensive expertise in diabetes research, computer vision, and machine learning, particularly in the field of ophthalmology. With over 17 years of experience, the LaTIM laboratory has demonstrated a proven track record in developing cutting-edge machine and deep learning solutions for ophthalmologic applications. The team's experience organizing and participating in prestigious international challenges is also noteworthy. They have successfully organized or co-organized three major ophthalmology challenges: the RIADD Challenge (2021), the IDRID Challenge (2018), and the CATARACTS Challenge (2017). Additionally, they have actively participated in various medical imaging challenges, consistently achieving top rankings. Their notable achievements include winning the first international challenge on DR screening (Retinopathy Online Challenge 2009), reaching second place in the Automatic Detection challenge on Age-related Macular degeneration (ISBI ADAM 2020), and achieving first place in the preliminary round of the MICCAI2023 STAGE Challenge.

Gwenolé Quellec, a member of our team, has served as a challenge reviewer for three editions of the MICCAI conference (2020, 2021, 2022) and was awarded the Best Challenge Reviewer Award in 2020.