The Federated Tumor Segmentation (FeTS) Challenge 2024: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

The Federated Tumor Segmentation (FeTS) Challenge 2024

Challenge acronym

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

FeTS 2024

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.

International challenges have become the standard for validation of biomedical image analysis methods. We argue, though, that the actual performance even of the winning algorithms on ``real-world`` clinical data often remains unclear, as the data included in these challenges are usually acquired in very controlled settings at few institutions. The seemingly obvious solution of just collecting increasingly more data from more institutions in such challenges does not scale well due to privacy and ownership hurdles.

We build upon the first-ever proposed federated learning challenge, Federated Tumor Segmentation (FeTS) 2021 and its follow-up in 2022, to introduce FeTS 2024, intending to address these hurdles, for the creation of tumor segmentation models. Specifically, the FeTS 2024 challenge will use clinically acquired, multi-institutional multi-parametric magnetic resonance imaging (mpMRI) scans from the RSNA-ASNR-MICCAI BraTS 2021 challenge, and BraTS 2023 Glioma challenge, as well as from various remote independent institutions included in the collaborative network of a real-world federation (www.fets.ai). The FeTS 2024 challenge focuses on innovating at the level of federated aggregation where locally trained models combine to form the consensus model for the segmentation of intrinsically heterogeneous (in appearance, shape, and histology) brain tumors, namely gliomas (and particularly the radiographically appearing glioblastomas). Compared to the BraTS challenge [1-4], the ultimate goal of FeTS is the creation of a consensus segmentation model that has gained knowledge from data of multiple institutions without pooling their data together (i.e., by retaining the data within each institution).

Since the conception of FeTS 2021 and FeTS 2022, Federated Learning has matured to a more active research field in biomedical AI. What separates FeTS 2024 challenge from those of previous years is that while FeTS 2021 and FeTS 2022 focused on semantic segmentation, now focuses on instance segmentation which in line with the findings of [11-12] stands superior due to the added value of evaluating multiple individual tumors per patient.

This year we plan to further broaden the final evaluation of the aggregation methods, by testing their

generalizability beyond segmentation tasks to a hidden (to the participants) task. This added evaluation will be a significant part of the final challenge paper which will provide detailed meta-analysis and inform us further insights about the developed aggregation methods. It will also inform further expansion of the tasks planned for next year's challenge proposal. For fairness, since the participants will only have access to the segmentation data, this added evaluation on a new task will neither be considered for the monetary awards, nor for the ranking. We will however announce performance on this hidden task during the challenge results presentation at MICCAI.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Federated Learning, Segmentation, Brain Tumors, Cancer, Collaborative Learning, Challenge

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.

We are in coordination with Workshop on Distributed, Collaborative, and Federated Learning (DeCaF), that a member of the organizing committee (Spyridon Bakas) is also an organizer and coordinated for the integration of the challenge into a one-hour time slot within the workshop's program.

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 50 submissions this year. FeTS 2021 received 16 submissions and FeTS 2022 received 32, and, after the absence of a FeTS challenge during 2023, there has been public demand for its return.
Taking into consideration the above, as well as the points raised below, we think that 50 participating teams is a conservative estimation.
i) the increasing interest for FL in medical imaging, which has now matured to a more active research field
ii) the value of FL in healthcare in lowering the barrier to accessing large-scale datasets while avoiding bureaucratic issues, as well as legal, privacy, and ownership concerns,
iii) the continuously increasing number of teams participating in the BraTS challenge during the past 10 years (2012:n=10, 2013: n=10, 2014: n=10, 2015: n=12, 2016: n=19, 2017: n=53, 2018: n=63, 2019: n=72, 2020: n=78, 2021: n=2,300),
iv) since last year's BraTS and the publication of our related FL manuscripts, we have received >50 requests for revealing the data contributions from the individual institutions to the BraTS 2021 data, and we expect these users to be interested in participating in the FeTS 2024 challenge.

In addition, we will advertise the event in related mailing lists (e.g., CVML; visionlist@visionscience.com; cvnet@mail.ewind.com; MIPS@LISTSERV.CC.EMORY.EDU) and we intend to send an email to all the above and notify them about this year's challenge.

Publication and future plans

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

We intend to coordinate 2 publication plans immediately after the challenge.
Plan 1: The configuration of combining the FeTS challenge with a workshop provides the FeTS participants with the option to publish their methods in the workshop's LNCS (Lecture Notes in Computer Science) post-conference proceedings. We have already performed this configuration for FeTS 2021 and FeTS 2022, and these papers are now with Springer.
Plan 2: Following up on the manuscript summarizing the results of FeTS 2021 and FeTS 2022 (which we have already submitted) we will focus on a comprehensive meta-analysis to inform the community about the obtained results of 2024 as well, emphasizing the value of instance segmentation compared to the semantic segmentation of previous years with a focus on innovative aggregation methods.

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.

Hardware requirements in case of an in-person meeting: 1 projector, 3 microphones, loudspeakers
FeTS is an off-site challenge and algorithms are run using the participants' computing infrastructure during the training and validation phase, followed by using the organizers' local infrastructure, as well as the infrastructure of the independent institutions involved in the www.fets.ai federation, during the testing phase. FL typically involves real-time and bandwidth restrictions. Participants will be provided clear rules on how we will simulate these compute restrictions such that no advantage is gained by using particular equipment.

TASK 1: Innovative Federated Learning Aggregation Methods

SUMMARY

Abstract

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

The specific focus of this task is to identify the best way to aggregate the knowledge coming from segmentation models trained on individual institutions, instead of identifying the best segmentation method. More precisely, the focus is on the methodological portions specific to federated learning (e.g., aggregation, client selection, training- per-round, compression, communication efficiency), and not on the development of segmentation algorithms (which has been the focus of the BraTS challenges). To facilitate this, an existing infrastructure for federated tumor segmentation using federated averaging will be provided to all participants indicating the exact places that the participants are allowed and expected to make changes. To prepare for this aggregation task, the participants will be provided with information on the data origin and acquisition protocols during the training phase of the challenge. The primary objective of this task is to develop methods for effective aggregation of local instance segmentation models, given the partitioning of the data into their real-world distribution. As an optional sub-task, participants will be asked to account for network communication outages, i.e., dealing with stragglers.

As previously mentioned, we'll also broaden the final evaluation of the aggregation methods,
by testing their generalizability beyond segmentation tasks to a hidden (to the participants) classification task.
The results will be included in our meta-analysis paper but will not be considered for the monetary awards, nor for the ranking.
We will announce performance on this hidden task during the challenge results presentation at MICCAI.

Keywords

List the primary keywords that characterize the task.

Federated Learning, Segmentation, Brain Tumors, Cancer, Collaborative Learning, Challenge

ORGANIZATION

Organizers

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

Spyridon Bakas
Akis Linardos
Sarthak Pati
Ujjwal Baid
IUPUI, USA

Micah Sheller
Brandon Edwards

Intel Labs / MLCommons

Alexandros Karargyris
IHU Strasbourg / MLCommons

Peter Mattson
Google / MLCommons

Bjoern Menze
University of Zurich, Switzerland
Clinical Evaluators & Annotation Approvers:

Michel Bilello
Suyash Mohan
UPENN, USA

+60 ASNR member neuroradiologists for ground truth generation (listed in BraTS 2021 paper)

Data Contributors:
John B. Freymann & Justin S. Kirby - TCIA, NCI, National Institutes of Health (NIH), USA
Christos Davatzikos
CBICA, UPENN, USA

Hassan Fathallah-Shaykh
University of Alabama at Birmingham, USA

Roland Wiest
University of Bern, Switzerland

Andras Jakab
University of Debrecen, Hungary

Rivka R. Colen
University of Pittsburgh Medical Center

Aikaterini Kotrotsou
MD Anderson Cancer Center, TX, USA

Daniel Marcus & Mikhail Milchenko & Arash Nazeri
WUSM, USA

Marc-Andre Weber
Heidelberg University, Germany

Abhishek Mahajan
Tata Memorial Center, Ind

b) Provide information on the primary contact person.

Spyridon Bakas, Ph.D.
Indiana University School of Medicine, Indiana, USA
spbakas@iu.edu

Life cycle type

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

Examples:

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

Repeated event with fixed submission 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.

synapse.org
Following our successful collaboration with the synapse platform (SAGE Bionetworks) during the FeTS 2022 challenge, we will continue using syapse.org as the challenge platform.

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

https://miccai2024.fets.ai/ - (Website will be publicly visible after the challenge approval). This will be similar to what we used in previous years for FeTS 2021: https://miccai2021.fets.ai/ and FeTS 2022: https://miccai2022.fets.ai/

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.

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 organizers and their immediate groups cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. Since organizing institutions are large, other employees from other labs/departments may participate and should be eligible for the awards and to be listed in the official leaderboard.

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

We are in coordination with Intel to confirm sponsorship of monetary awards ($5K) for the top 3 teams.

e) Define the policy for result announcement.

Examples:

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

Top 3 performing methods will be announced publicly at the conference and the participants will be invited to present their method.

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).

The configuration of combining the FeTS challenge with the BrainLes workshop provides the FeTS participants with the option to publish their methods in the workshop's LNCS post-conference proceedings.
Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of FeTS 2024, making comparative assessment with the summary results of the previous FeTS challenges.

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 participants are required to send the output of their methods to the evaluation platform for the scoring to occur during the training and the validation phases. At the end of the validation phase, the participants are asked to identify the method they would like to evaluate in the final testing/ranking phase. To facilitate this ranking, the participants will be requested to submit to the organizers:
1. their aggregation algorithm, and
2. the weights of the selected model.
The organizers will then confirm the results that the participants will be reporting during the validation phase, and those that match will then be evaluated by the organizers in a hidden testing dataset. This will enable

confirmation of reproducibility and comparison with results obtained by algorithms trained on pooled datasets, as well as other FL aggregation algorithms, thereby maximizing the benefit towards solving the problem of federated learning for tumor segmentation.
We appreciate that this configuration might be considered restrictive to specific programming frameworks, and we are working on making this submission as inclusive as we can in terms of used programming languages or tools. Specifically, we are in open discussions with numerous developer teams of open-source FL libraries, towards writing a common set of instructions to enable the use of these frameworks by the challenge participants. Notably, Docker or other containerization technology will probably not be available due to IT restrictions in the federation of participating clinical sites.
During the training and validation phases, the participants will have the chance to test the functionality of their submission through the FeTS platform (which will provide the implementation of the evaluation metrics). The aggregation algorithm requirements, specifications, and examples will be described in: miccai2024.fets.ai. The organizers will provide to all participants a code package comprising a complete framework solution for federated training on the provided multi-institutional data. This code package will comprise 1) a challenge-specific adapter layer (utilizing the FeTS platform including OpenFL as the backend library), and 2) instructions on how to run a FL simulation on a single system.
Furthermore, specific instructions will be given to the participants on the parts/functions that they would need to alter the federated algorithm in the following ways:
1. The aggregation function used to fuse the collaborator model updates.
2. Which collaborators are chosen to train in each federated round.
3. The training parameters for each collaborator for each federated round.
Each of these functions will have a complete default implementation, such that participants can choose which aspects they want to explore. These functions do not relate to data loading, so participants who stay "in-bounds" should not have to worry about direct data leakage.
When participants submit their implementations, they will only include the implementations of the updated functions. They must not override any other parts of the base package at runtime. This will reduce the risk that they either accidentally or intentionally leak data across sites. It will also simplify manual inspection of submissions after the validation phase when all top-ranking participants (from the validation phase) will have their training reproduced (by us) using their code submission. Specifically, when reproducing a participant's results, we will start with a clean copy of the base package, then overwrite only these specific code files from that participant. This reproduction run will ensure only the specified functions were modified. In cases where the participant's reported results differ significantly* from our replicated results, we will investigate the cause, and if we find that the participant violated the rules of the competition, they may be disqualified. We reserve the right to determine an honest mistake was made and allow the participant to resubmit.
*we will determine this exact margin and include it in the rules for the participants. We may not be able to ensure completely deterministic runs on all platforms.

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

We intend to release a validation set of 219 cases, allowing participants to tune their methods in unseen data. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform will be allowed.

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: Registration opens & Training and Validation data is released
Participants will be able to register for the challenge in synapse.org, from April 1, 2024 until the short paper submission deadline at July 31, 2024. Data will also be made available for training (with ground truth labels) and validation (without ground truth labels).
- 1 July 2024: Submission of 1) short papers reporting method & preliminary results and 2) source code & aggregated model.*
*With regards to point 2, there is a possibility we will instead ask for containers instead of source code this year. We will update the synapse site accordingly in that case.
- 05 July - 12 August 2024: Evaluation on testing data (by the organizers - only for participants with submitted papers).
- 18 August 2024: Contacting top performing methods for preparing slides for oral presentation.
- 6-10 October 2024: Announcement of final top 3 ranked teams: Challenge at MICCAI
- 30 October 2024: Camera-ready submission of extended papers for inclusion in the associated workshop proceedings (incl. results on testing data)

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 use the RSNA-ASNR-MICCAI BraTS challenge data, the training and validation data of which has been released via The Cancer Imaging Archive (TCIA) of the National Institutes of Health (NIH), following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions).

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.
Additional comments: CC-BY, but if any of the non-TCIA contributors object to this license, the specific subset of the BraTS data will be released under a CC-BY-NC license.

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 metrics and the ranking code used during the whole challenge's lifecycle will be made available through the FeTS platform.

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

The participants are required to submit their aggregation algorithm implemented in PyTorch, together with their final submitted results, during the validation phase. Specific instructions for the aggregation algorithm specification will be provided at miccai2024.fets.ai
All participants of the challenge will be required to accept an agreement through the synapse.org website that participation in the testing phase will automatically mean that the organizers can make the submitted method publicly available through our challenge webpage.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

We coordinate with Intel to offer monetary awards.
Spyridon Bakas, Akis Linardos, Sarthak Pati, Ujjwal Baid, and the clinical evaluators will have access to the validation, and 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
-

Treatment planning, Training, Assistance, CAD, Intervention planning, Surgery, Diagnosis, Research

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.

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with mpMRI acquisition protocol including i) unenhanced and ii) contrast-enhanced T1-weighted, iii) T2- weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.
b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.
Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with mpMRI acquisition protocol including i) unenhanced and ii) contrast-enhanced T1-weighted, iii) T2- weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

MRI

Imaging modality(ies)

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

Partitioning of the data according to individual contributing institutions, and acquisition equipment.

Context information

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

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

Age, Gender.

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

Brain multi-parametric MRI scans.

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.

Tumor in the brain.

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.

Tumor in the brain.

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)).

Precision, Sensitivity, Specificity, Dice, Hausdorff 95% - per lesion (instance) segmentation
Additional points: Additional points: Find the optimal weight aggregation method for brain tumor segmentation in MRI images.

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 exact scanners and their technical specifications used for acquiring the TCIA cohort has been listed in the data reference published in the related Nature Scientific Data manuscript of ours [2]. Since then, multiple institutions have contributed data to the create the current RSNA-ASNR-MICCAI BraTS dataset and these are listed in the latest BraTS arxiv paper [3]. We have made the complete BraTS 2021 dataset permanently available through TCIA. All the acquisition details are included together with the data availability in TCIA.
[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, 4:170117 (2017) DOI: 10.1038/sdata.2017.117
[3] S.Bakas, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv:1811.02629

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

The acquisition protocol is different for each different institution as the scans we use are representative of real clinical protocols.
The exact scanners and their technical specifications used for acquiring the TCIA cohort has been listed in the data reference published in the related Nature Scientific Data manuscript of ours [2]. Since then, multiple institutions have contributed data to the create the current RSNA-ASNR-MICCAI BraTS dataset and these are listed in the latest BraTS arxiv paper [3]. We have made the complete BraTS 2021 dataset permanently available through TCIA. All the acquisition details are included together with the data availability in TCIA.
All sequences included for each case of our dataset represent the best image quality available in the acquiring institution for this particular case. There was no inclusion/exclusion criterion applied that related to 3d acquisitions, or (for example) MPRAGE. We instead accepted all types of T1 acquisitions (with the exception of T1 FLAIR, as we did not want to mix the fluid suppressed values with non-flair scans) and then we applied the harmonized preprocessing protocol we have been using in the BraTS dataset, across the complete dataset.

[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, 4:170117 (2017) DOI: 10.1038/sdata.2017.117
[3] S.Bakas, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv:1811.02629

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 provided data describe pre-operative multi-parametric MRI scans, acquired with different clinical protocols and various scanners from:
1) University of Pennsylvania (PA, USA),
2) University of Alabama at Birmingham (AL, USA),
3) Heidelberg University (Germany),
4) University of Bern (Switzerland),

5) University of Debrecen (Hungary),
6) Henry Ford Hospital (MI, USA),
7) University of California (CA, USA),
8) MD Anderson Cancer Center (TX, USA),
9) Emory University (GA, USA),
10) Mayo Clinic (MN, USA),
11) Thomas Jefferson University (PA, USA),
12) Duke University School of Medicine (NC, USA),
13) Saint Joseph Hospital and Medical Center (AZ, USA),
14) Case Western Reserve University (OH, USA),
15) University of North Carolina (NC, USA),
16) Fondazione IRCCS Instituto Neuroligico C. Besta, (Italy),
17) MD Anderson Cancer Center (TX, USA),
18) Washington University in St. Louis (MO, USA),
19) Tata Memorial Center (India),
20) Ivy Glioblastoma Atlas Project.
21) UCSF
22) Unity Health
23) University Hospital of Zurich
Note that data from institutions 6-16, and 20 are provided through The Cancer Imaging Archive (TCIA - http://www.cancerimagingarchive.net/), supported by the Cancer Imaging Program (CIP) of the National Cancer Institute (NCI) of the National Institutes of Health (NIH).

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).

People involved in MRI acquisition for suspected brain tumors during standard clinical practice.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case describes multi-parametric MRI scans for a single patient at a single timepoint. The exact scans included for one case are i) unenhanced and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

FeTS 2024 training, validation, and testing data will be leveraging the RSNA-ASNR-MICCAI BraTS 2021 data, augmented with metadata related to the institution the data originates from and acquisition protocols. We plan to use the same data as in FeTS 2022. The exact numbers of 2024:

Training data: 1251 patients
Validation data: 219 patients
Testing data: 570 patients

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

Based on availability. The data was split in these numbers between training, validation, and testing after considering the number of cases used as test cases in previous instances of BraTS and the fact that the organizers did not want to reveal ground truth labels of previous test cases, to avoid compromising ranking the participants.

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.

N/A

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.

Reference approved from at least 2 experienced neuroradiologists, following annotations from 60 clinical neuroradiologists (volunteers from ASNR)

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 data considered in the FeTS 2024 challenge is the BraTS 2023 glioma challenge data.
The annotation of these data followed a pre-defined clinically approved annotation protocol, which was provided to all clinical annotators, describing in detail instructions on what the segmentations of each tumor sub-region should describe (see below for the summary of the specific instructions).
The annotators were given the flexibility to use their tool of preference for making the annotations, and also follow either a complete manual annotation approach, or a hybrid approach where an automated approach is used to produce some initial annotations followed by their manual refinements.

Summary of specific instructions:
i) the farthest tumor extent including the edema (what is called the whole tumor), delineates the hyperintense regions with homogeneous signal on T2 & T2-FLAIR.
ii) the tumor core (including the enhancing, non-enhancing, and necrotic tumor) delineates regions of lower T2 signal.
iii) the enhancing tumor delineates the hyperintense signal of the T1-Gd, after excluding the vessels.

iv) the necrotic core outlines regions appearing dark in both T1 and T1-Gd images (denoting necrosis/cysts), and dark regions in T1-Gd and bright in T1 (e.g., hemorrhage).

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.

Each case was assigned to a pair of annotator-approver. Annotators spanned across various experience levels and clinical/academic ranks, while the approvers were the 2 experienced board-certified neuroradiologists (with >13 years of experience), listed in the Organizers section as clinical evaluators and annotation approvers. The annotators were given the flexibility to use their tool of preference for making the annotations, and also follow either a complete manual annotation approach, or a hybrid approach where an automated approach is used to produce some initial annotations followed by their manual refinements. Once the annotators were satisfied with the produced annotations, they were passing these to the corresponding approver. The approver is then responsible for signing off these annotations. Specifically, the approver would review the tumor annotations, in tandem with the corresponding MRI scans, and if the annotations were not of satisfactory quality they would be sent back to the annotators for further refinements. This iterative approach was followed for all cases, until their respective annotations reached satisfactory quality (according to the approver) for being publicly available and noted as final ground truth segmentation labels for these scans.

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.

No Aggregation.

Data pre-processing method(s)

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

The exact preprocessing pipeline applied to all the data considered in the FeTS 2024 challenge is identical with the one evaluated and followed by the BraTS 2017-2021, FeTS 2021, and FeTS 2022 challenges. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format [5], we first perform a re-orientation of all input scans (T1, T1-Gd, T2, T2-FLAIR) to the LPS/RAI orientation, and then register all of them to the same anatomical atlas (i.e., SRI-24 [6]) and interpolating to the same resolution as this atlas (1 mm^3). The exact registration process comprises the following steps:
N4 Bias field correction (notably the application of N4 bias field correction is a temporary step. Taking into consideration we have previously [2] shown that use of non-parametric, non-uniform intensity normalization (i.e., N4) to correct for intensity non-uniformities caused by the inhomogeneity of the scanner's magnetic field during image acquisition obliterates the MRI signal relating to the abnormal/tumor regions, we intentionally use N4 bias field correction in the preprocessing pipeline to facilitate a more optimal rigid registration across the difference MRI sequences. However, after obtaining the related information (i.e., transformation matrices), we discard the bias field corrected scans, and we apply this transformation matrix towards the final co-registered output images used in the challenge).
Rigid Registration of T1, T2, T2-FLAIR to the T1-Gd scan, and obtain the corresponding transformation matrix.
Rigid Registration of T1-Gd scan to the SRI-24 atlas [6], and obtain the corresponding transformation matrix.
Join the obtained transformation matrices and applying aggregated transformation to the LPS-oriented scans.

After completion of the registration process, we perform brain extraction to remove any apparent non-brain tissue (e.g., neck fat, skull, eyeballs) based on a deep-learning approach we developed in house, focusing on scans with apparent diffuse gliomas and exhaustively evaluated it in both private and public multi-institutional data [7]. We then manually assessed all scans for confirming the correct brain extraction (i.e., skull stripping), where the complete brain region is included, and all non-brain tissue is excluded.
This whole pipeline, and its source code are available through the FeTS platform (https://github.com/FETS-AI/Front-End).
[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, 4:170117, 2017. DOI: 10.1038/sdata.2017.117
[5] R.Cox, J.Ashburner, H.Breman, K.Fissell, C.Haselgrove, C.Holmes, J.Lancaster, D.Rex, S.Smith, J.Woodward, A (Sort of) new image data format standard: NIfTI-1: WE 150, Neuroimage, 22, 2004.
[6] T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.
[7] S.Thakur, J.Doshi, S.Pati, S.Rathore, C.Sako, M.Bilello, S.M.Ha, G.Shukla, A.Flanders, A.Kotrotsou, M.Milchenko, S.Liem, G.S.Alexander, J.Lombardo, J.D.Palmer, P.LaMontagne, A.Nazeri, S.Talbar, U.Kulkarni, D.Marcus, R.Colen, C.Davatzikos, G.Erus, S.Bakas, Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-institutional Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training, NeuroImage, 220: 117081, 2020. DOI: 10.1016/j.neuroimage.2020.117081

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.

Study and evaluation of the effect of this error is addressed by the uncertainty task of BraTS 2019-2020 (i.e., to quantify the uncertainty in the tumor segmentations) and is outside the scope of the FeTS challenge.

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

N/A

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), 95% Hausdorff distance (HD), - per lesion evaluation.
Communication/Budget time provided to the participants, defined as the product of bytes sent/received multiplied by the number of federated rounds.
Sensitivity, Specificity, Precision

The regions evaluated using these metrics describe the whole tumor, the tumor core, and the enhancing tumor. Note that the tumor core includes the part of the tumor that is typically resected (i.e., enhancing, non-enhancing, and necrotic tumor), and the whole tumor describes all tumor sub-regions (i.e., tumor core and edema/invasion).

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

In terms of the assessed and evaluated tumor sub-regions:
i) the enhancing tumor describes the regions of active tumor and based on this, clinical practice characterizes the extent of resection.
ii) the tumor core (including necrosis) describes what is typically resected during a surgical procedure.
iii) the whole tumor as it defines the whole extent of the tumor (radiographically defined by the abnormal hyperintense
T2-FLAIR signal), including the peritumoral edematous tissue and highly infiltrated area.
In terms of evaluation metrics we use:
i) the Dice Similarity Coefficient, which is commonly used in the assessment of segmentation performance,
ii) the 95% Hausdorff distance as opposed to standard HD, in order to avoid outliers having too much weight,
iii) The budget will capture the communication cost for each algorithm
iv) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or undersegment.

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.

For ranking of multidimensional outcomes (or metrics), for each team, we will compute the summation of their ranks across the average of the 7 metrics described above as a univariate overall summary measure. This measure will decide the overall ranking for each specific team. To visualize the results in an intuitive fashion, we propose to visualize the outcome via an augmented version of radar plot [8].
Notably, since all teams are ranked per patient, whereas the communication cost is only accounted once for the complete training phase, the communication cost will be weighted according to the number of testing subjects in order to give it equal importance to the quality of the tumor segmentations.
[8] Duan R, Tong J, Lin L, Levine LD, Sammel MD, Stoddard J, Li T, Schmid CH, Chu H, Chen Y. PALM:
Patientcentered
Treatment Ranking via Large-scale Multivariate Network Meta-analysis. medRxiv. 2020 Jan 1

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

If an algorithm fails to produce a result metric for a specific test case, this metric will be set to its worst possible value (0 for the DSC and the image diagonal for the HD).

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

Following discussions with the biostatisticians involved in the design of this challenge (Dr Kun Huang - Chair of Dept of Biostatistics, IU), and also while considering transparency and fairness to the participants.

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.

Similarly to BraTS 2021, uncertainties in rankings will be assessed using permutational analyses [3]. Performance for the segmentation task will be assessed based on relative performance of each team on each tumor tissue class and for each segmentation measure, while the temporal element of the algorithmic convergence during training of the consensus model will be taken into account independently with the related weighted time-based metric.
These will be combined by averaging ranks for the measures, and statistical significance will be evaluated only for the segmentation performance measures and will be quantified by permuting the relative ranks for each segmentation measure and tissue class per subject of the testing data.
[3] S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

This permutation testing would reflect differences in performance that exceeded those that might be expected by chance.

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.

[1] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", IEEE Transactions on Medical Imaging 34(10), 1993-2024 (2015) DOI: 10.1109/TMI.2014.2377694
[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, 4:170117 (2017) DOI: 10.1038/sdata.2017.117
[3] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, et al., "Identifying the Best Machine Learning

Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge", arXiv preprint arXiv:1811.02629 (2018)
[4] U.Baid, et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification", arXiv:2107.02314, 2021.
[5] R.Cox, J.Ashburner, H.Breman, K.Fissell, C.Haselgrove, C.Holmes, J.Lancaster, D.Rex, S.Smith, J.Woodward, "A (Sort of) new image data format standard: NIfTI-1: WE 150", Neuroimage, 22, 2004.
[6] T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.
[7] S.Thakur, J.Doshi, S.Pati, S.Rathore, C.Sako, M.Bilello, S.M.Ha, G.Shukla, A.Flanders, A.Kotrotsou, M.Milchenko, S.Liem, G.S.Alexander, J.Lombardo, J.D.Palmer, P.LaMontagne, A.Nazeri, S.Talbar, U.Kulkarni, D.Marcus, R.Colen, C.Davatzikos, G.Erus, S.Bakas, "Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-institutional Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training", NeuroImage, 220: 117081, 2020. DOI: 10.1016/j.neuroimage.2020.117081
[8] Duan R, Tong J, Lin L, Levine LD, Sammel MD, Stoddard J, Li T, Schmid CH, Chu H, Chen Y. PALM: Patient-centered Treatment Ranking via Large-scale Multivariate Network Meta-analysis. medRxiv. 2020 Jan 1
[9] Wiesenfarth, Manuel, Annika Reinke, Bennett A. Landman, Matthias Eisenmann, Laura Aguilera Saiz, M. Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. "Methods and open-source toolkit for analyzing and visualizing challenge results." Scientific reports 11, no. 1 (2021): 1-15.
[10] L. Maier-Hein et al., "Why rankings of biomedical image analysis competitions should be interpreted with care," Nat. Commun., vol. 9, no. 1, pp. 1-13, Dec. 2018, doi: 10.1038/s41467-018-07619-7.
[11] Reinke, Annika, et al. "Understanding metric-related pitfalls in image analysis validation." ArXiv (2023).
[12] Maier-Hein, Lena, and Bjoern Menze. "Metrics reloaded: Pitfalls and recommendations for image analysis validation." arXiv. org 2206.01653 (2022).

Further comments

Further comments from the organizers.

N/A