Multi-Class Segmentation of Aortic Branches and Zones in Computed Tomography Angiography: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Multi-Class Segmentation of Aortic Branches and Zones in Computed Tomography Angiography

Challenge acronym

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

Aorta-CTA

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.

The aorta, the body's largest artery, can face potential threats like dissection and aneurysm, requiring prompt surgical intervention. Traditional surgical techniques for aortic disease often carry significant risks. Recent advancements in medical imaging, particularly computed tomography angiography (CTA), and minimally invasive approaches like endovascular grafting, offer a promising alternative. Accurate 3D segmentation of the aorta and its branches and zones on CTA is crucial for successful interventions. Inaccurate segmentation can lead to critical errors in surgical planning and endograft design, jeopardizing patient safety and treatment outcomes.

While machine learning has revolutionized 3D medical image analysis, its potential in acute uncomplicated type B aortic dissection (auTBAD), the most common aortic emergency, remains largely unexplored. In the clinical realm, auTBAD is categorized using SVS/STS zones, a detailed classification system defined by specific zones of the aorta in relation to aortic branches. Current methods for aortic segmentation often treat it as a binary segmentation problem, neglecting the essential differentiation between individual aortic branches and their relationships to SVS/STS zones.

This challenge addresses these limitations by offering the first large-scale dataset of 100 CTA volumes paired with detailed annotations for the aorta, its branches, and the clinically relevant SVS/STS zones. Participating teams will have the opportunity to develop innovative algorithms for accurate, automated, and multi-class segmentation of this intricate vascular structure.

By fostering advancements in image analysis techniques for CTA, this challenge aims to:
(1) Improve clinical care for patients with aortic diseases by enabling accurate diagnosis, more precise surgical planning, and potentially safer, minimally invasive interventions.

(2) Bring greater attention and research focus to auTBAD, a relatively rare and challenging disease, potentially leading to novel treatment strategies.
(3) Bridge interdisciplinary communication between researchers in medical image analysis, computer vision, and machine learning, paving the way for collaborative solutions to overcome technical barriers in complex aortic segmentation tasks.

In summary, this challenge has three main features:
(1) Task: this is the first challenge for the segmentation of aortic branches and zones in CTA scans.
(2) Dataset: we provide the largest annotated dataset for aortic segmentation, including 100 3D CTA scans.
(3) Evaluation: we focus on both segmentation accuracy and segmentation efficiency.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Aorta Segmentation, Computed Tomography Angiography, Branch and Zone Segmentation

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

N/A

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect to have 50 participating teams. The related Segmentation of the Aorta (SEG.A.) MICCAI challenge hosted in 2023 had 21 participating teams. We anticipate that our challenge will attract more participants since (1) our challenge allows for multi-class segmentation of 23 aortic branches instead of binary segmentation, and (2) we offer a much larger dataset of 100 CT scans.

Publication and future plans

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

We will submit a research paper on the challenge results to a top-tier journal, such as IEEE Transactions on Medical Imaging, Medical Image Analysis, or Nature Communications.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

The Grand-Challenge platform (https://grand-challenge.org) will host our challenge. Access to all data for training, validation, and testing will be through the grand-challenge platform. All algorithms will undergo testing on the identical computing platform provided by the Grand-Challenge. We welcome participation in various formats, such as purely virtual (e.g., via Zoom meetings), hybrid, and in-person.

TASK 1: Multi-Class Segmentation of Aortic Branches and Zones on Computed Tomography Angiography

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

The body's largest artery, the aorta, is susceptible to threats like dissection and aneurysm, necessitating surgical intervention. Recent advances in medical imaging, particularly computed tomography angiography (CTA), and minimally invasive approaches such as endovascular grafting offer a promising alternative. Accurate 3D segmentation of the aorta on CTA is crucial for successful interventions, as inaccurate segmentation can lead to critical errors in surgical planning and jeopardize patient safety.

Current aortic segmentation methods often overlook the differentiation between individual aortic branches and their relationships to clinically relevant zones. This task provides a large-scale dataset of 100 CTA volumes with detailed annotations for 23 different aortic branches and SVS/STS zones.

In summary, the task focuses on the segmentation of aortic branches and zones in CTA scans, provides the largest annotated dataset for aortic segmentation, and evaluates both segmentation accuracy and efficiency.

Keywords

List the primary keywords that characterize the task.

Aorta Segmentation, Computed Tomography Angiography, Branch and Zone Segmentation

ORGANIZATION

Organizers

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

Muhammad Imran (Department of Medicine, University of Florida, United States)
Jonathan R. Krebs (Department of Surgery, University of Florida, United States)
Michol A. Cooper (Department of Surgery, University of Florida, United States)
Jun Ma (Department of Computer Science, University of Toronto, Canada)
Yuyin Zhou (Department of Computer Science and Engineering, UC Santa Cruz, United States)
Wei Shao (Department of Medicine, University of Florida, United States)

b) Provide information on the primary contact person.

Wei Shao (University of Florida) wei.shao@medicine.ufl.edu

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Open call (challenge opens for new submissions after conference deadline)

Challenge venue and platform

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

N/A

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

grand-challenge.org (Type 2)

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

TBA

Participation policies

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

Fully automatic

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

No additional training dataset is allowed. This ensures fairness and consistency across submissions from all participating teams.

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.

All participating teams that successfully submit their Docker container and a 4-page technical paper will receive a participation certificate. The top three teams will receive award certificates and cash prizes (1st place: $2,000, 2nd place: $1,000, 3rd place: $500). Teams that achieve a test score higher than the baseline 3D U-Net model will be invited to co-author a journal paper about the challenge.

e) Define the policy for result announcement.

Examples:

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

The performance results of all teams will be made publicly available.

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).

Each participating team will be invited to submit a 4-page paper to describe their methods and results for the challenge. Their papers will undergo a brief review by the organizers. The challenge organizers will submit a research paper on the challenge results to a top-tier journal, such as IEEE Transactions on Medical Imaging, Medical Image Analysis, or Nature Communications. Teams that successfully submit their 4-page technical paper and achieve a test score higher than the baseline 3D U-Net model will be invited to co-author the journal paper.

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.

Participating teams are required to submit their solution via Docker container.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

We will make the validation dataset available in May, enabling participating teams to develop and validate their models. It is important to highlight that the ground truth aorta segmentations for scans in the validation dataset will not be provided to participants. Please be aware that the validation dataset is not meant for use in algorithm ranking.

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

Challenge website open and training data release: April 1, 2024
Validation data release and validation submission open: May 20, 2024
Test submission open: July 15, 2024
Short paper and final docker container submission deadline: August 15, 2024
Invite top 5 teams for oral presentation: September 1, 2024
Announcement of winning teams: October 6, 2024

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

We have received the Institutional Reviewing Board (IRB) approval (IRB202301895) from the University of Florida (UF) to collect and use these data for the purpose of research. Each participating team is required to submit a signed data usage agreement (DUA) with verified identity and agrees on all conditions specified in the Terms of Use.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC.
Additional comments: During the challenge, the dataset available for participating teams to download for model development consists of 50 labeled 3D CT images for training and 10 labeled 3D CT images for validation. All teams are required to sign a data usage agreement (DUA) before downloading data under the IRB protocol. After the challenge, participating teams can continue to use the data for non-commercial research purposes and must cite the challenge paper if they incorporate this dataset into their own publications.

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.

The model evaluation code will be released simultaneously with the training data. For transparency, the code used to calculate the final ranking will be made available after the closing date of the challenge.

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

Participating teams will be required to submit their 4-page technical paper along with a URL link to the GitHub repository containing the code for their implementations, in order to be considered as a co-author of the challenge paper.

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.

None

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Diagnosis; Prognosis; Research; Screening

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction

- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation

Cohorts

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

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

The target cohort is composed of patients with symptoms concerning for aortic dissection in whom CTA is medically indicated, or patients with a previous diagnosis of Type B aortic dissection in whom operative intervention is being considered and and CTA is indicated for operative planning purposes.

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

The challenge cohort is composed of patients with acute uncomplicated Type B aortic dissection that was diagnosed by CTA.

Imaging modality(ies)

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

Computed tomography angiography (CTA)

Context information

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

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

All the images will be deidentified by using the NIfTI format, stored as .nii.gz files.

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

All patients had a diagnosis of acute uncomplicated Type B aortic dissection without a prior history of aortic surgery and without aberrant aortic arch anatomy.

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.

CTA of the chest, abdomen, and pelvis.

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 aorta and its main branches (innominate, left common carotid, left subclavian, celiac, superior mesenteric, right renal, left renal, right common iliac, left common iliac, right external iliac, left external iliac), in addition to the anatomic zones of the aorta (zone 0 = aortic root through the innominate artery, zone 1 = between innominate artery and left common carotid, zone 2 = between left common carotid and left subclavian, zone 3 = first 2 cm distal to the left subclavian, zone 4 = end of zone 3 to the sixth thoracic vertebra, zone 5 = end of zone 4 to the celiac artery, zone 6 = celiac artery to superior mesenteric artery, zone 7 = superior mesenteric to renal arteries, zone 8 = end of zone 7 to the lower renal artery, zone 9 = end of zone 8 to the aortic bifurcation into left and right common iliac arteries, zone 10 = common iliac arteries, zone 11 = external iliac arteries).

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, Reliability, Runtime, Sensitivity, Specificity

DATA SETS

Data source(s)

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

The CTA chest, abdomen, and pelvis scans all included the standard non-contrast, arterial, and delayed phases. Because most patients were transferred to our institution, many CTAs were sourced from various local imaging centers with slight variations in imaging protocols. All CTAs were manually inspected for quality with a maximum slice thickness of 3mm.

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

Standard CTA comprises three phases: the non-contrast phase, the arterial phase, and the delayed phase. Initially, the non-contrast phase is taken to detect any hematoma or plaque on the vessel wall that might be concealed by iodinated contrast. This is followed by the arterial phase, where iodinated contrast medium (ICM) is rapidly injected for optimal arterial vessel visualization, ensuring the imaging coincides with the peak aortic/aortic branch contrast arrival. The delayed phase, accurately timed, follows to evaluate slow-filling and venous structures. As mentioned, given that many initial CTAs from TBAD patients were sourced from various local imaging centers with differing protocols, we manually reviewed images to ensure that all three phases were captured in slices of thickness less than 3mm. After reviewing image quality, the scans were de-identified and converted from deidentified DICOM files to a single NIFTI format file.

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.

CTA scans were obtained from patients admitted to the University of Florida Health Shands Hospital with a diagnosis of TBAD.

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).

CTA data were labeled by one of three image analysts: a board-certified vascular surgeon, or two general surgery residents with four years of dedicated post-graduate surgical training.

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.

In this challenge, a case comprises a 3D computed tomography angiography scan of a patient at a single time point, and the corresponding manually annotated label volume that delineates 23 different aortic branches and zones.

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

This dataset comprises 100 cases from the University of Florida. 50 cases will be used for training, 10 for validation, and 40 for final testing. This distribution ensures sufficient data for the reliable evaluation of model performance.

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

The dataset consists of 100 cases, with a 50/10/40 split for training, validation, and testing. The training set (50) provides the model with enough data to learn complex relationships and achieve robust performance. This is crucial for reliable application in real-world scenarios. The large testing set (40) allows rigorous assessment of the model's generalizability to unseen data. This balanced approach allows us to identify potential weaknesses in the model's performance for different types of cases, enabling further refinement and optimization before deployment. Ultimately, this data distribution prioritizes both model learning and performance evaluation, ensuring a well-trained and reliable tool for aortic segmentation analysis.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

At least half of the testing cases will be complex cases with unusual anatomy or thoracoabdominal aortic aneurysms, which are less common in clinical practice and underrepresented in our dataset. This enables us to evaluate both the accuracy and robustness of segmentation models.

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.

Manual image annotation

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

CTAs were collected and reviewed by a board-certified vascular surgeon to ensure adequate image quality. The three annotators, all physicians with clinical backgrounds in managing patients with aortic disease, underwent 3D Slicer training with the study PI to understand the basic commands for semi-automated segmentation and labeling. The annotators were asked to segment the aorta and its 13 main anatomical branches (innominate, left common carotid, left subclavian, celiac, superior mesenteric, right renal, left renal, right common iliac, left common iliac, right external iliac, left external iliac), in addition to the anatomic zones of the aorta (zone 0 = aortic root through the innominate artery, zone 1 = between the innominate artery and left common carotid, zone 2 = between left common carotid and left subclavian, zone 3 = first 2cm distal to left subclavian, zone 4 = end of zone 3 to the sixth thoracic vertebra, zone 5 = end of zone 4 to the celiac artery, zone 6 = celiac artery to superior mesenteric artery, zone 7 = superior mesenteric to renal arteries, zone 8 = end of zone 7 to the lower renal artery, zone 9 = end of zone 8 to aortic bifurcation into left and right common iliac arteries, zone 10 = common iliac arteries, zone 11 = external iliac arteries). A color-coded reference system was used to ensure label consistency and accuracy. A board-certified vascular surgeon reviewed all segmentations for accuracy by comparing the CTA to the segmented output, and segmentations were further refined based on that feedback.

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.

CTA data were labeled by one of three image analysts, all physicians: a board-certified vascular surgeon, and two general surgery residents, each with four years of dedicated post-graduate surgical training, using a semi-automated segmentation tool on 3D Slicer.

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.

CTA data were labeled by one of three image analysts: a board-certified vascular surgeon, or two general surgery residents, each with four years of dedicated post-graduate surgical training. All cases annotated by the surgical residents were reviewed for accuracy by the vascular surgeon.

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.

PHI was removed from the DICOM metadata header.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

The SVS/STS aortic zone classification system is based on a 2D representation of the aorta, so the main sources of error are overlapping consecutive zones of the aorta based on the orientation of the aorta used for segmentation. In addition, the branch points of the main aortic branches are highly variable between patients, and in some cases, they may be diminutive or have poor contrast filling, leading to variable interpretations of the branch origins. Unlike solid organs like the liver, spleen, and kidney, aortic anatomy is highly variable, particularly in patients with aortic pathology like in this dataset. Because of this, acceptable inter- and intra-annotator dice coefficients of 0.7 may be acceptable.

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

Other sources of error include misidentification of the aortic wall, which may result in the undersegmentation of the aorta. By definition, an aortic dissection is a tear in the aorta, meaning that two separate lumens are created which have distinct appearances on contrasted scans like CTA. In certain patients, the outer wall of the aorta can be difficult to differentiate from surrounding tissue.

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)

We will use the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) for evaluation. Specifically, we compute the Dice Similarity Score (DSC) and the Normalized Surface Distance (NSD) for each class and then calculate the mean of DSC and NSD across all classes.

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

The metrics are complementary. Specifically, DSC and NSD are used to measure the region error and boundary error, respectively.

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.

This is a popular ranking scheme that can aggregate different metrics. A similar ranking scheme was also employed in the MICCAI BraTS 2017-2023 and FLARE 2021-2023 Challenges.

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

Missing results on test cases will result in the ranks being set to the worst (i.e., the number of participating teams). Specifically, missing results will get a zero value for DSC, NSD and the equivalent worst value for running time, and area under GPU memory-time curve.

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

The ranking scheme includes the following three steps:
Step 1. For each testing case, we compute the DSC and NSD metrics.
Step 2. Rank participants for each of the 40 testing cases and each metric; Thus, each participant will have 80 (40x2) rankings.
Step 3. Average all these rankings.

Statistical analyses

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

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

When the segmentation result is missing for a class label, we will assign the poorest performance, i.e., a DSC of 0 and a NSD of 1. This ensures that there is no missing data for the statistical analysis. Moreover, bootstrapping will used to assess the variability of rankings.

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

Bootstrap is a simple nonparametric method that relies on minimal assumptions.

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 will analyze combining the top performing algorithms via model ensembling.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

Further comments

Further comments from the organizers.

N/A