Multi-class Brain Hemorrhage Segmentation in Non-contrast Computed Tomography under Limited Annotations: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Multi-class Brain Hemorrhage Segmentation in Non-contrast Computed Tomography under Limited Annotations

Challenge acronym

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

MBH-Seg

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.

Multi-class brain hemorrhage segmentation in biomedical imaging is pivotal for accurate diagnosis and treatment planning. Different types of brain hemorrhages (like subdural, epidural, and intracerebral) have distinct implications for patient care. Accurately segmenting and identifying these types can lead to more personalized and effective treatment strategies. Current diagnostic methods, primarily reliant on expert radiologists interpreting non-contrast CT scans, face challenges like variability in interpretation and time constraints. The proposed challenge aims to revolutionize this process by leveraging advanced segmentation techniques, enabling rapid, more consistent, and accurate diagnosis, potentially reducing mortality and improving patient outcomes.

Technically, this proposed challenge addresses a key limitation in medical imaging: the scarcity of detailed annotations. Developing algorithms capable of accurate segmentation with limited data is a significant leap in machine learning, particularly in medical applications. It pushes the boundaries of semi-supervised and unsupervised learning models, fostering innovation in algorithm development and data efficiency.

The envisioned impact is multidimensional, enhancing medical imaging software, aiding clinical decision-making, and establishing new standards in computational diagnostics. This challenge drives technological advancements and has the potential to transform patient care in neurology and emergency medicine, making it a pivotal intersection of technology and healthcare.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Multi-class Brain Hemorrhage Segmentation, Non-contrast Computed Tomography, Limited Annotations

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

None

Duration

How long does the challenge take?

Half day.

Expected number of participants

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

Reflecting on the participation in similar themed challenges at previous MICCAI events, we anticipate attracting between 200 to 300 teams for this year's challenge. Engagement efforts have already been made with approximately 10 potential teams. This forecast is based on our networking reach, marketing strategies, and the appeal of the challenge's theme. We aim to bring together professionals and teams from various research areas and backgrounds, thereby enriching the diversity and competitive spirit of the challenge.

Publication and future plans

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

Yes, coordinating a publication of the challenge results is planned. This publication will provide a detailed overview of the methodologies, findings, and innovations presented in the challenge. It will serve as a valuable resource for the broader research community and contribute to advancing the field, particularly in medical imaging and machine learning applications. The publication is expected to be in a peer-reviewed journal or as part of the conference proceedings, ensuring wide dissemination and recognition of the participants' work and the challenge's impact.

Space and hardware requirements

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

We are setting a stringent limit on the computational resources available for the solutions. The models must be efficient enough to run within a maximum timeframe of 120 seconds per input sample on the Grand Challenge backend. Any solutions exceeding these time constraints will not be eligible for leaderboard consideration. This policy is adopted for multiple reasons: Firstly, to minimize costs, as each model evaluation on Grand Challenge entails financial implications. Secondly, to mirror real-world conditions where computational and memory resources and algorithm runtime are often substantially constrained. Lastly, this limitation ensures more

equitable competition, particularly for those participants who might not have access to extensive computational power.

TASK 1: Multi-class Brain Hemorrhage Segmentation with Limited Labeld data and Large-scale Unlabeled data

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.

This challenge addresses a critical need in neurology: the precise segmentation of different types of brain hemorrhages in non-contrast CT scans, a key to accurate diagnosis and effective treatment planning. The difficulty of acquiring large-scale, pixel-level annotated medical imaging data, which is crucial for training deep learning models, is well-recognized. Such data acquisition is not only expensive but also demands specialized medical expertise. Conversely, unlabeled data is more readily available but underutilized due to the complexity of extracting meaningful information without annotations. In this task, we provide a unique dataset comprising a limited amount of meticulously labeled data and a significantly larger pool of unlabeled data. This setup challenges participants to innovate in semi-supervised learning, advancing the application of deep learning in medical imaging, and paving the way for more accurate, accessible, and efficient diagnostic tools in clinical practice.

Keywords

List the primary keywords that characterize the task.

Multi-class Brain Hemorrhage Segmentation, Non-contrast Computed Tomography, Limited Annotations, Unlabeled data

ORGANIZATION

Organizers

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

Organizers
1. Yutong Xie, Researcher, University of Adelaide, Australia
2. Minh-Son To, Medical Practitioner, Flinders Health and Medical Research Institute, Flinders University, Australia
3. Chenyu Wang, Researcher, Brain and Mind Centre, University of Sydney, Australia
4. Dongang Wang, Researcher, University of Sydney, Australia
Advisor
1. Qi Wu, Assoiate Professor, University of Adelaide, Australia
2. Yong Xia, Professor, Northwestern Polytechnical University, China
Contributors
1. Biao Wu, Student, University of Adelaide, Australia
2. Zeyu Zhang, Student, Australian National University, Australia
3. Vu Minh Hieu Phan, Researcher, University of Adelaide, Australia
4. Yifan Liu, Lecturer, University of Adelaide, Australia

b) Provide information on the primary contact person.

Yutong Xie, email: yutong.xie678@gmail.com

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

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

N/A

Participation policies

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

Only automatic methods are allowed

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

No additional data is allowed. We will require all participants to provide extensive documentation on their development processes, including data and methods used. Top-ranked participants must submit their training code and checkpoints for verification. Publicly available pre-trained models are allowed, ensuring a level playing field and transparency in the competition's evaluation process.

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 the leaderboard.

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

1. Certificates/Awards for the top-3 teams, the team with the most innovative method, and the teams with the careful method description.

2. The top-10 teams will be invited to give oral presentations at the MICCAI 2024 Event (either virtually or in person).
3. Co-author of the challenge paper, which will be submitted to a top journal (MedIA/TMI).

e) Define the policy for result announcement.

Examples:

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

The top-10 teams will be notified one week before the challenge day to prepare their presentations. Final results and awards will be announced on the challenge day.

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).

1. Authorship Qualification: Members of top-10 teams who have made significant contributions to the conceptualization, design, execution, or interpretation of the research presented are qualified as authors.
2. Independent Publications: Participating teams are allowed to publish their results separately. However, this should be done in a manner that respects the collective efforts of the challenge.
3. Embargo Period: An embargo period allows the challenge organizers to publish a comprehensive challenge paper first. During this period, individual teams are encouraged to refrain from publishing their complete findings independently.

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.

Algorithm container submission (type 2) on Grand Challenge.

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.

This challenge allows the submission of multiple results, and only the last run is officially counted to compute challenge results. We will provide a publicly available leaderboard for the 1st validation phase. The results of the 2nd-Testing phase can only be seen by organizers and participants during the submission phase. The final leaderboard for the 2nd-Testing phase will be announced on the challenge day.

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

1. Release of training data: May 15th, 2024
2. Submission for the 1st-Validation phase: July 15th, 2024 - August 15th, 2024
3. Submission for the 2nd-Testing phase: August 15th, 2024 - September 1th, 2024
4. Announcement of invited presentations: September 28th, 2024
5. Announcement of final leaderboards: October 6th, 2024 @ MICCAI 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 reannotate the open datasets in this challenge, hence no formal ethics review and IRB approval was required. Each institution contributing to the open datasets secured the approval of its institutional review board and institutional compliance Officers.

Data usage agreement

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

Examples:

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

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

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 evaluation software developed by the organizers for this challenge will be made accessible to all participants. The code, which is used for producing rankings and evaluating submissions, will be hosted on a public repository (e.g., GitHub). This will include detailed instructions for its use and the supported platforms, ensuring transparency and fairness in the evaluation process. A link to the repository will be provided on the challenge website, allowing participants to download, review, and use the code for preparing their submissions. This approach ensures that all participants have equal access to the tools necessary for participating in the challenge.

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

For the participating teams in the challenge, it is encouraged to make their code publicly available upon completion of the challenge. Top-ranked participants must submit their training code and checkpoints for verification. This transparency fosters a collaborative environment and facilitates further research. Teams can host their code on platforms like GitHub or Bitbucket. Upon challenge completion, teams should provide a link to their code repository, ensuring it is accessible to the broader research community. This policy not only enhances the reproducibility of the research but also contributes to the collective advancement of the field.

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.

The funding of this challenge is from the Sydney Neuroimaging Analysis Centre, Australia.
Only organizers can access the test case labels.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Diagnosis; Research; Training

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 for this task involves patients who have undergone non-contrast CT scans for suspected brain hemorrhage. This group would primarily include individuals exhibiting symptoms of intracranial bleeding, such as sudden severe headaches, loss of consciousness, or neurological deficits. The cohort is not restricted by sex or age, as brain hemorrhages can occur across various demographics. However, certain subgroups like older adults or individuals with a history of hypertension may be more prevalent. The final biomedical application is aimed at this diverse patient population, ensuring the developed models are broadly applicable and beneficial in real-world clinical scenarios.

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

This cohort includes a diverse set of patients who have undergone CT scans for suspected brain hemorrhages, encompassing various types and severities of brain hemorrhages. The data set includes both a small subset of pixel-level labeled images and a larger set of unlabeled images, reflecting the real-world scenario of limited labeled data availability in medical settings. This cohort was carefully selected to provide a realistic and challenging dataset for developing and testing advanced machine learning models in this critical area of medical imaging.

Imaging modality(ies)

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

Non-contrast Computed Tomography

Context information

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

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

Annotations of Tumor volume

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

Image Acquisition Details: Technical specifications like the make and model of the CT scanner, scan settings, and date of the scan.

Target entity(ies)

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

The image data is acquired from the brain region of subjects, as shown in computed tomography (CT) data. This encompasses the entire cranial cavity, focusing on detecting and segmenting various brain hemorrhages. The data origin for both the target and challenge cohort is consistent in this respect, utilizing brain CT scans to develop and validate segmentation models. This specific focus on brain imaging ensures the relevance and applicability of the models to real-world clinical scenarios involving brain hemorrhage diagnosis.

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.

Brain hemorrhage segmentation for multiple classes: extradural hemorrhage, subdural hemorrhage, subarachnoid hemorrhage, intraparenchymal hemorrhage, and intraventricular hemorrhage

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)).

Find highly accurate brain hemorrhage segmentation algorithm for non-contrast CT images. Corresponding metrics include Dice score, Hausdorff distance, Sensitivity, and Runtime.

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).

Did not retain the manufacturer, model information

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

1. Image Plane - Axial only
2. Resolution - 512 x 512 original
Original for at least one site
Downsampled/re-sampled to 5mm for one site
3. Pre-processing
Not performed by sites submitting data
4. Burned-in PHI
CTP Anonymizer and other programs used

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

The original CT brain images was compiled from the clinical picture archiving and communication system archives from three institutions: Stanford University (Palo Alto, Calif), Universidade Federal de São Paulo (São Paulo, Brazil) and Thomas Jefferson University Hospital (Philadelphia, Pa). In this challenge, we annotate this dataset and provide pixel-level labels.

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).

Each contributing institution secured the approval of its institutional review board and institutional compliance officers. Methods for examination identification, extraction, and anonymization were unique to each contributing institution: Universidade Federal de São Paulo provided all brain CT examinations performed during a 1-year period, Stanford University preselected examinations based upon a normal versus abnormal assessment of radiology reports to provide a 50/50 sample of positive (for any abnormality) to negative examinations, and Thomas Jefferson University Hospital extracted cases using simple natural language processing on radiology reports mentioning specific hemorrhage subtypes.

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" refers to the entire set of data associated with an individual patient's CT scan of the brain. Each case includes the CT image data, which is the primary focus for the segmentation task. The training cases come with limited pixel-level annotations identifying different types of brain hemorrhages, while the test cases might have either similar weak annotations or no annotations at all, depending on the specific design of the challenge. This setup allows for the development and testing of models that can work effectively with varying levels of available labeled data.

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

A high-quality medical imaging dataset comprising 2192 high-resolution 3D CT scans of the brain, each containing between 24 to 40 slices of 512 × 512 pixels in size. It contains 192 volumes with pixel-level annotations and 2000 CT scans without any annotations. We split the labeled images into a training set with 96 cases, a validation set with 48 cases and a test set with 48 cases. All unlabeled images are used for training.

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

These sets were chosen to balance the number and types of bleeds. All splits provided will be with a mixture of health and patients.

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.

In the challenge, the training, validation, and test cases are characterized by the following:
1. Class Distribution: The cases are selected to reflect the real-world distribution of different types of brain hemorrhages. This choice ensures that the developed models apply to clinical scenarios and can handle the variability seen in real patient data.
2. Variability in Imaging Quality: The cases include a range of imaging qualities and conditions to simulate the diversity encountered in clinical practice. This includes variations due to different CT scanners and patient conditions.
3. Annotation Quality: The training cases feature pixel-level annotations, but with limited availability to mimic the common scenario in medical settings. The test cases may have minimal or no annotations to challenge the algorithms' ability to generalize and perform in less-than-ideal conditions.
These characteristics ensure that the challenge realistically reflects clinical conditions and prepares the algorithms for practical deployment.

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 by three medical imaging experts.

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.

None

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.

Three medical imaging experts performed pixel-level annotations in two stages. Hemorrhages on individual head scans were independently segmented using ITK-SNAP by two trained medical imaging experts and radiology residents, both with over one year of experience reading CT head scans, using the original image-level hemorrhage annotations as a guide. These annotations were then reviewed by a board-certified radiologist with over five years of post-fellowship experience, ensuring the quality of the annotations.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Not performed by sites submitting data

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.

1. Inter-Annotator Variability: Differences in how different annotators perceive and mark the same features in an image. This variability can be significant, especially in complex cases with subtle features.
2. Intra-Annotator Variability: The inconsistency of a single annotator when labeling similar features across different images or at different times.
3. Annotation Ambiguity: In cases where the hemorrhage boundaries are unclear, annotations could be ambiguous.

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

Technical Limitations: Errors introduced due to limitations of the annotation tools or image resolution.

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)

Dice Similarity Coefficient (DSC), Hausdorff distance, Sensitivity, and Inference Time.

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

1. Dice Similarity Coefficient (DSC): DSC measures the overlap between the predicted segmentation and the ground truth. It's highly relevant for medical image segmentation to ensure accurate delineation of affected areas.
2. Hausdorff Distance: This measures the maximum distance of the set of points of one segmentation from the nearest point in the other segmentation. It's crucial for assessing the accuracy of boundary delineation in brain hemorrhages.
3. Sensitivity: This metric assesses the true positive rate, which is important in medical applications to ensure all relevant areas are identified without missing critical information.
4. Inference Time: In a clinical setting, the speed of obtaining results is vital for timely diagnosis and treatment.

These metrics collectively ensure that the developed models are accurate, reliable, and clinically applicable.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

The chosen ranking scheme is justified as it balances the need to accurately reflect the performance of each algorithm comprehensively. The scheme provides a holistic view of each submission's capabilities by aggregating results across all test cases and metrics. This approach ensures that the final ranking is not only based on one aspect of performance but also considers accuracy, sensitivity, and practical applicability, which are crucial for clinical implementation. Such a ranking method aligns with the challenge's goal of identifying robust and reliable solutions for real-world medical applications.

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

For submissions with missing results on test cases, we will penalize them by assigning a score equivalent to the worst performance observed in that metric among all submissions. For example, if the lowest Dice score across all submitted and complete cases is 0.5, then missing Dice scores will be assigned a 0.5. A disqualification threshold is set at more than 20% missing data across all test cases; submissions exceeding this threshold will not be considered for final ranking. This ensures accountability and completeness in submissions.

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

The performance rank of submitted algorithms is computed by aggregating the results obtained per case for each metric. The aggregation method will involve calculating the mean score for each metric across all test cases. The final ranking will be based on a weighted sum of these mean scores, with weights assigned according to the importance of each metric for the task.

To set specific weights for the evaluation metrics of the Dice Similarity Coefficient, Hausdorff Distance, Sensitivity, and Inference Time, we plan to use one approach:
--Dice Similarity Coefficient (DSC): 40% - Given its importance in measuring the overlap between the predicted

segmentation and the ground truth, a higher weight reflects its significance in accurate segmentation.
--Hausdorff Distance: 25%- This metric is crucial for evaluating the maximum distance between the predicted and actual boundaries, important for precise segmentation.
--Sensitivity: 25% - Critical for ensuring that the algorithm accurately identifies all relevant areas of hemorrhage, weighted equally with Hausdorff to emphasize accuracy.
--Inference Time: 10% - While speed is important, in clinical settings, the accuracy and precision of segmentation take precedence.
These weights are designed to prioritize accuracy and precision in segmentation, reflecting the critical nature of these aspects in medical diagnosis while still considering the practical aspect of inference speed.

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.

In the challenge analysis, the statistical methods include:
1. Handling Missing Data: We might use multiple imputation techniques to handle missing data, ensuring that the analysis remains robust despite incomplete information.
2. Variability of Rankings: Statistical measures like standard deviation or confidence intervals for each algorithm's performance across different metrics can be calculated to assess the variability. Additionally, non-parametric tests like the Wilcoxon signed-rank test could be employed to compare the performance of different algorithms.
3. Assessment of Data Assumptions: Before applying any statistical method, tests like the Shapiro-Wilk test for normality or Levene's test for homogeneity of variances can be conducted. These tests ensure that the data meets the assumptions of the chosen statistical methods.
4. Statistical Software: Software like R or Python's SciPy and Pandas libraries can be used for all data analysis. These platforms offer robust and versatile tools for statistical analysis and are widely accepted in scientific research.

This approach ensures a comprehensive and scientifically rigorous evaluation of the challenge submissions.

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

The chosen statistical methods are justified as they offer a comprehensive and unbiased approach to evaluating the challenge submissions. Multiple imputations for missing data ensure a complete analysis dataset, preserving statistical power. Assessing the variability of rankings through non-parametric tests provides insights into the consistency and reliability of the algorithms. Validating data assumptions guarantees the appropriateness of the statistical techniques used. Utilizing robust statistical software like R or Python's libraries ensures the accuracy and reproducibility of the analysis. This combination of methods ensures a fair, rigorous, and scientifically sound evaluation process.

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.

N/A

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.

This challenge is based on the papers:
[1]. Wu B, Xie Y, Zhang Z, et al. BHSD: A 3D Multi-class Brain Hemorrhage Segmentation Dataset[C]//International Workshop on Machine Learning in Medical Imaging. Cham: Springer Nature Switzerland, 2023: 147-156.
[2] Flanders A E, Prevedello L M, Shih G, et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge[J]. Radiology: Artificial Intelligence, 2020, 2(3): e190211.

Further comments

Further comments from the organizers.

N/A