Ischemic Stroke Lesion Segmentation Challenge 2024: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Ischemic Stroke Lesion Segmentation Challenge 2024

Challenge acronym

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

ISLES'24

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.

Accurate segmentation of ischemic lesions in stroke is essential both during acute stages to guide treatment decisions (e.g., determine a patient's eligibility for thrombectomy treatment), and during sub-acute and chronic stages for evaluating disease outcomes, clinical follow-up, and for defining optimal therapeutic and rehabilitation strategies to maximize critical windows for recovery. The Ischemic Stroke Lesion Segmentation (ISLES) Challenge (https://www.isles-challenge.org/), an initiative aimed at advancing stroke image analysis, involves neurointervention, radiology, and computer science professionals and researchers from multiple leading institutions in the field. The ISLES challenge has been hosted at five MICCAI conferences (2015, 2016, 2017, 2018, 2022). In its inaugural edition, ISLES'15 [1], participants were tasked with segmenting sub-acute ischemic stroke lesions from post-interventional MRI and acute perfusion lesions from pre-interventional MRI. Subsequent editions, ISLES'16 and ISLES'17 [2], focused on stroke outcome prediction, requiring the segmentation of follow-up stroke lesions from acute multimodal MR imaging and the estimation of patient outcome disability scores. In ISLES'18 [3], acute stroke segmentation was approached indirectly and in a cross-modality fashion, with teams predicting the core tissue delineated in concomitant MRI from acute perfusion CT series. The most recent challenge, ISLES'22, addressed acute, sub-acute, and chronic ischemic stroke segmentation in over 2000 MRI scans [4,5]. Over the years, ISLES events have garnered significant attention from the research community. There were 120 database downloads until the ISLES'15 challenge day with 14 participating teams, and the number of participating teams was roughly duplicated in the ISLES'18 edition. The datasets released in ISLES'22 have been downloaded over 2000 times each, serving as reference datasets in prominent clinical and image analysis research [6,7,8]. The ISLES challenge has played a pivotal role in stroke image analysis for over eight years, contributing to the development of open stroke imaging datasets and benchmarking state-of-the-art image processing algorithms. Based on the experience gained from these previous editions, ISLES'24 seeks to benchmark final infarct segmentation algorithms. Unlike ISLES'16-17 editions, which share common points with this year's one, ISLES'24 makes use of standard-of-care acute stroke CT imaging (including non-contrast CT, perfusion CT, and CT angiography) and sub-acute stroke MRI (follow-up DWI with delineated infarct labels),

coupled with clinical and demographic tabular data, highlighting its clinical relevance. Furthermore, we will release four times more imaging data through this challenge than ISLES'17. ISLES'24 aims to identify prominent final infarct segmentation algorithms from pre-interventional data, thus providing outputs that could, from a clinical standpoint, help optimize reperfusion treatment decision-making. Overall, the diversity of ISLES'24 imaging and clinical data and the clinical relevance of the task will provide participants with a unique challenge.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

ischemic stroke, final infarct, 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.

ISLES'24 is part of the half-day Stroke Workshop on Imaging and Treatment CHallenges (SWITCH2024). During SWITCH24, a one-hour slot is dedicated to the ISLES'24 challenge, where top-3 ranked teams will present their methods and the results of the challenge will be announced. Note that no additional space/time is required.

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 estimate 30 participating teams based on:
- Previous ISLES challenge editions (2015 - 2022). There were >400 registrations to each of the two tasks from the previous challenge editions, and 20 submissions to the final test phase.
- The large amount of database download requests received in previous challenge editions. The ISLES'22 [4] and the ATLAS v1.2 datasets [5] have been downloaded, each of them, more than 2000 times up to now.

Besides, we will advertise the challenge in existing mailing lists, social media, and we aim to also send emails to previous ISLES participants.

Publication and future plans

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

Yes. As done in previous ISLES editions, where the benchmarking of algorithms has been published in top ranked journals (as Medical Image Analysis for ISLES'15 [2], Stroke for ISLES'18, Scientific Data for ISLES'22 dataset). For this year's edition, we aim to publish two works from this challenge: one describing the dataset and one with the

challenge benchmarking results.

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.

ISLES'24 is an off-site challenge to be hosted in https://grand-challenge.org/
Participants will run their algorithmic docker submissions in computers with GPU capabilities. For the challenge day during SWITCH'24 workshop, we will require a projector, two microphones and speakers.

TASK 1: Final, post-treatment infarct segmentation from pre-treatment acute imaging and clinical 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.

Background and aims:
Clinical decisions regarding the treatment of ischemic stroke patients depend on the accurate estimation of core (irreversibly damaged tissue) and penumbra (salvageable tissue) volumes [14]. The clinical standard method for estimating perfusion volumes is deconvolution analysis, consisting of i) estimating perfusion maps through perfusion CT (CTP) deconvolution and ii) thresholding the perfusion maps [15]. However, the different deconvolution algorithms, their technical implementations, and the variable thresholds used in software packages significantly impact the estimated lesions [16]. Moreover, core tissue tends to expand over time due to irreversible damage of penumbral tissue, with infarct growth rates being patient-specific and dependent on diverse factors such as thrombus location and collateral circulation. Understanding the core's growth rate is clinically crucial for assessing the relevance of transferring a patient to a comprehensive stroke center based on transport times [17]. Moreover, since not every reperfusion treatment with mechanical thrombectomy achieves complete reperfusion, predicting infarct growth might provide interventional radiologists with insights into the potential benefits of additional reperfusion attempts. Therefore, anticipating the temporal core evolution from acute imaging data is key for clinical decision-making [17]. This challenge task aims to segment the final stroke infarct from pre-interventional acute stroke data. It is worth noting that, unlike the clinical standard for core estimation, participants in this challenge will have access to non-contrast CT (NCCT) and CT angiography (CTA) modalities. While NCCT might help identify infarct areas not visible in CTP (e.g., in patients with spontaneous reperfusion), CTA can provide insights into the collateral circulation status and the thrombus location.

Challenge Data:
Data inputs include acute CT images (NCCT, CTP and CTA) and tabular data (demographic and clinical data). Each team decides which data modalities are used as input to their algorithms. The output of the algorithm is a binary infarct segmentation mask.

Challenge Algorithm Submission:
Algorithms are submitted as Dockers through our challenge platform.

Keywords

List the primary keywords that characterize the task.

Final infarct, segmentation, lesion evolution

ORGANIZATION

Organizers

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

Ezequiel de la Rosa
University of Zurich, Switzerland

Jan S. Kirschke
Klinikum rechts der Isar, Technical University of Munich, Germany

Benedikt Wiestler
Klinikum rechts der Isar, Technical University of Munich, Germany

Roland Wiest
University of Bern, Switzerland

Mauricio Reyes
University of Bern, Switzerland

Ruisheng Su
Erasmus University Medical Center, The Netherlands

Susanne Wegener
University of Zurich, Switzerland

Bjoern Menze
University of Zurich, Switzerland

b) Provide information on the primary contact person.

Ezequiel de la Rosa (ezequieldlrosa@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

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'24

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

www.grand-challenge.org

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

https://www.isles-challenge.org/

Participation policies

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

Fully automatic.

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

Participants are allowed to use private data. However, if they choose to train models with private data, they must also include in their GitHub repositories a second model that is trained exclusively with the challenge dataset.

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.

Members of the organizers institutes that are not associated with the challenge may participate and are eligible for awards.

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

N/A

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 three top-ranked methods will be announced publicly during SWITCH2024.

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).

We will write a paper that summarizes the results of this challenge.
1) Two authors from every submission (first and last) will be included as coauthors of this paper.
2) Selected approaches based on innovation and performance will be invited to describe their methods in the challenge manuscript.
3) Participating teams can submit their results separately without any 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.

Participants submit a docker through our evaluating platform. Submission instructions will be shared through our website. Besides, we will release (via Git) a docker template that participants must use to build their solutions. Under exceptional deployment failures, participants will be contacted to fix and resubmit their dockers.

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.

There are 3 phases for this challenge:
- Train phase: Teams can evaluate the performance of their trained models by themselves. With this purpose, we will release together with the first batch of training data, a Python evaluation script that computes the performance metrics defined in this document (please check the Assessment Methods section). Evaluation scripts will be shared through GitHub. It is important to mention that there is no 'validation' set for ISLES'2024. However, participants are strongly encouraged to take validation sub-sets from the training data in order to validate their models.

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

- Test phase: Participants submit a docker that will be locally evaluated by our team over the test data. Only one submission to this phase is allowed. No evaluation or ranking will be shared until the submission system is closed by the end of the challenge. For consistency, the same evaluation scripts provided during the 'train phase' will be used for computing the different teams' performance metrics and the leaderboard.

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

- Release of Training data (1st batch): 15th of May 2024
- Release of Training data (2nd batch): 15th of June 2024
- Opening of submission system for dockers: 1st of July 2024
- Closing of submission system for dockers: 15th of August 2024

Ethics approval

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

All data are derived from studies that were approved by their local ethics committee and were conducted in accordance with the 1964 Declaration of Helsinki. The ethics committee at the receiving site (the University of Zurich) approved sharing of the de-identified data, which do not contain any of 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)

CC BY

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

A Python script for evaluating the results will be shared together with the 1st batch of train-phase data. Code will be made available through a Github repository.

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

In order to ensure reproducibility and credibility of the algorithms, and in order to make the best-performing methods available for the research community, algorithms submitted to this challenge must use a working Github repository with a permissive open source license (e.g. Apache 2.0).

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.

No conflicts of interest. Jan S. Kirschke and Benedikt Wiestler will have access to a subset of the labeled dataset by the time the first-batch of data is released. Susanne Wegener and Ezequiel de la Rosa will have access to the full labeled dataset by the time the second-batch of data is released.

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, intervention planning, final outcome prognosis

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.

Acute ischemic stroke patients in routine clinical practice with a standard-of-care acute clinical imaging protocol (NCCT, CTA, CTP).

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

Retrospective cohort of acute ischemic stroke patients that evidenced brain infarct lesions in follow-up (1-to-7 days) MRI (DWI) and that received successful reperfusion treatment at acute stroke stages.

Imaging modality(ies)

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

- CT (Non-contrast CT, perfusion CT, and CT angiography).
- MRI (DWI and FLAIR)

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).

Clinical tabular data will also be released, including demographics (age and sex), medical history when available (diabetes mellitus, hypertension, etc), acute clinical data (NIHSS at admission, time since symptoms onset, TICI 2b/3 recanalisation, etc.) and outcome clinical data (90-day mRS, NIHSS at discharge).

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 CT (CTP, CTA and NCCT) and MRI (DWI, FLAIR)

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 infarcts.

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 accurate final post-treatment infarct segmentation algorithms from pre-treatment CT imaging data.

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).

CT scanners:
- Center 1: >80% of the data have been acquired with Siemens scanners (Somatom Force and Somatom Xcite).
- Center 2: >80% of the data was acquired with one of the following scanners: Siemens Somatom AS+, Philips Brilliance 64, Philips Ingenuity.

MRI scanners:
- Center 1: >90% of the data have been acquired with Philips or Siemens, 1.5 or 3T scanners.
- Center 2: >90% of the data have been acquired at one of the following scanners: Siemens Verio, Philips Achieva, Philips Ingenia, all 3T.

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

CT scans include three imaging modalities (CTP, CTA and NCCT). MRI scans include DWI and FLAIR modalities. All images are obtained with acquisition parameters spanning within normal ranges used for clinical purposes.

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: University Hospital of Zurich, Switzerland
Center 2: University Hospital of Munich, Germany

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 the CT and MRI devices 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.

Training and test cases represent CT scans (CTP, NCCT and CTA) of the brain. Training cases additionally include a DWI-MRI with its corresponding ground-truth (voxel-wise binary mask with True labels within the infarcted area and False labels within the non-infarcted areas).

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

Train set: 200 cases.
Test set: 100 cases.

c) 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 (segmentation task).
-The effort needed to identify and retrieve the data at two time-points (acute and follow-up) and the clinical tabular data from the centers,
-The effort needed to manually annotate the ground-truth at a voxelwise level.
From our experience in previous challenges, we believed that 300 scans is a reasonable amount of data to devise and evaluate algorithms for the task at hand.
The data is split in an approximate train-test proportion of 2/3-1/3. Both the train and test sets include a wide range of stroke lesions (infarct sizes and vascular territories affected). 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.

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

The train and test phases comprise data from centers 1 and 2. In the train phase, 100 cases are retrieved from center 1 and 50 cases from center 2. In the test phase, 50 cases are retrieved from center 1 and 50 cases from center 2. The data used within the test-case will belong to different patients. This ensures that images coming from a same patient, but acquired at different times, are not included in both the train and test phases.

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.

ISLES24 ground-truth annotation consists in a hybrid AI-human approach, as performed in the previous ISLES'22 [4] challenge edition. A robust, leading infarct segmentation algorithm devised in ISLES'22 will be used to pre-annotate DWI infarctions. Afterwards, the pre-delineated infarct masks will be revised and corrected, when necessary, by (at least) one out of three experienced neuroradiologists with more than 10 years of experience in the field.

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.

The protocol used for infarct annotation is the same protocol used for the ISLES22 challenge. The difference with ISLES'22 regards the challenge task. While in ISLES'22 participants predicted infarction from DWI input sequences, participants in ISLES'24 are asked to predict the (MRI-derived, post-treatment) infarct masks using pre-treatment, acute CT imaging.
Protocol definition for annotation of MRI infarctions: DWI-lesion segmentation of lesions at least two 2x2x2mm^3 voxels with definitive hypointensity in the ADC. Artefacts will not be segmented. To judge whether a lesion is true or just an artefact, the FLAIR sequence will additionally be reviewed.

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.

1) Algorithm pre-segmentation: We use a deep-learning ensemble model from leading ISLES'22 participants for pre-segmenting the entire database. The model has been devised using the train-set data from ISLES'22 and tested over the ISLES'22 test-set.
2) Manual correction of the segmentations: All segmentations provided in step 1) are visually inspected by an experienced neuroradiologist from University Hospital of Munich or University Hospital of Zurich. Segmentations are manually corrected in 2D, when necessary, using the software ITK-SNAP.

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.

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.

All images will be released anonymized and defaced as NIFTI files using the BIDS convention [9]. Data preprocessing steps consist of:
-Temporal resampling (1 frame/second) of the 4D CTP data.
- Deriving perfusion maps (CBF, CBV, MTT and Tmax) from the 4D CTP series using standard deconvolution analysis.
- CTA, NCTT, and CTP images will be linearly co-registered to the DWI image space.
- All images (CT-based, perfusion maps, and MRI) will be skull-stripped. For doing so, standard algorithms such as HD-BET [10] or SynthStrip [11] will be used to generate an MRI brain mask that will be later propagated over the remaining, co-registered CT images.

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 small brain ischemic lesions, there might be cases where an apparent lesion is doubtfully a true one. In those cases, a second neuroradiologist will also review the images. If there is still no consensus between them, the scans will be reviewed by the third neuroradiologist, and the apparent lesion will be labeled by majority-voting.

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

None.

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)

Voxel-wise metrics:
- Segmentation: Dice Similarity Coefficient.
- Volume: Absolute difference.
Lesion-wise metrics:
- Detection: F1-Score
- Count: Absolute lesion difference

Metrics are defined as follows:
-Dice Similarity Coefficient (voxelwise) = Detection F1 (elementwise) = 2TP/ (2TP + FN + FP)
-Absolute volume difference (voxelwise) = Absolute lesion count difference (elementwise) = abs(Total predicted - Total rater).

The lesion count in the annotated/predicted images is obtained by computing the number of connected components per case. We will release a Python script that computes all these metrics when releasing the first batch of image data.

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

The metrics considered for this task are mainly chosen from this challenge's technical and clinical motivations. We aim to evaluate the accuracy of the algorithms to localize and detect most of the infarct lesions. From this perspective, typical segmentation metrics (e.g. Dice Coefficient) may not be sufficient, as small lesions might be neglected (not driving relevant changes in Dice terms) under the presence of large ones [12]. Thus, we consider metrics that are often of relevance for neuroradiologists, such as the final infarct volume (measured in terms of absolute differences), the presence/absence of a lesion (i.e, detection with F1-Score) and the accurate count of the lesion burden (absolute lesion count difference). Besides, the inclusion of classical segmentation metrics, such as Dice Similarity Coefficient, are used to globally measure the performance overlap between ground truth and predictions.

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 ranking scheme considered for ISLES'2024 is the same one used in our previous ISLES editions [1-4], and provides fairness and transparency for the teams. The strategy received positive feedback from the research community and has been widely adopted for other medical image challenges.

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

Missing results on test cases will result in ranks for the corresponding metrics to be set to the worst possible value.

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

Similarly, as for previous editions of the ISLES challenge, ranking will be produced using a "rank then aggregate" approach. In a nutshell, it consists in comparing each metric at the case level. Metrics are calculated for each case, followed by establishing metric-specific ranks separately for each dataset. A mean rank over all metrics is then obtained to obtain the team's rank for each case. The final teams' rank is the mean over all case-specific ranks.

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.

-A 1000 bootstraps ranking will be obtained using challengeR [13].
-T-test or Wilcoxon (for non-uniformly distributed data) will be used to check for statistically significant differences among submissions. After inspecting the data distribution, the choice of a parametric or non-parametric test will be decided.

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

We will use paired tests in order to evaluate algorithm-rater and inter-algorithm differences for the considered metrics. We also aim to identify statistically outperforming models.

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 provide inter-rater variability for the considered metrics. For the evaluation and visualization of the different algorithms performance, we will make use of the challengeR library (e.g [13]). An ensemble using the

top-leading solutions will be devised, and its performance will be compared against single algorithmic solutions.

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. Maier, Oskar, et al. "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI." Medical image analysis 35 (2017): 250-269.
2. Winzeck, Stefan, et al. "ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI." Frontiers in neurology 9 (2018): 679.
3. Hakim, Arsany, et al. "Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: Lessons from the ISLES challenge." Stroke 52.7 (2021): 2328-2337.
4. Hernandez Petzsche, Moritz R., et al. "ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset." Scientific data 9.1 (2022): 762.
5. Liew, Sook-Lei, et al. "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms." Scientific data 9.1 (2022): 320.
6. Iglesias, Juan E., et al. "SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry." Science advances 9.5 (2023): eadd3607.
7. Liew, Sook-Lei, et al. "Association of Brain Age, Lesion Volume, and Functional Outcome in Patients With Stroke." Neurology 100.20 (2023): e2103-e2113.
8. Ashtari, Pooya, et al. "Factorizer: A scalable interpretable approach to context modeling for medical image segmentation." Medical image analysis 84 (2023): 102706.
9. Gorgolewski, Krzysztof J., et al. "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments." Scientific data 3.1 (2016): 1-9.
10. Isensee, Fabian, et al. "Automated brain extraction of multisequence MRI using artificial neural networks." Human brain mapping 40.17 (2019): 4952-4964.
11. Hoopes, Andrew, et al. "SynthStrip: Skull-stripping for any brain image." NeuroImage 260 (2022): 119474.
12. Maier-Hein, L., et al. "Metrics reloaded: Recommendations for image analysis validation. arXiv 2023." arXiv preprint arXiv:2206.01653.
13. Wiesenfarth, Manuel, et al. "Methods and open-source toolkit for analyzing and visualizing challenge results." Scientific reports 11.1 (2021): 2369.
14. Albers, Gregory W., et al. "Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging." New England Journal of Medicine 378.8 (2018): 708-718.
15. Lin, Longting, et al. "Whole-brain CT perfusion to quantify acute ischemic penumbra and core." Radiology 279.3 (2016): 876-887.
16. Fahmi, F., et al. "Differences in CT perfusion summary maps for patients with acute ischemic stroke generated by 2 software packages." American journal of neuroradiology 33.11 (2012): 2074-2080.
17. Robben, David, et al. "Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning." Medical image analysis 59 (2020): 101589.

Further comments

Further comments from the organizers.

Several members of the organization committee have participated in the organization of other MICCAI challenges (ISLES, Brats, Verse, among others) and workshops (BrainLes, iMIMIC, SWITCH, etc.) We finally would like to thank the reviewers for taking the time to evaluate the ISLES 2024 proposal.