Multi-class Bi-atrial Segmentation from 3D Contrast-Enhanced Magnetic Resonance Imaging: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Multi-class Bi-atrial Segmentation from 3D Contrast-Enhanced Magnetic Resonance Imaging

Challenge acronym

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

Multi-class bi-atrial segmentation challenge (MBAS)

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.

Persistent or long-standing persistent atrial fibrillation (AF) is associated with progressive atrial structural remodelling, including chamber dilatation, fibrosis and atrial wall thickness variation. Left atrial (LA) fibrosis quantified with late gadolinium-enhanced (LGE) magnetic resonance imaging (MRI) or estimated via low voltage map has been used to guide adjunctive ablation beyond pulmonary vein isolation in recent clinical trials with mixed clinical outcomes. Many factors may contribute to this, such as the absence of right atrial (RA) fibrosis as a potential ablation target in these trials. Clinical studies found that a majority of patients with persistent AF have AF drivers in the RA. In addition, the reconstructed bi-atrial chambers can be used to guide ablations and extract key structural biomarkers, such as the LA and RA volume and wall thickness for patient stratification and targeted treatment alongside fibrosis.
Expanded from our previous LA challenge in 2018 [Xiong et al., Medical Image Analysis, 2021, with 185 citations so far], this new challenge aims to test multi-class machine learning approaches for both LA and RA cavity and atrial wall directly from LGE-MRIs to improve targeted AF ablation. In addition, the proposed methods will be tested for their performances across segmentation and biomarker extraction tasks (such as the LA/RA volume and fibrosis) on cross-center LGEMRI datasets. We will use 200 multi-centre 3D LGE-MRIs for this challenge, the largest one in the field so far. More importantly, the precious data were carefully labelled in consensus with three independent experts for each LGE-MRI scan to obtain one segmentation per scan.
The developed AI approaches and clinical pipelines may be transferrable to other challenging medical tasks. The proposed challenge and data are a paradigm shift for cardiac structural analysis and may accelerate the search for optimised ablation strategies for patients with persistent and long-standing persistent AF.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Atrial fibrillation, contrast-enhanced MRI, gadolinium, atrial chamber, fibrosis, multi-class machine learning, medical images

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.

STACOM

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.

The proposed medical image data and challenge will be interesting to a broader research community. In our previous challenge for LA segmentation in 2018, we attracted 28 teams with about 100 researchers involved [Xiong et al., 2021]. While that may make it difficult to predict, we estimate that due to the importance of the topic, and general interest, there could be 30+ or 40 teams in this challenge.

Publication and future plans

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

Publication: A manuscript summarising the design, implementation, and results of our Benchmark will be submitted for publication following MICCAI 2024. Participating teams may have representatives as co-authors. For this purpose, we will consider MICCAI's recommended publication mechanism and journal.

Future plans: At completion of MICCAI 2024, we will make the training dataset publicly available through our GitHub repositories permanently to allow continuous submissions and impacts to help developers to benchmark the performance of their AI approaches.

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 MBAS Challenge will be conducted online in the cloud using Codalab Competition platform (https://codalab.lisn.fr/). Codalab will provide expertise in hosting the challenge.

TASK 1: Segment both RA and LA cavities and atrial walls directly from LGE-MRIs

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.

Expanded from our previous left atrium (LA) challenge in 2018, this new challenge aims to test multi-class machine learning approaches for both LA and right atrium (RA) cavity and atrial wall directly from the late gadolinium-enhanced (LGE)-MRIs to improve targeted ablation for patients with atrial fibrillation. In addition, the developed methods will be tested for their performances across segmentation and biomarker extraction tasks (such as the LA/RA volume and fibrosis) on cross-center LGEMRI datasets. We will use 200 multi-centre 3D LGE-MRIs for this challenge, the largest one in the field so far. More importantly, the precious data were carefully labelled in consensus with three independent experts for each LGE-MRI scan to obtain one segmentation per scan. The developed AI approaches and clinical pipelines may be transferrable to other challenging medical tasks. The proposed challenge and data are a paradigm shift for cardiac structural analysis. They may accelerate the search for optimised ablation strategies for patients with persistent and long-standing persistent AF. We plan to publish a challenge report and keep training data and labels free and available permanently on our GitHub website following MICCAI 2024.

Keywords

List the primary keywords that characterize the task.

Atrial fibrillation, contrast-enhanced MRI, gadolinium, atrial chamber, fibrosis, multi-class machine learning, medical images

ORGANIZATION

Organizers

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

Associate Professor Jichao Zhao (Auckland Bioengineering Institute);
Fangqiang Xu (Auckland Bioengineering Institute);
Fan Feng (Auckland Bioengineering Institute)

b) Provide information on the primary contact person.

Jichao Zhao (j.zhao@auckland.ac.nz)

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

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.

Codalab Competition platform (https://codalab.lisn.fr/)

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

We will develop our GitHub challenge website upon acceptance of this proposal.

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 data allowed, so it will be fair for all participants since we focus on algorithm development (supervised deep learning) and comparison in the challenge. Otherwise, it will disadvantage participants with no knowledge or access to private LGE-MRI training data.

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.

We will provide awards for the top 3 winners, and participants in the challenge will be represented as co-authors on the challenge report to be submitted for peer-review publication. We will provide prizes: First place: $500; Second place: $300; and Third place: $200.

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.

We will publish the complete ranking list of all participants online via our GitHub challenge website and invite the top five ranked methods to be presented at the STACOM 2024 challenge session. In addition, organisers of the challenge will also summarise the challenge and key results in the STACOM workshop.

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).

A benchmark manuscript, developed by the organising team, will be drafted soon after the completion of MICCAI 2024. Each of the teams participating in the challenge may have 2-4 key members (e.g., lead approach developer and senior team member) participate as co-authors in the report. Participating teams may publish their own results separately, but only after an embargo period of nine months, to allow challenge organisers to publish their report first. Participants will be required to submit an abstract, describing their approach and results from the training and validation phases, due after the validation phase (as the STACOM workshop paper). The paper can be updated at the end of the challenge, which will help the organisers draft the challenge publication that will include a summary of each approach. All challenge policies will be stated on the GitHub challenge website, and participants must agree to the terms of the challenge before they can participate.

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.

The submission method for validation and test phases will be through a docker submission of algorithm to the Codalab platform. Instructions will be provided in the GitHub challenge website (URL and content to be established upon acceptance of this proposal).

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.

Each participant/team may submit up to a maximum of five results during the test phase of the challenge. Only the last run will officially count as their final challenge result. Each participant can only be in one team.

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 April, 2024: Challenge website open begins
15 April, 2024: Challenge data are released, and the training phase begins
15 June, 2024: The training phase begins
28 June, 2024: Workshop abstract submission deadline
15 July, 2024: Challenge testing phase begins
31 July, 2024: Workshop paper submission
15 August, 2024: Challenge ends
22 August, 2024: Workshop camera ready
6 October 2024: Winner announcement during the STACOM workshop

All deadlines are at 11:59 PM UTC on the corresponding day unless otherwise noted. The competition organisers reserve the right to update the contest timeline if they deem it necessary.

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).

The de-identified LGE-MRI data from each clinical centre already had their ethics approval. The proposed challenge focused on the approach development and utilised the existing de-identified dataset. Therefore, no additional ethics is needed.

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 code will be a open source and available in our GitHub challenge website.

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

The participating teams will be encouraged to provide their code as open-source, or at a minimum, as dockers.

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 organisers declare no conflicts of interest with regards to this challenge. We have no confirmed sponsors for the proposed challenge yet. All three organisers have 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

Research, Intervention assistance

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 represents multi-centre LGE-MRIs from both male and female subjects with atrial fibrillation, originally acquired for research protocols related to patient screening, ablation treatment planning, and therapy response assessment.

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

The challenge LGE-MRI cohort was acquired from patients with atrial fibrillation, with all the attributed noted in the target cohort (above).

Imaging modality(ies)

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

Patients with atrial fibrillation had pre-ablation or post-ablation LGE-MRI scans (3-13 months) at a spatial resolution of 1.25×1.25×2.5 mm3 to evaluate atrial fibrosis/scars. Approximately 20-25 minutes after injection of gadolinium contrast, high-resolution LGE-MRIs of bi-atrial chambers were acquired.

Context information

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

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

Original LGE-MRI data were acquired for purposes of atrial fibrosis/scar screening detection, therapy guidance, and response assessment.

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

Subjects include patients with atrial fibrillation. Those patient demographics, gender, and medical history were very diverse, though almost all patients were relative old (>50 years old). However, this was not critical to the purpose of our challenge, which is to address atrial segmentation.

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 original LGE-MRI datasets (target/challenge cohort) used in the challenge were in the DICOM format, obtained from different clinical centres (such as Waikato Hospital, New Zealand and The University of Utah). These datasets included the heart and a part of body torso from patients with atrial fibrillation.

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 participating algorithms will be required to focus on segmenting different atrial components (both RA and RA cavities, atrial wall and fibrosis) from the raw LGE-MRIs. These will be described in detail through the challenge instructions to the participating teams.

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

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 LGE-MRI challenge data were obtained at various clinical centres using either a 1.5 Tesla Avanto or 3.0 Tesla Verio whole-body MRI scanner (Siemens Medical Solutions, Erlangen, Germany). The imaging protocol was developed originally by the University of Utah and remained the same across the clinical centres.

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

The LGE-MRI imaging protocol was the same among the clinical centres. All patients underwent LGE-MRI scanning to define the atrial structure and fibrosis distribution prior to and post-ablation treatment (McGann et al., 2014; McGann et al., 2011). The clinical images were acquired with either a 1.5 or 3.0 Tesla whole-body MRI scanner. High-resolution LGE-MRIs of bi-atrial chambers were acquired approximately 20 to 25 min after the injection of 0.1 mmol/kg gadolinium contrast (Multihance, Bracco Diagnostics Inc., Princeton, NJ) using a 3D respiratory

navigated, inversion recovery prepared gradient echo pulse sequence. An inversion pulse was applied every heartbeat, and fat saturation was applied immediately before data acquisition. To preserve magnetisation in the image volume, the navigator was acquired immediately after the data acquisition block. Typical scan times for the LGEMRI study were between 8 and 15 min at 1.5 T and 6 to 11 min using the 3T scanner (for Siemens sequences) depending on patient respiration (McGann et al., 2014; McGann et al., 2011).

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 LGE-MRI data come from a wide range of clinical studies performed at institutions/hospitals across the US, China, Japan and New Zealand.

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).

Original clinical imaging data had been acquired by clinicians for assessment of patients with atrial fibrillation prior to and post ablation treatment at major research clinical centers. These studies had been referred by cardiogists.

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 the challenge, a case refers to one 3D LGE-MRI image sequence (series of 2D images) performed using a well established LGE imaging protocol (see previously) for patients with atrial fibrillation for stratification, treatment guidance and outcomes assessment. Studies are stored in the DICOM standard.

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

Training cases = 70
Validation cases = 30
Test cases = 100

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

A LGE-MRI dataset from one single centre (100 3D LGE-MRIs) was chosen for the training and validation dataset that is used for fine tuning of algorithms. Another cohort of LGE-MRI data from different clinical centres (100 3D LGE-MRIs) was chosen to test the developed approaches on a more diverse dataset, mimicking real world application. The challenge design reflects reality facing most of applications.

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.

We will ensure that distribution of cases across the training, validation, and test are balanced as far as possible by analysing the characteristics of the dataset. For instance, the LGE-MRI dataset contained a range of imaging qualities, with the differences mostly being attributed to varying patient characteristics and magnetic fields. In order to explore the quality of the dataset, the signal to noise ratio (SNR), contrast ratio (CR), and heterogeneity (HET) between the foreground containing the atria and the background was assessed. The three metrics used were in agreement as SNR had a strong positive correlation with CR and HET and while CR and HET had a strong negative correlation. Distributions of the quality measurements on all data showed that less than 15% of the data was of high quality (SNR > 3), 70% of the data was of medium quality (SNR = 1 to 3), and over 15% of the data was of low quality (SNR < 1). We will split the LGE-MRI data evenly across the training, validation, and test in terms of imaging qualities.

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, three

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 LA and RA cavities and bi-atrial walls were manually segmented in consensus with three trained observers for each LGE-MRI scan to obtain one segmentation per scan. The atrial cavities were manually segmented, and the atrial endocardial surfaces were traced. Then cavities were morphologically dilated to include atrial walls. The four pulmonary veins (PVs) were included in the LA segmentation, but were limited to the PV antrum region. The mitral and tricuspid valves were delineated by a 3D plane to create a smooth linear surface and separate the atrial and ventricular chambers. Due to the complex atrial anatomy (e.g., PVs and valves), manual editing was also required to obtain the outer boundary of the wall. In addition, the atrial septum was manually traced to connect the walls of the LA and RA. The atrial fibrosis ground truths were extracted using an adaptive threshold approach for each slice of the manually segmented 3D LGE-MRI to mask out high-intensity pixels. This was performed separately for the LA and RA. The thresholds for each image slice were computed as a set number of standard deviations above the mean pixel intensity of the non-fibrotic tissue.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

The lead annotator for the proposed challenge dataset had more than four years of experience in segmenting 3D LGE-MRIs. He was aided by several senior colleagues who had many years of experience in LGE-MRIs.

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.

There was only one set of annotations, as the nature of this challenge does not require multiple annotations.

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.

There was no pre-processing method for data used in this challenge.

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.

A potential source of error could have been the segmentation of atrial chambers in the challenge dataset. The magnitude of this error was hard to estimate given it was impossible to obtain the absolute ground truths for the clinical data. However, given the experience of the team in the LGE-MRI field, the error was estimated very small (<0.1% in Dice scores), much less than the difference in Dice scores for ranking the teams.

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)

Both Dice Similarity Coefficient (DSC) and Hausdorff distance (95%) (HD95) will be used to computing the ranking.

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

In our analysis for the 2018 LA segmentation challenge [Xiong et al. 2021], we found that a high level of agreement between the main technical measures of DSC of the LA diameter and volume error showed that the top-performing algorithms were capable of producing anatomically accurate segmentation, which is highly important in clinical applications. The most widely used segmentation metrics: the Dice score and Jaccard Index/IoU produced fairly consistent rankings amongst the top 5 contestants in our 2018 LA chanllege. However, the sensitivity, and specificity produced significantly different rankings in comparison. A potential explanation of this discrepancy is that the sensitivity and specificity do not consider both positive and negative pixels simultaneously, leading to biased measurements. This also likely explains why these metrics are not commonly used for evaluating segmentation accuracies.
Furthermore, Hausdorff Distance (95%) (HD95) evaluatives model by pinpointing the maximum deviation between the prediction result and the ground truth, which highlights key differences and emphasizes the importance of boundary accuracy in segmentation tasks. Consequently, given the complementary strengths of these metrics in capturing both overall performance and boundary precision, we employ both DSC and HD95 in the upcoming challenge. It aims to foster a more robust and comprehensive assessment of segmentation performance, catering

to the detailed requirements of clinical application.

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.

Ranking will be based on the output labels of the submitted approach from each team for the 100 3D multi centre LGE-MRI testing data evaluated using DSC and HD95 (with same weighting).

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

A missing result will count as a zero score.

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

For segmentation tasks, DSC and HD95 are widely used to evaluate the performance of each approach. The two new published sources are used as metric selection guides:
https://arxiv.org/abs/2206.01653
https://metrics-reloaded.dkfz.de/

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.

For the statistical methods used in the scope of the challenge analysis, regression plots, Bland Altman plots and bias will be used to measure the pixel-wise accuracy. Statistical significance was measured using the two-tailed t-test. In our previous analysis, we found the most widely used segmentation metrics: the Dice score, Jaccard Index/IoU produced fairly consistent rankings. However, Hausdorff Distance uniquely influenced rankings by accurately measuring boundary differences. Considering these insights and the comprehensive evaluation strengths of DSC and HD95, we will utilize both metrics for performance evaluation in this new challenge.

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

Our selection of these statistical methods is driven by their ability to assess outcomes from diverse perspectives, coupled with their extensive application in the medical domain.

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.

Ranking variability will be based on the categorical differences described above.

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. 2018 LA Segmentation Challenge: https://www.cardiacatlas.org/atriaseg2018-challenge/
2. Z Xiong, A Nalar, K Jamart, M Stiles, V Fedorov, J Zhao. Fully automatic 3D bi-atria segmentation from late gadolinium-enhanced MRIs using double convolutional neural networks. Lecture Notes in Computer Science, 11395 LNCS, 63-71, 2020.
3. Z Xiong, Q Xia, Z Hu, N Huang, C Bian, S Vesal, N Ravikumar, A Maier, X Yang, PA Heng, D Ni, C Li, Q Tong, W Si, E Puybareau, Y Khoudli, T Géraud, C Chen, W Bai, D Rueckert, L Xu, X Zhuang, X Luo, S Jia, M Sermesant, Y Liu, K Wang, D Borra, A Masci, C Corsi, C Vente, M Veta, R Karim,C Preetha, S Engelhardt, M Qiao, Y Wang, Q Tao, M Nuñez-Garcia, O Camara, N Savioli, P Lamata, J Zhao. A global benchmark of algorithms for segmenting late gadolinium-enhanced cardiac MRI. Medical Image Analysis, 67(101832), 2021.
4. C McGann, E Kholmovski, J Blauer, S Vijayakumar, THaslam, J Cates, E DiBella, N Burgon, B Wilson, A Alexander, M Prastawa, M Daccarett, G Vergara, N Akoum, D Parker, R MacLeod, N Marrouche. Dark regions of no-reflow on late gadolinium enhancement magnetic resonance imaging result in scar formation after atrial fibrillation ablation. J Am Coll Cardiol. 5;58(2):177-85, 2011.
5. C McGann, N Akoum, A Patel, E Kholmovski, P Revelo, K Damal, B Wilson, J Cates, A Harrison, R Ranjan, N Burgon, T Greene, D Kim, E Dibella, D Parker, R Macleod, N Marrouche. Atrial fibrillation ablation outcome is predicted by left atrial remodeling on MRI. Circ Arrhythm Electrophysiol. 7(1):23-30, 2014.
6. https://arxiv.org/abs/2206.01653.7
7. https://metrics-reloaded.dkfz.de/

Further comments

Further comments from the organizers.

N/A