TopAneu 2026: Multimodal Vessel-Specific Intracranial Aneurysm Classification and Segmentation Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

TopAneu 2026: Multimodal Vessel-Specific Intracranial Aneurysm Classification and Segmentation Challenge

Challenge acronym

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

TopAneu 2026

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.

Intracranial aneurysms affect an estimated 3% of the global population. Although many remain asymptomatic, rupture can cause subarachnoid hemorrhage with high mortality and long-term disability. Early detection and accurate characterization of intracranial aneurysms are therefore essential. Clinically, the treatment of aneurysms depends on the size, morphology, and location in the brain vessel anatomy. However, there remain difficulties in clinical practice: small aneurysms are often missed, and there are currently no multiclass segmentation methods that classify aneurysms by their specific vessel location. Automated detection and volumetric segmentation from MR angiography (MRA) and CT angiography (CTA) have the potential to reduce workload, standardize measurements, and support more objective shape-based risk assessment.

In recent years, there have been several aneurysm segmentation or detection challenges [1-3] and datasets [5-7]. However, previous datasets and tasks had binary-only aneurysm segmentation labels or with limited aneurysm vessel location labels. More recently, RSNA organized a Kaggle competition focused on classifying aneurysms based on thirteen anatomical locations [2]. While this was an excellent and impactful contribution to the field, we believe there are opportunities for further enhancement. For example, the dataset could encompass a broader range of aneurysm types, including fusiform and dissecting aneurysms in addition to saccular ones. It would also be valuable to include aneurysm segmentation masks and to increase the anatomical granularity of the vessel locations.

To this end, we propose a MICCAI 2026 challenge centered on two interconnected tasks: (1) aneurysm multilabel classification across 50+ fine-grained vessel anatomy locations; (2) aneurysm segmentation with its vessel location

class. For Task 1, algorithms classify an MRA or CTA image based on the presence of aneurysms with a specified vessel location class. For Task 2, algorithms provide voxel-level segmentation masks of detected aneurysms with their corresponding vessel location class. Both tasks inherently evaluate detection accuracy, but prioritize different clinical outputs: Task 1 focuses on high-level diagnostic screening and anatomical localization, while Task 2 emphasizes precise morphological quantification of the aneurysm. Evaluation is conducted per anatomical location, as the vessel location of an aneurysm influences both its characteristics and potential severity. Task 2 will additionally evaluate the performance with stratified aneurysm size groups. Clinically, both the location and size of the aneurysm are key determinants of rupture risk and treatment strategy and should be evaluated in a fine-grained manner.

The challenge will provide a multi-center, multi-modality (CTA and MRA) dataset comprising over 300 new images from organizing institutions, including Lausanne University Hospital (CHUV), University Hospital Geneva (HUG), and Mie Chuo Medical Center; and over 200 cases with new annotations from prior-released public datasets. The internal test sets will comprise private data of over 350 cases. After the result-announcement, additional external private test sets from new centers will be used to evaluate the top methods to rigorously assess robustness and generalizability across scanners, centers, and populations. This setup will enable a comprehensive benchmark for state-of-the-art algorithms.

In summary, this MICCAI 2026 challenge will introduce a new, richly annotated dataset with fine-grained vessel location classes and vessel-specific segmentation masks of aneurysms. It aims to establish a new benchmark for anatomy-aware aneurysm detection, localization, and segmentation. We believe this challenge and the dataset can foster the development of anatomically informed and clinically meaningful image-analysis methods for improved diagnosis, guidance, and treatment decision support in aneurysms and other neurovascular diseases.

Challenge keywords

List the primary keywords that characterize the challenge.

Neurovascular image analysis, image-guided vascular interventions, deep learning, vascular topology, image processing, intracranial aneurysms, detection, classification, segmentation

Year

2026

Novelty of the challenge

Briefly describe the novelty of the challenge.

This challenge introduces a comprehensive, clinically oriented pipeline for intracranial aneurysm analysis, combining detection, fine-grained localization, and voxel-level segmentation across CT and MR angiography modalities. Unlike existing datasets, our challenge includes over 50 predefined vessel-specific classes, enabling precise anatomical localization, and stratifies aneurysms by size, reflecting clinically meaningful risk factors. The use of both gold-standard (from human-expert) and silver-standard (model-generated [8]) anatomical vessel annotations allows development of models that are robust to real-world label variability.

To our knowledge, no publicly available challenge simultaneously addresses multi-modality aneurysm detection, vessel-location-aware segmentation, and size-based evaluation in a single unified framework. This design

encourages development of methods that not only identify aneurysms but also provide clinically actionable morphological information, bridging the gap between automated image analysis and intervention planning.

Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

Automated intracranial aneurysm analysis has direct clinical applications in screening, diagnosis, and treatment planning. Task 1 supports vessel-class-encoded detection and localization of aneurysm candidates in CT or MR angiography scans, enabling radiologists to prioritize review and quickly identify high-risk regions. Task 2 provides voxel-level, vessel-specific aneurysm segmentation with size information, which is critical for assessing rupture risk, planning surgical or endovascular intervention, and monitoring aneurysm growth over time. Together, the two tasks reflect a clinically realistic workflow, from initial detection to anatomically precise and actionable characterization of aneurysms.

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

We have coordinated with the organizers of three other related challenges/workshops: ISLES challenge 2026, TopBrain 2026 challenge, and SWITCH+ 2026 workshop, and we are happy to merge into a joint, comprehensive neurovascular-themed event.

Duration

How long does the challenge take?

2 Hours

In case you selected half or full day, please explain why you need a long slot for your challenge.

N/A

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 expect at least 20 final submissions for TopAneu. This estimation is based on recent similar neurovascular MICCAI challenges in which the organizers were involved, as well as a recently concluded RSNA Kaggle aneurysm competition, which attracted over 1000 teams. With the new features and improvements introduced in the new challenge, and with the outreach from partnering neurovascular challenges, we anticipate considerable interest and participation from the community.

Publication and future plans

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

The challenge results will be summarized as a single benchmark paper for publication in a suitable journal.

MICCAI LNCS proceedings

Indicate if you want to offer MICCAI Springer LNCS proceedings to the participants. Publishing a proceedings volume is optional and at the discretion of each challenge's organizers. At a minimum, organizers must ensure that a description of each participant's submission is publicly available. Organizers who wish to publish MICCAI Springer LNCS proceedings must adhere to the MICCAI Satellite events publication process.

No

Collaboration with European Society of Radiology (ESR)

In collaboration with European Society of Radiology (ESR), we announce special clinical interest topics with associated clinicians who can help with the preparation of the proposals; the best 3 challenge proposals on these topics will get the opportunity to present their challenges at the European Congress of Radiology (ECR) 2027 in a special session. If you want to organize a challenge in collaboration with ESR on one of these topics, please reach out to the MICCAI Challenges Team (miccai-challenges-2026@dkfz-heidelberg.de) and we will put you in contact with the corresponding clinician.

Challenge in collaboration with ESR. Ticking 'Yes' implies that the challenge has been prepared in collaboration with the clinical contact point.

No

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.

For the in-person event, we need projectors, microphones, loudspeakers, and a camera/video conferencing system for hybrid participation.

The challenge will be hosted on grand-challenge.org. Participants use their own computing resources for the algorithm training and development. The organizers will use grand-challenge.org and local servers for the Docker evaluation during the testing phase.

TASK 1: Task-1: Vessel-Specific Aneurysm Classification

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.

Task 1 aims to detect the presence or absence of intracranial aneurysm at more than 50 predefined vessel location classes in CT or MR angiography images. This task is formulated as a multi-label classification problem, where each image is associated with a set of one-hot encoded binary labels indicating the presence or absence of aneurysms at each anatomically defined vessel location. Algorithm performance is evaluated using per-location precision, recall, and Matthews correlation coefficient (MCC). The vessel locations are based on the vessel anatomy classes from the TopBrain dataset [4]. As auxiliary training labels, each training image also comprises either verified (gold-standard, 50 cases) or model-predicted (silver-standard) TopBrain vessel segmentation masks [8].

Task 1 emphasizes vessel-specific evaluation of intracranial aneurysms, an aspect that is rarely addressed in existing benchmarks. By decoupling aneurysm detection and assessment according to more than 50 predefined vessel locations, the challenge aligns algorithm evaluation with clinically meaningful distinctions in aneurysm diagnosis and treatment planning. By benchmarking multi-label classification at fine-grained vessel locations, the challenge promotes the development of anatomy-aware models that handle heterogeneous imaging data and complex vascular anatomy. The inclusion of both gold- and silver-standard annotations further encourages robust learning under label uncertainty and weak supervision.

While Task 1 is formulated as a multi-label classification problem, the solution space is not restricted to classification-only algorithms. Participants are free to adopt diverse methodological paradigms. This design choice is motivated by recent evidence from the closely related RSNA Kaggle intracranial aneurysm detection challenge [2], which—despite being posed as a multi-label classification task—saw top-ranked solutions emerge from a wide range of modeling strategies, including hybrid segmentation–classification and multi-task learning methods.

In contrast to dense voxel-level or pixel-level segmentation annotations, which are costly and time-consuming to acquire, multi-label classification annotations can be generated more efficiently while still capturing clinically meaningful information. This enables the inclusion of a larger and more diverse data cohort, which is critical for the quality of our benchmark.

Although Task 1 only requires identifying the presence or absence of aneurysms rather than producing explicit spatial coordinates, the predefined vessel locations are highly fine-grained, encompassing more than 50 anatomically distinct vessel segments. This level of anatomical specificity provides a meaningful form of localization and is clinically valuable for screening and diagnostic workflows, where detection of the aneurysm presence at the affected vessel class is often sufficient to guide further evaluation and decision-making. As such, Task 1 strikes a deliberate balance between annotation scalability, methodological flexibility, and clinical relevance and utility.

Keywords

List the primary keywords that characterize the task.

Multilabel classification, Intracranial aneurysm, Detection, Anatomical location

ORGANIZATION

Organizers

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

[Eindhoven University of Technology (TU/e), The Netherlands]
Ruisheng Su

[UMC Utrecht, the Netherlands]
Irene van der Schaaf, Jiaxin Zhang, Chantal M.W. Tax

[Department of Radiology, Lausanne University Hospital, Switzerland]
Jonas Richiardi, Guillaume Saliou, Patric Hagmann

[Zurich University of Applied Sciences (ZHAW), Switzerland]
Norman Juchler, Ekaterina Golubeva, L'Allinec Vincent, Sven Hirsch

[Geneva University Hospitals, Switzerland]
Philippe Bijlenga

[University of Zurich, Switzerland]
Ezequiel de la Rosa, Kaiyuan Yang, Houjing Huang, Suprosanna Shit, Bjoern Menze

[Medical Image Insights, China]
Pengcheng Shi

[Zhongnan Hospital of Wuhan University, China]
Yihui Ma

[Charité Lab for Artificial Intelligence in Medicine (CLAIM), Germany]
Orhun Utku Aydin, Satoru Tanioka, Dietmar Frey

[Department of Radiology, Ankara University, Ankara, Turkey]
Oktay Algin

[German Cancer Research Center (DKFZ) Heidelberg, Germany]
Stefan Denner, Andres Martinez Mora, Alexandra Ertl, Maximilian R. Rokuss, Yannick Kirchhoff, Klaus Maier-Hein

b) Provide information on the primary contact person.

Ruisheng Su, r.su@tue.nl, Eindhoven University of Technology, The Netherlands.

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes. Clinicians are integral members of the organizing team and are responsible for:

Definition of Annotation Targets: Identifying clinically significant anatomical and pathological targets and defining the annotation protocol.

Quality Assurance & Verification: Manually reviewing and validating annotations to ensure high-fidelity ground truth.

Domain Expertise: Providing the clinical perspective and context to bridge the gap between algorithmic performance and clinical utility.

Clinical Impact Design: Designing downstream analysis and key objectives to assess how the proposed solutions could be integrated into diagnostic or surgical workflows.

Data Governance & Contribution: As data owners from various medical institutions, they oversee the curation, de-identification, and ethical sharing of the multi-center datasets.

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 conference submission deadline

Challenge venue and platform

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

MICCAI 2026

b) Report the platform used to run the challenge.

grand-challenge.org

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

No URL available online yet. Tentative: www.topaneu-26.grand-challenge.org

Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

No user interaction is allowed at any step.

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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

Participants are allowed to use any other public datasets and private in-house data, or modify the supplied training data, provided that they disclose any additional or modified training datasets in their description of the submitted algorithm.

For teams that choose to train models with additional data, we require them to provide models trained exclusively on the challenge dataset for post-challenge analysis.

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

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

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

Top three independent teams will be publicly named, and given a certificate and a small souvenir at the in-person challenge event. There will be no monetary awards given.

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.

Top performing submissions are announced at the in-person challenge event. However, the participating team can choose whether their results will be made public any time before the day of announcement. The top five teams will be invited to prepare a 5-minute presentation for the challenge session to present and discuss their methods.

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).

Top five teams from each task are invited to contribute to our challenge publication. Each team can have maximum three co-authors for the challenge paper. Additional authors from the submissions can be included upon request with justification according to the ICMJE authorship guidelines. Participating teams may publish their own results separately without any publication embargo.

Submission method

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

Examples:

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

Submission for evaluation will be done on the test datasets via submitted docker containers, i.e., type 2 submissions on Grand Challenge.

Along with the docker containers, each participating team is encouraged to submit a 1-page summary describing their methods and approaches. This summary is required for co-authorship in the final challenge journal paper.

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.

Sanity-check phase: Consists of a 'toy' example docker submission phase. It is solely intended for teams to test whether their devised dockers work in the Grand Challenge cloud environment. Multiple submissions to this phase are allowed.

The test phases only allow one submission per team. Each team is given only one opportunity to upload their containers for the test set. In case of technical issues, we allow the participants to try their docker submissions again.

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

Preliminary Schedule:
- Challenge website online: April 15, 2026
- Release of training data: June 15, 2026
- Submit to test phases: August 15 – September 05, 2026
- Contacting top performing teams and planning for the in-person session: Sep 05 – Oct 01, 2026

- In-person challenge event: Oct 04 or Oct 08, 2026

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 data are derived from studies that were approved by their local ethics committee. The data is anonymized (removal and anonymization of relevant patient information) in accordance with the IRB regulations. This includes de-facing and cropping procedures to ensure patient privacy in the image data, and the data do not contain any personal identifiers.

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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

Open use. Must provide the source. Use for commercial purposes requires permission of the data owner. (see https://opendata.swiss/en/terms-of-use)

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.

Our challenge evaluation will be transparent to all participants. We will use a public GitHub repo to update and synchronize any changes to the evaluation code.

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

The submitted docker containers will be made publicly available with permissions from the participating teams. We highly encourage the participants to make their code public.

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.

We will not give monetary awards.
Only the main organizers and their local annotation team will have access to all test labels and the private test datasets.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research,Diagnosis,Education,Screening,Intervention planning

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

Detection,Classification

Cohorts

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

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

Any patient undergoing a routine MRA or CTA scan, as an accurate intracranial aneurysm detection can be beneficial for incidental finding and screening. Intracranial aneurysm patients in routine clinical practice with a standard-of-care acute clinical imaging protocol (CTA, MRA).

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

Intracranial aneurysm patients in routine clinical practice with a standard-of-care acute clinical imaging protocol (CTA, MRA).

Imaging modality(ies)

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

CT Angiography, MR Angiography

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.

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

None.

Target entity(ies)

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

Brain shown in angiographic CT and MR scans.

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

The primary targets for the participating algorithms are intracranial aneurysms, categorized by their specific vessel anatomical locations. The vessel location classes are based on the vessel anatomy from the TopBrain dataset [4] that covers the entire cerebrovascular tree. The vessel segments are distinguished between the left and right sides, resulting in over 50 unique location-specific classes.

The vessel locations include:

VA trunk (left and right)
VA-PICA junction (left and right)
PICA trunk (left and right)
VA-BA junction
BA-AICA junction (left and right)
AICA trunk (left and right)
BA trunk
BA-SCA junction (left and right)
SCA trunk (left and right)
BA tip
P1P2 (left and right)
P3P4 (left and right)
ICA-infraclinoid C1-C5 (left and right)
ICA-C6-OA junction (left and right)
ICA-C6-nonOA (left and right)
ICA-C7-Pcom junction (left and right)
ICA-C7-nonPcom (left and right)
ICA-C7-AChA junction (left and right)
ICA-terminus (left and right)
A1 (left and right)
Acom
A2 (left and right)
A2/A3-CMA junction (left and right)
A3 (left and right)
Distal ACA branches (left and right)
M1 (left and right)
M1-M2 junction (left and right)
M2-M3 and distal (left and right)

Abbreviations: Vertebral artery (VA); Posterior inferior cerebellar artery (PICA); Anterior inferior cerebellar artery (AICA); Basilar artery (BA); Superior cerebellar artery (SCA); Segments of posterior cerebral artery (P1-P4); Internal carotid artery (ICA); segments of ICA (C1-C7); Ophthalmic artery (OA); Posterior communicating artery (Pcom); Anterior choroidal artery (AChA); Anterior cerebral artery (ACA); segments of ACA (A1-A3); Anterior communicating artery (Acom); Callosomarginal artery (CMA); Segments of middle cerebral artery (M1-M3).

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)).

Accuracy,Precision,Robustness,Sensitivity,Specificity

DATA SETS

Data source(s)

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

Data were acquired with CT and MRI systems from various manufacturers and scanner types during routine clinical practice. For MRI, this includes Siemens 1.5T and 3T systems.

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

All images are obtained with acquisition parameters spanning within normal ranges used for clinical purposes. These may differ between institutions providing datasets, but for example, magnetic resonance angiography is typically acquired with time-of-flight angiography.

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.

Center 1: Lausanne University Hospital (CHUV), Switzerland
Center 2: Geneva University Hospital (HUG), Switzerland
Center 3: University Medical Center Utrecht (UMCU), The Netherlands
Center 4: Mie Chuo Medical Center (Mie-Chuo), Japan
Samples from prior public dataset candidates: INSTED 2024 challenge [3], Lausanne TOF-MRA aneurysm dataset [6], and Royal Brisbane TOF-MRA aneurysm dataset [7].

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).

Healthcare professionals operating CT and MRI systems in clinical routine.

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 in Task 1 is a 3D angiographic imaging scan of a human brain in either CTA or MRA modality. Both training and test cases are fully annotated with the list of labels indicating the presence or absence of aneurysms at predefined vessel locations by anatomy. To support model development, all training cases are supplemented with separate vessel segmentation masks, provided in two tiers: Gold-Standard masks, which have been verified by human experts, and Silver-Standard masks, which are generated by a high-performance prediction model. Furthermore, while Task 1 focuses on classification, the majority of the training cases include high-quality aneurysm segmentation masks—shared with Task 2—to allow participants to leverage precise segmentation masks for improved detection and localization accuracy.

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

Task 1 data is a superset of Task 2 data.

Train set: 500 cases
By centers and modalities:
Center 1 (CHUV): 100 (MRA); Center 2 (HUG): 100 (CTA/MRA); Center 4 (Mie-Chuo): 100 (CTA); Re-labeled public data: 200 (CTA/MRA).
All annotated with one-hot encoded binary labels indicating the presence or absence of aneurysms at each anatomically defined vessel location, with auxiliary vessel TopBrain segmentation labels, either verified (gold-standard, 50 cases, 25 for CTA + 25 for MRA) or model-predicted (silver-standard).

Test set: 350 cases
By centers and modalities:
Center 1 (CHUV): 50 (MRA); Center 2 (HUG): 100 (CTA/MRA); Center 3 (UMCU): 100 (CTA/MRA); Center 4 (Mie-Chuo): 100 (CTA).
All annotated with one-hot encoded binary labels indicating the presence or absence of aneurysms at each anatomically defined vessel location, with auxiliary vessel TopBrain segmentation labels, either verified (gold-standard, 20 cases, 10 for CTA + 10 for MRA) or model-predicted (silver-standard).

c) How much of the data are already annotated (stratified by train test in percentage)?

Train set: ~400 (80%) annotated with aneurysm detection in the form of masks or metadata text, to be labeled with vessel location class.
Test set: ~250 (71%) annotated as in the train set.

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

The database size is conveyed by taking into account:
-The purpose of the challenge (detection, classification, and segmentation).
-The effort needed to identify and retrieve the imaging data of patients from the centers.

-The effort needed to manually annotate the ground-truth at a voxel-wise level.

We have maximized the amount of data we could retrieve from each of the participating centers. From our experience in related challenges, we believe that ~850 scans (500 for training + 350 for testing) are a reasonable amount of data to develop and evaluate algorithms for the task at hand.

The data is split in an approximate train-test proportion of 60-40 split. Both the train and test sets include a wide range of intracranial aneurysms (with various sizes and locations). It is worth mentioning that we do not provide a validation set. However, we encourage participants to consider a subset of the training set to validate their algorithms in, e.g., a cross-validation fashion.

e) 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 dataset will be specifically curated to ensure sufficient representation of aneurysms on the included anatomical vessel locations. We will implement a stratified sampling approach to guarantee that each of the anatomical vessel locations (including both left and right) contains at least three positive samples in both training and test sets.

The dataset will also be curated especially to ensure at least one-fifth of the data is non-aneurysm control cases in both training and test sets.

Justification for this choice: By ensuring a minimum number of occurrences per location, we provide a baseline for cross-validation and ensure that the winning models demonstrate generalized diagnostic capability across the entire vascular tree, rather than only performing well on the most common vessel sites. Similarly, by providing non-aneurysm controls, we can evaluate for specificity on a patient-level.

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

Both the training and test sets from all participating centers are new. All vessel-specific multi-class labels used in this challenge are newly created. In the training set, we include 200 publicly available cases, for which we created entirely new multi-class vessel annotations.

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.

Aneurysms were initially identified within the 3D volumes based on either prior clinical diagnoses or a de novo screening by a member of our clinical experts. Then for detected aneurysms, the anatomical vessel location name is labeled by our clinical experts. At least one clinical expert reviews and verifies each detected aneurysm and its vessel location name.

The auxiliary vessel segmentation masks for the training data will be provided by the partnering TopBrain

challenge, either verified by human experts (gold-standard) or generated (silver-standard) using an organizer-developed strong baseline model [8].

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.

All annotators underwent a standardized training and calibration phase for the aneurysm location scheme prior to the formal labeling of the dataset. This phase included a comprehensive review of our specified vessel anatomical locations, which provides the exact definitions of aneurysm locations for over 50 anatomical vessel segments and their junctions. Annotators are instructed to label aneurysms based on the parent vessel of origin.

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 annotations are verified and approved by senior (over 10 years of experience) neurosurgeons or neuroradiologists from the organizing team.

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.

None, since each label is generated by a single clinician verifier.

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 data were anonymized (removal and anonymization of relevant DICOM patient information). Additional de-facing and cropping procedures were performed to ensure patient privacy in the image data after converting the DICOM to NIfTI format. The defaced and cropped image includes only the brain region. All images are re-oriented to LPS+ orientation. Other than the defacing, the cropping to brain region, and re-orientation steps, no further preprocessing of the data is performed to keep the data as close to the clinical setting as possible.

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.

- In case the aneurysm is located close to a bifurcation, or when an aneurysm has multiple touching adjacent vessels, the annotator may have high variability in deciding which vessel location class to assign to the aneurysm.
- For CTA modality, aneurysms along the internal carotid artery can be hard to detect, classify, and segment due to their proximity to the bones.
- The left and right sides of the anterior cerebral artery along the inter-hemispheric fissure can be hard to distinguish, especially for intertwined vessels. This can cause uncertainty in the aneurysm parent vessel location label.
- Images from different centers are not labeled by the same annotator, and different annotation styles can introduce variability.

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

CTA and MRA can have different resolutions. MRA might amplify the effect of stenosis in vessels. Both modalities might contain common artifacts such as flow-dependent signal cancellation artifacts, noise artifacts, ringing artifacts, pulsation artifacts, and beam hardening artifacts. These artifacts might cause the vessels and aneurysms to be over- or under-segmented, and might affect the diagnosis.

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)

Metrics for the multi-label binary classification task:

Per-location precision
Per-location recall (sensitivity)
Per-location Matthews correlation coefficient (MCC)

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

Precision, recall, and Matthews correlation coefficient (MCC) are metrics that can handle class imbalance of aneurysm prevalence in certain vessel locations. High precision ensures a low false alarm rate. High recall ensures a low miss rate. MCC takes into account the true negative performance. All metrics are highly clinically relevant and commonly reported in the literature. Furthermore, we report the metrics per-location since the focus of our challenge is to evaluate algorithms for each fine-grained anatomical location.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

The ranking for the awards of the MICCAI event will be based on the leaderboards displayed on the grand-challenge website, which rank the teams by sorting their performance on each metric. The leaderboard uses equal weights for each metric and ranks the submissions based on the mean rank position across all metrics. In case of identical mean ranks, per-location MCC serves as a tie-breaking metric due to its balanced handling of class imbalance.

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

Missing results will result in an increase in false negative detections.

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

The decision to utilize a mean rank across all metrics is driven by the multi-dimensional and cohort-based nature of the evaluation. This challenge evaluates performance across over 50 distinct anatomical locations, and the detection metrics are computed per-location class across the cohort. An average rank of per-location precision,

recall, and Matthews correlation coefficient (MCC) rankings helps identify the most clinically useful and robust algorithm.

Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

Post-challenge analysis will use bootstraps of the test set and analyze the ranking stability for bootstrapped average ranks.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

During challenge evaluation, bootstrap resampling with 1,000 iterations will be used to estimate 95 percent confidence intervals for the primary evaluation metrics on the full test set. For each algorithm and metric, we will report the mean performance together with its bootstrap-derived confidence interval and dispersion measure. This ensures that leaderboard results are accompanied by uncertainty estimates and do not rely solely on point estimates.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

Variability of algorithm performance across test cases will be addressed at two stages.
During the challenge evaluation, dispersion across cases will be quantified by reporting the mean and standard deviation of metrics to ensure that leaderboard scores reflect not only average performance but also stability across the test set.

For post-challenge analysis, a more detailed variability analysis will be conducted. This will include examination of performance heterogeneity across clinically meaningful subgroups, such as aneurysm location category, by computing subgroup-specific means with 95 percent confidence intervals. Where appropriate, distribution-based visualizations will be used to identify potential outliers, which will be qualitatively reviewed to assess whether reduced performance is associated with anatomical complexity, small lesions, or annotation uncertainty.

This staged approach ensures a robust and fair ranking during the challenge while allowing for deeper interpretability analyses after completion.

Provide a description of how variability of rankings is assessed.

Variability of rankings will be assessed through bootstrap resampling of the test set using 1,000 iterations. For each bootstrap sample, evaluation metrics will be recomputed on the resampled cases using the submitted predictions, and algorithms will be ranked according to the predefined ranking scheme. This process yields a distribution of ranks for each algorithm across resampled test sets. Ranking stability will be summarized using the median rank and the 95 percent rank interval, thereby quantifying the robustness of the relative ordering of algorithms to sampling variability in the test set.

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Post challenge, pairwise comparisons between the top five performing algorithms will be conducted using statistical tests appropriate to the metric type. For cohort-level detection metrics derived from binary outcomes such as precision, recall, and MCC, McNemar's test will be used to compare paired classification disagreements between algorithms using contingency tables.

Provide a description of the missing data handling.

Not applicable since our data are images with verified labels.

Indicate any software product that is used for all data analysis methods.

Scipy and Numpy with custom Python scripts.

Further analyses

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

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

We can perform ensembling of podium-finish top algorithms, and evaluate the ensemble prediction in the benchmark.

Post challenge, we will evaluate the top-performing models on an external private test set. This second phase is designed to evaluate the generalization capability of the algorithms. The results from this private evaluation represent a meaningful measure of a model's potential performance in a real-world clinical environment.

Post challenge, we will perform inter-rater agreement analysis on the annotations.

TASK 2: Task-2: Vessel-Specific Aneurysm Segmentation

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.

Task 2 focuses on voxel-level multiclass segmentation of intracranial aneurysms in CT and MR angiography images. Given an angiographic volume, participants are required to identify the presence of aneurysms and generate a voxel-wise aneurysm segmentation mask, where each aneurysm voxel is assigned to one of more than 50 predefined intracranial vessel location classes from Task 1. Task 2 is formulated as a multiclass semantic segmentation problem with vessel-specific aneurysm classes. The train and test images for Task 2 are a subset of the ones in Task 1, with Task 2 providing additional verified aneurysm segmentation annotations.

Algorithm performance is evaluated for both detection and segmentation. Detection performance is assessed using per-location precision, recall, and Matthews correlation coefficient (MCC), the same as for Task 1. Aneurysms are further stratified into three size groups based on diameter, and detection performance is additionally analyzed across size categories. For true-positive detections, segmentation performance is evaluated using overlap-based, surface-distance–based, and volume-difference metrics, reported per vessel location and per size group.

Clinically, both the location and size of the aneurysm are key determinants of rupture risk and treatment strategy. Aneurysms in certain vessels carry a higher rupture risk and may be more challenging to treat. Similarly, larger aneurysms require closer monitoring or intervention. Task 2 emphasizes fine-grained, voxel-level segmentation with vessel-specific labels and size stratification, providing clinically meaningful information that can support risk assessment, treatment planning, and longitudinal follow-up. Task 2 complements Task 1 by moving from detection to clinically actionable segmentation. While Task 1 identifies the presence and anatomical location of aneurysms, Task 2 provides voxel-level segmentations with vessel-specific labels and size information. This fine-grained information is essential for assessing rupture risk, planning surgical or endovascular intervention, and monitoring aneurysm growth over time. In other words, Task 1 answers “which vessel location has an aneurysm?”, whereas Task 2 answers more questions: “how large is it, what is its precise shape?”. Together, the two tasks form a clinically meaningful pipeline from detection to anatomically precise characterization.

Keywords

List the primary keywords that characterize the task.

Multiclass segmentation, Intracranial aneurysm, Detection, Anatomical location, Aneurysm size

ORGANIZATION

Organizers

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

[Eindhoven University of Technology (TU/e), The Netherlands]
Ruisheng Su

[UMC Utrecht, the Netherlands]
Irene van der Schaaf, Jiaxin Zhang, Chantal M.W. Tax

[Department of Radiology, Lausanne University Hospital, Switzerland]
Jonas Richiardi, Guillaume Saliou, Patric Hagmann

[Zurich University of Applied Sciences (ZHAW), Switzerland]
Norman Juchler, Ekaterina Golubeva, L'Allinec Vincent, Sven Hirsch

[Geneva University Hospitals, Switzerland]
Philippe Bijlenga

[University of Zurich, Switzerland]
Ezequiel de la Rosa, Kaiyuan Yang, Houjing Huang, Suprosanna Shit, Bjoern Menze

[Medical Image Insights, China]
Pengcheng Shi

[Zhongnan Hospital of Wuhan University, China]
Yihui Ma

[Charité Lab for Artificial Intelligence in Medicine (CLAIM), Germany]
Orhun Utku Aydin, Satoru Tanioka, Dietmar Frey

[Department of Radiology, Ankara University, Ankara, Turkey]
Oktay Algin

[German Cancer Research Center (DKFZ) Heidelberg, Germany]
Stefan Denner, Andres Martinez Mora, Alexandra Ertl, Maximilian R. Rokuss, Yannick Kirchhoff, Klaus Maier-Hein

b) Provide information on the primary contact person.

Ruisheng Su, r.su@tue.nl, Eindhoven University of Technology, The Netherlands.

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes. Clinicians are integral members of the organizing team and are responsible for:

Definition of Annotation Targets: Identifying clinically significant anatomical and pathological targets and defining the annotation protocol.

Quality Assurance & Verification: Manually reviewing and validating annotations to ensure high-fidelity ground truth.

Domain Expertise: Providing the clinical perspective and context to bridge the gap between algorithmic performance and clinical utility.

Clinical Impact Design: Designing downstream analysis and key objectives to assess how the proposed solutions could be integrated into diagnostic or surgical workflows.

Data Governance & Contribution: As data owners from various medical institutions, they oversee the curation, de-identification, and ethical sharing of the multi-center datasets.

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 conference submission deadline

Challenge venue and platform

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

MICCAI 2026

b) Report the platform used to run the challenge.

grand-challenge.org

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

No URL available online yet. Tentative: www.topaneu-26.grand-challenge.org

Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

No user interaction is allowed at any step.

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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

Participants are allowed to use any other public datasets and private in-house data, or modify the supplied training data, provided that they disclose any additional or modified training datasets in their description of the submitted algorithm.

For teams that choose to train models with additional data, we require them to provide models trained exclusively on the challenge dataset for post-challenge analysis.

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

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

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

Top three independent teams will be publicly named, and given a certificate and a small souvenir at the in-person challenge event. There will be no monetary awards given.

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.

Top performing submissions are announced at the in-person challenge event. However, the participating team can choose whether their results will be made public any time before the day of announcement. The top five teams will be invited to prepare a 5-minute presentation for the challenge session to present and discuss their methods.

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).

Top five teams from each task are invited to contribute to our challenge publication. Each team can have maximum three co-authors for the challenge paper. Additional authors from the submissions can be included upon request with justification according to the ICMJE authorship guidelines. Participating teams may publish their own results separately without any publication embargo.

Submission method

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

Examples:

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

Submission for evaluation will be done on the test datasets via submitted docker containers, i.e., type 2 submissions on Grand Challenge.

Along with the docker containers, each participating team is encouraged to submit a 1-page summary describing their methods and approaches. This summary is required for co-authorship in the final challenge journal paper.

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.

Sanity-check phase: Consists of a 'toy' example docker submission phase. It is solely intended for teams to test whether their devised dockers work in the Grand Challenge cloud environment. Multiple submissions to this phase are allowed.

The test phases only allow one submission per team. Each team is given only one opportunity to upload their containers for the test set. In case of technical issues, we allow the participants to try their docker submissions again.

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

Preliminary Schedule:
- Challenge website online: April 15, 2026
- Release of training data: June 15, 2026
- Submit to test phases: August 15 – September 05, 2026
- Contacting top performing teams and planning for the in-person session: Sep 05 – Oct 01, 2026
- In-person challenge event: Oct 04 or Oct 08, 2026

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 data are derived from studies that were approved by their local ethics committee. The data is anonymized (removal and anonymization of relevant patient information) in accordance with the IRB regulations. This includes de-facing and cropping procedures to ensure patient privacy in the image data, and the data do not contain any personal identifiers.

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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

Open use. Must provide the source. Use for commercial purposes requires permission of the data owner. (see https://opendata.swiss/en/terms-of-use)

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.

Our challenge evaluation will be transparent to all participants. We will use a public GitHub repo to update and synchronize any changes to the evaluation code.

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

The submitted docker containers will be made publicly available with permissions from the participating teams. We highly encourage the participants to make their code public.

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.

We will not give monetary awards.
Only the main organizers and their local annotation team will have access to all test labels and the private test datasets.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research,Diagnosis,Education,Screening,Intervention planning

Task category(ies)

State the task category(ies)

Examples:

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

Detection,Segmentation

Cohorts

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

defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

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

Any patient undergoing a routine MRA or CTA scan, as an accurate intracranial aneurysm detection can be beneficial for incidental finding and screening. Intracranial aneurysm patients in routine clinical practice with a standard-of-care acute clinical imaging protocol (CTA, MRA).

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

Intracranial aneurysm patients in routine clinical practice with a standard-of-care acute clinical imaging protocol (CTA, MRA).

Imaging modality(ies)

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

CT Angiography, MR Angiography

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.

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

None.

Target entity(ies)

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

Brain shown in angiographic CT and MR scans.

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

The primary targets for the participating algorithms are intracranial aneurysms, categorized by their specific vessel anatomical locations. The vessel location classes are based on the vessel anatomy from the TopBrain dataset [4] that covers the entire cerebrovascular tree. The vessel segments are distinguished between the left and right sides, resulting in over 50 unique location-specific classes.

The vessel locations include:

VA trunk (left and right)
VA-PICA junction (left and right)

PICA trunk (left and right)
VA-BA junction
BA-AICA junction (left and right)
AICA trunk (left and right)
BA trunk
BA-SCA junction (left and right)
SCA trunk (left and right)
BA tip
P1P2 (left and right)
P3P4 (left and right)
ICA-infraclinoid C1-C5 (left and right)
ICA-C6-OA junction (left and right)
ICA-C6-nonOA (left and right)
ICA-C7-Pcom junction (left and right)
ICA-C7-nonPcom (left and right)
ICA-C7-AChA junction (left and right)
ICA-terminus (left and right)
A1 (left and right)
Acom
A2 (left and right)
A2/A3-CMA junction (left and right)
A3 (left and right)
Distal ACA branches (left and right)
M1 (left and right)
M1-M2 junction (left and right)
M2-M3 and distal (left and right)

Abbreviations: Vertebral artery (VA); Posterior inferior cerebellar artery (PICA); Anterior inferior cerebellar artery (AICA); Basilar artery (BA); Superior cerebellar artery (SCA); Segments of posterior cerebral artery (P1-P4); Internal carotid artery (ICA); segments of ICA (C1-C7); Ophthalmic artery (OA); Posterior communicating artery (Pcom); Anterior choroidal artery (AChA); Anterior cerebral artery (ACA); segments of ACA (A1-A3); Anterior communicating artery (Acom); Callosomarginal artery (CMA); Segments of middle cerebral artery (M1-M3).

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)).

Accuracy,Precision,Robustness,Sensitivity,Specificity

DATA SETS

Data source(s)

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

Data were acquired with CT and MRI systems from various manufacturers and scanner types during routine clinical practice. For MRI, this includes Siemens 1.5T and 3T systems.

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

All images are obtained with acquisition parameters spanning within normal ranges used for clinical purposes. These may differ between institutions providing datasets, but for example, magnetic resonance angiography is typically acquired with time-of-flight angiography.

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.

Center 1: Lausanne University Hospital (CHUV), Switzerland
Center 2: Geneva University Hospital (HUG), Switzerland
Center 3: University Medical Center Utrecht (UMCU), The Netherlands
Samples from prior public dataset candidates: INSTED 2024 challenge [3], Lausanne TOF-MRA aneurysm dataset [6], and Royal Brisbane TOF-MRA aneurysm dataset [7].

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).

Healthcare professionals operating CT and MRI systems in clinical routine.

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 in Task 2 is a 3D angiographic imaging scan of a human brain, acquired via either CTA or MRA modalities. Both training and test cases are fully annotated with aneurysm segmentation masks, where the voxel-level labels

correspond to specific anatomical vessel locations defined in Task 1. To support model development, all training cases are supplemented with separate vessel segmentation masks, provided in two tiers: Gold-Standard masks, which have been verified by human experts, and Silver-Standard masks, which are generated by a high-performance prediction model. While these two mask types are provided as separate files, participants should treat the aneurysm pathology as the primary objective; in case of spatial overlap between a vessel and an aneurysm, the aneurysm label should take priority.

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

Task 2 data is a subset of Task 1 data.

Train set: 300 cases
By centers and modalities:
Center 1 (CHUV): 100 (MRA); Center 2 (HUG): 100 (CTA/MRA); Re-labeled public data: 100 (CTA/MRA).
All annotated with voxel-level multi-class aneurysm segmentation masks, with auxiliary vessel TopBrain segmentation labels, either verified (gold-standard, 50 cases, 25 for CTA + 25 for MRA) or model-predicted (silver-standard).

Test set: 250 cases
By centers and modalities:
Center 1 (CHUV): 50 (MRA); Center 2 (HUG): 100 (CTA/MRA); Center 3 (UMCU): 100 (CTA/MRA).
All annotated with voxel-level multi-class aneurysm segmentation masks, with auxiliary vessel TopBrain segmentation labels, either verified (gold-standard, 20 cases, 10 for CTA + 10 for MRA) or model-predicted (silver-standard).

c) How much of the data are already annotated (stratified by train test in percentage)?

Train set: ~250 (83%)
Test set ~200 (80%)

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

The database size is conveyed by taking into account:
-The purpose of the challenge (detection, classification, and segmentation).
-The effort needed to identify and retrieve the imaging data of patients from the centers.
-The effort needed to manually annotate the ground-truth at a voxel-wise level.

We have maximized the amount of data we could retrieve from each of the participating centers. From our experience in related challenges, we believe that ~550 scans (300 for training + 250 for testing) are a reasonable amount of data to develop and evaluate algorithms for the task at hand.

The data is split in an approximate train-test proportion of 60-40 split. Both the train and test sets include a wide range of intracranial aneurysms (with various sizes and locations). It is worth mentioning that we do not provide a validation set. However, we encourage participants to consider a subset of the training set to validate their algorithms in, e.g., a cross-validation fashion.

e) 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 dataset will be specifically curated to ensure sufficient representation of aneurysms on the included anatomical vessel locations. We will implement a stratified sampling approach to guarantee that each of the anatomical vessel locations (including both left and right) contains at least three positive samples in both training and test sets.

The dataset will also be curated especially to ensure at least one-fifth of the data is non-aneurysm control cases in both training and test sets.

Justification for this choice: By ensuring a minimum number of occurrences per location, we provide a baseline for cross-validation and ensure that the winning models demonstrate generalized diagnostic capability across the entire vascular tree, rather than only performing well on the most common vessel sites. Similarly, by providing non-aneurysm controls, we can evaluate for specificity on a patient-level.

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

Both the training and test sets from all participating centers are new. All vessel-specific multi-class labels used in this challenge are newly created. In the training set, we include 100 publicly available cases, for which we created entirely new multi-class vessel annotations.

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.

Aneurysms were initially identified within the 3D volumes based on either prior clinical diagnoses or a de novo screening by a member of our clinical experts. Then for detected aneurysms, the aneurysm will be segmented by either manual annotation or correction based on a state-of-the-art binary aneurysm segmentation model prediction [8]. The class of an aneurysm will be assigned to its anatomical vessel location name from Task 1 annotations. At least one clinical expert will review and verify the segmentation masks of each detected aneurysm.

The auxiliary vessel segmentation masks for the training data will be provided by the partnering TopBrain challenge, either verified by human experts (gold-standard) or generated (silver-standard) using an organizer-developed strong baseline model [8].

The size of the aneurysm is calculated from the diameter of the minimal bounding sphere via post-processing of the segmentation mask.

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.

Segmentations are manually corrected in 2D, when necessary, using the software ITK-SNAP or 3D Slicer. Most types of aneurysm (saccular, fusiform, dissecting) are segmented.

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 annotations are verified and approved by senior (over 10 years of experience) neurosurgeons or neuroradiologists from the organizing team.

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.

None, since each label is generated by a single clinician verifier.

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 data were anonymized (removal and anonymization of relevant DICOM patient information). Additional de-facing and cropping procedures were performed to ensure patient privacy in the image data after converting the DICOM to NIfTI format. The defaced and cropped image includes only the brain region. All images are re-oriented to LPS+ orientation. Other than the defacing, the cropping to brain region, and re-orientation steps, no further preprocessing of the data is performed to keep the data as close to the clinical setting as possible.

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.

- In case the aneurysm is located close to a bifurcation, or when an aneurysm has multiple touching adjacent vessels, the annotator may have high variability in deciding which vessel location class to assign to the aneurysm.
- For CTA modality, aneurysms along the internal carotid artery can be hard to detect, classify, and segment due to their proximity to the bones.
- The left and right sides of the anterior cerebral artery along the inter-hemispheric fissure can be hard to distinguish, especially for intertwined vessels. This can cause uncertainty in the aneurysm parent vessel location label.
- Images from different centers are not labeled by the same annotator, and different annotation styles can introduce variability. Aneurysm boundaries can be identified differently by radiologists from different centers, especially in small or low-contrast aneurysms (common in TOF-MRA), due to differences in judgment when distinguishing the aneurysm neck from the parent vessel or defining its extent in ambiguous regions.

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

CTA and MRA can have different resolutions. MRA might amplify the effect of stenosis in vessels. Both modalities might contain common artifacts such as flow-dependent signal cancellation artifacts, noise artifacts, ringing artifacts, pulsation artifacts, and beam hardening artifacts. These artifacts might cause the vessels and aneurysms to be over- or under-segmented, and might affect the diagnosis.

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)

Metrics are divided into detection and segmentation metrics, with each metric further subdivided by size groups. We adopt the same detection definition as the ADAM challenge [1], where a predicted aneurysm is considered a true positive if it spatially overlaps with the reference mask of the corresponding vessel-location class. Size groups are based on the diameter of the ground-truth aneurysm (e.g., <5mm, 5-10mm, >10mm) as measured by the diameter of the minimal bounding sphere.

For each of the aneurysm size groups (all-size, small-size, medium-size, and large-size), we compute the following detection and segmentation metrics:

Detection metrics:

Per-location precision
Per-location recall (sensitivity)
Per-location Matthews correlation coefficient (MCC)

Segmentation metrics (only for true positively detected aneurysms):

Per-location Dice Similarity Coefficient (DSC)
Per-location Hausdorff Distance 95% Percentile (HD95)
Per-location Volumetric Similarity (VS)

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

Similar to Task 1, we use precision, recall, and Matthews correlation coefficient (MCC) for per-location detection performance. Additionally, we report the detection metrics for stratified size groups. This is clinically important since smaller aneurysms can often be missed in screening and diagnosis.

Dice Similarity Coefficient (DSC): Measures the spatial overlap between the prediction and the ground truth. It is a simple and intuitive metric for volumetric agreement.

HD95 is more robust and reliable for small structures than DSC.

Volume similarity (VS) evaluates whether the model correctly estimated the total volume, regardless of the exact spatial overlap. This is critical for clinical follow-up and growth monitoring.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

The ranking for the awards of the MICCAI event will be based on the leaderboards displayed on the grand-challenge website, which rank the teams by sorting their performance on each metric. The leaderboard uses equal weights for each metric and ranks the submissions based on the mean rank position across all metrics. In case of identical mean ranks, the mean of per-location MCC and DSC serves as a tie-breaking metric due to its balanced handling of class imbalance.

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

Missing results will result in an increase in false negative detections. Missing results will not affect segmentation evaluation since we only evaluate segmentation metrics for true positive detections.

A method that fails to output a segmentation for a true aneurysm is counted as a false negative in detection evaluation, directly lowering recall and MCC for that anatomical class. Since detection/classification and segmentation metrics are equally weighted in the final ranking, strategies that sacrifice detection to inflate segmentation quality are penalized overall.

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

The decision to utilize an average rank across all metrics is driven by the multi-dimensional and cohort-based nature of the evaluation. This challenge evaluates performance across over 50 distinct anatomical locations and three different size tiers, and the detection metrics are computed per-location class across the cohort. Task 2 has six detection rankings and six segmentation rankings, ensuring the average rank reflects the most clinically robust performance.

Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

Post-challenge analysis will use bootstraps of the test set and analyze the ranking stability for bootstrapped average ranks.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

During challenge evaluation, bootstrap resampling with 1,000 iterations will be used to estimate 95 percent confidence intervals for the primary evaluation metrics on the full test set. For each algorithm and metric, we will report the mean performance together with its bootstrap-derived confidence interval and dispersion measure. This ensures that leaderboard results are accompanied by uncertainty estimates and do not rely solely on point estimates.

As post-challenge analysis, additional bootstrap-based analyses will be performed for clinically relevant subgroups, such as aneurysm size. Subgroup aggregated metrics will be reported as mean plus standard deviation with corresponding confidence intervals to assess performance heterogeneity.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

Variability of algorithm performance across test cases will be addressed at two stages.
During the challenge evaluation, dispersion across cases will be quantified by reporting the mean and standard deviation of metrics to ensure that leaderboard scores reflect not only average performance but also stability across the test set.

For post-challenge analysis, a more detailed variability analysis will be conducted. This will include examination of performance heterogeneity across clinically meaningful subgroups, such as aneurysm location or size category, by computing subgroup-specific means with 95 percent confidence intervals. Where appropriate, distribution-based visualizations will be used to identify potential outliers, which will be qualitatively reviewed to assess whether reduced performance is associated with anatomical complexity, small lesions, or annotation uncertainty.

This staged approach ensures a robust and fair ranking during the challenge while allowing for deeper interpretability analyses after completion.

Provide a description of how variability of rankings is assessed.

Variability of rankings will be assessed through bootstrap resampling of the test set using 1,000 iterations. For each bootstrap sample, evaluation metrics will be recomputed on the resampled cases using the submitted predictions, and algorithms will be ranked according to the predefined ranking scheme. This process yields a distribution of ranks for each algorithm across resampled test sets. Ranking stability will be summarized using the median rank and the 95 percent rank interval, thereby quantifying the robustness of the relative ordering of algorithms to sampling variability in the test set.

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Post challenge, pairwise comparisons between the top five performing algorithms will be conducted using statistical tests appropriate to the metric type. For paired case-level continuous metrics such as DSC, HD95, and VS, a two-sided Wilcoxon signed rank test will be used to test the null hypothesis that there is no difference in median paired performance between algorithms. The significance threshold will be set at alpha equal to 0.05 and 95 percent confidence intervals will be reported. Where multiple comparisons are performed, Benjamini–Hochberg false discovery rate (FDR) correction [9] will be applied. For cohort-level detection metrics derived from binary outcomes such as precision, recall, and MCC, McNemar's test will be used to compare paired classification disagreements between algorithms using contingency tables.

Provide a description of the missing data handling.

Not applicable since our data are images with verified labels.

Indicate any software product that is used for all data analysis methods.

Scipy and Numpy with custom Python scripts.

Further analyses

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

- combining algorithms via ensembling,
- inter-algorithm variability,

- common problems/biases of the submitted methods, or
- ranking variability.

We can perform ensembling of podium-finish top algorithms, and evaluate the ensemble prediction in the benchmark.

Post challenge, we will evaluate the top-performing models on an external private test set. This second phase is designed to evaluate the generalization capability of the algorithms. The results from this private evaluation represent a meaningful measure of a model's potential performance in a real-world clinical environment.

Post challenge, we will perform inter-rater agreement analysis on the annotations.

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] Timmins, Kimberley M., et al. "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: the ADAM challenge." Neuroimage 238 (2021): 118216.
[2] Jeff Rudie, Evan Calabrese, Robyn Ball, Peter Chang, Rennie Chen, Errol Colak, Maria Correia de Verdier, Luciano Prevedello, Tyler Richards, Rachit Saluja, Greg Zaharchuk, Jason Sho, Maryam Vazirabad and . RSNA 2025 Intracranial Aneurysm Detection. https://kaggle.competitions/rsna-2025-intracranial-aneurysm-detection, 2025. Kaggle. RSNA Intracranial Aneurysm Detection. https://kaggle.com/competitions/rsna-intracranial-aneurysm-detection, 2025. Kaggle.
[3] INSTED 2024 challenge, https://www.codabench.org/competitions/2139
[4] TopBrain 2025 MICCAI Challenge, https://topbrain2025.grand-challenge.org
[5] Large IA segmentation dataset, https://zenodo.org/records/6801398
[6] Di Noto, Tommaso, et al. "Towards automated brain aneurysm detection in TOF-MRA: open data, weak labels, and anatomical knowledge." Neuroinformatics 21.1 (2023): 21-34., https://openneuro.org/datasets/ds003949/versions/1.0.1
[7] de Nys, Chloe M., et al. "Time-of-Flight MRa of Intracranial aneurysms with Interval Surveillance, Clinical Segmentation and annotations." Scientific Data 11.1 (2024): 555., https://openneuro.org/datasets/ds005096/versions/1.0.3
[8] https://www.kaggle.com/competitions/rsna-intracranial-aneurysm-detection/writeups/2nd-place-solution
[9] Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal statistical society: series B (Methodological) 57, no. 1 (1995): 289-300.

Further comments

Further comments from the organizers.

We would like to express our gratitude to the reviewers for taking the time to evaluate our proposal.