Federated Tumor Segmentation: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Federated Tumor Segmentation

Challenge acronym

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

FeTS

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.

As the first challenge to ever be proposed for federated learning in medicine, the Federated Tumor Segmentation (FeTS) challenge 2021 intends to address these hurdles, both for the creation and the evaluation of tumor segmentation models. Specifically, the FeTS 2021 challenge uses clinically acquired, multi-institutional MRI scans from the BraTS challenge, as well as from various remote independent institutions included in the collaborative network of a real-world federation (www.fets.ai). The FeTS challenge focuses on the construction and evaluation of a consensus model for the segmentation of intrinsically heterogeneous (in appearance, shape, and histology) brain tumors, namely gliomas [1]. Compared to the BraTS challenge [2-4], the ultimate goal of FeTS is 1) 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), and 2) the evaluation of segmentation models in such a federated configuration (i.e., in the wild).

The FeTS 2021 challenge is structured in two tasks:
- Task 1 ("Federated Training") aims at effective weight aggregation methods for the creation of a consensus model, given a pre-defined segmentation algorithm for training, while also (optionally) accounting for network outages.
- Task 2 ("Federated Evaluation") aims at robust segmentation algorithms, given a pre-defined weight aggregation method, evaluated during the testing phase on unseen datasets from various remote independent institutions of the collaborative network of the fets.ai federation.
To prepare for both these tasks, the participants can use the information provided on data origin and acquisition

settings during the training phase of the challenge.
We intend to add a third task in the FeTS challenge 2022 to account for adversaries during the training phase.

The clinical relevance and importance of the FeTS challenge is that it addresses challenges related to privacy, legal, bureaucratic, and ownership concerns. Ground truth reference annotations are created and approved by expert neuroradiologists for every subject included in the training, validation, and testing datasets to quantitatively evaluate the performance of the participating algorithms.

Participants are free to choose whether they want to focus on only one or multiple tasks.

Challenge keywords

List the primary keywords that characterize the challenge.

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

Year

The challenge will take place in ...

2021

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

Distributed and Collaborative Learning (DCL) and/or Brain-Lesion Workshop (BrainLes)

We would appreciate the recommendation of the Area Chair to combine the FeTS challenge with the DCL workshop as a full-day event, since both these satellite events revolve around collaborative learning and hence refer to the same community.

We intend to combine the FeTS 2021 proceedings with the BrainLes workshop LNCS proceedings to leverage their agreement with Springer for proceedings submitted for publication after MICCAI, thereby giving more time to the FeTS challenge participants to develop their methods.

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 conservatively estimate participation from (at least) 30 teams, considering:
i) the increasing interest for federated learning (FL) in medical imaging,
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 8 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),
iii) since last year's BraTS and the publication of our related FL manuscripts, we have received >30 requests for revealing the data contributions from the individual institutions to the BraTS 2020 data, and we expect these users to be interested in participating in the FeTS 2021 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 extend their individual papers to 12-14 pages, and hence publish their methods in the workshop's LNCS post-conference proceedings.

Plan 2:
Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of FeTS 2021, making a comprehensive meta-analysis to inform the community about the obtained results, in particular focusing on the evaluation of each aggregation method and the unique federated evaluation that was never attempted in a MICCAI challenge before.

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, 2 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. Federated learning 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: Federated Training (Weight Aggregation) for Glioma Segmentation

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 is the focus of the BraTS challenge). 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. The primary objective of this task is to develop methods for effective aggregation of local 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.

Keywords

List the primary keywords that characterize the task.

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

ORGANIZATION

Organizers

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

Organizing team:

Spyridon Bakas, Ph.D., – [Lead Organizer - Contact Person]
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Micah Sheller
Intel Labs

Sarthak Pati, M.Sc.
CBICA, University of Pennsylvania, Philadelphia, PA, USA

Brandon Edwards, Ph.D.
Intel Labs

G Anthony Reina, M.D.
Intel Internet of Things Group

Ujjwal Baid, Ph.D.
CBICA, University of Pennsylvania, Philadelphia, PA, USA

Yong Chen, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Russell (Taki) Shinohara, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Jason Martin
Intel Labs

Bjoern Menze, Ph.D.
University of Zurich, Switzerland

Shadi Albarqouni, Ph.D.
Helmholtz AI & Technical University of Munich, Germany

Clinical Evaluators and Annotation Approvers:

Michel Bilello, MD, Ph.D.,
University of Pennsylvania, Philadelphia, PA, USA

Suyash Mohan, MD, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Data Contributors:

John B. Freymann & Justin S. Kirby - on behalf of The Cancer Imaging Archive (TCIA), Cancer Imaging Program, NCI, National Institutes of Health (NIH), USA

Christos Davatzikos, Ph.D.,
CBICA, University of Pennsylvania, Philadelphia, PA, USA

Hassan Fathallah-Shaykh, M.D., Ph.D.,
University of Alabama at Birmingham, AL, USA

Roland Wiest, M.D.,
University of Bern, Switzerland

Andras Jakab, M.D., Ph.D.,

University of Debrecen, Hungary

Rivka R. Colen, M.D.
University of Pittsburgh Medical Center

Aikaterini Kotrotsou, Ph.D.,
MD Anderson Cancer Center, TX, USA

Daniel Marcus, Ph.D., & Mikhail Milchenko, Ph.D., & Arash Nazeri, M.D.,
Washington University School of Medicine in St.Louis, MO, USA

Marc-Andre Weber, M.D.,
Heidelberg University, Germany

Abhishek Mahajan, M.D. & Ujjwal Baid, Ph.D.,
Tata Memorial Center, Mumbai, India

Philipp Vollmuth, M.D.
Heidelberg University Clinics

b) Provide information on the primary contact person.

Spyridon Bakas, Ph.D., – [Lead Organizer - Contact Person]
CBICA, University of Pennsylvania, Philadelphia, PA, USA
sbakas@upenn.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).

MICCAI.

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

In consistency with the BraTS 2017-2020 challenge, we will be using the University of Pennsylvania's Image Processing Portal (ipp.cbica.upenn.edu) for running the challenge.

Implementation of exact evaluation metrics will be made publicly available in favor of transparency, through FeTS.

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

www.fets.ai/miccai2020 - (Website will be publicly visible after the challenge approval)

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 will not be eligible for awards.

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 leaderboard. Note that any potential participating Intel employee (outside the immediate groups of the organizers) should not be eligible for any awards.

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

Intel will sponsor a $5K award for the top 3 teams.

Additionally, the top 3 algorithms will be incorporated in the FeTS platform and utilized in the first ever real world federation across >55 institutions (www.fets.ai) to create the largest FL model focusing on the segmentation of brain glioma. The authors of the top-ranked participating team will also be co-authors of the manuscript that will summarize the results of this analysis.

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 extend their papers to 12-14 pages, and hence 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 2020, making comparative assessment with the summary results of the BraTS 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 algorithm in the form of an FeTS/OpenFL component, 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. 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 both the FeTS platform (which will provide the implementation of the evaluation metrics), as well as via the online evaluation platform (ipp.cbica.upenn.edu).

The FeTS/OpenFL component requirements, specifications, and examples will be described in:
www.fets.ai/miccai2020/openflcomponents

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) the FeTS platform including OpenFL as the backend library, and 2) instructions on how to run a federation learning simulation on a single system. (Note that the package will include a complete multi-node federated learning implementation that participants can use along with the OpenFL documentation if they choose.)
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.
4. The validation metrics to be computed each round (which can then be used as inputs to the other functions).
5. Compression of model update uploads/downloads.

6. Logic to determine whether to end a round early due to a slow collaborator (this relates to the optional part of Task 1).

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 in June, 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 (ipp.cbica.upenn.edu) 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

- Registration dates: From challenge's approval until submission deadline of short papers reporting method and preliminary results (see below).
- 01 May 2020: Expected release of training data
- 15 June 2020: Expected release of validation data
- 16 July 2020: Submission of 1) short papers reporting method & preliminary results and 2) source code & aggregated model.
- 19 July - 27 August 2020: Evaluation on testing data (by the organizers - only for participants with submitted papers).
- 04 September 2020: Contacting top performing methods for preparing slides for oral presentation.

- 27 September 2020: Announcement of final top 3 ranked teams: Challenge at MICCAI
- 30 October 2020: Camera-ready submission of extended papers for inclusion in the associated workshop proceedings

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 MICCAI BraTS challenge data, that are currently coordinated for release 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, 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 algorithm as a FeTS/OpenFL component implemented in PyTorch together with their final submitted results, during the validation phase. Specific instructions for creating FeTS/OpenFL components will be provided at www.fets.ai/miccai2020/openflcomponents

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.

Monetary awards are expected by Intel.
Spyridon Bakas, 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

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

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation.
Federated Learning Weight Aggregation.

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

multi-parametric MRI acquisition protocol including i) native 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 multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Imaging modality(ies)

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

MRI

Context information

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

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

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

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

Age, Gender.

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 multi-parametric MRI scans.

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)).

Specificity, Sensitivity.

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 [4]. Since then, multiple institutions have contributed data to the create the current BraTS dataset and these are listed in the current BraTS arxiv paper [3]. We are currently in coordination with TCIA to make the complete BraTS dataset permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[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
[4] 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

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 these 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 [4]. Since then multiple institutions have contributed data to the create the current BraTS dataset and these are listed in the current BraTS arxiv paper [3]. We are currently in coordination with TCIA to make the complete BraTS dataset permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[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
[4] 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

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.

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) native 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 2021 training, validation, and testing data will be leveraging the BraTS data, augmented with metadata related to the institution the data originates from and acquisition protocols. We plan to use the BraTS 2021 training data, but since these are still in process due to the BraTS challenge being an RSNA challenge in 2021, we report the numbers for the BraTS 2020 dataset below, which is the minimum amount of data to be expected. The exact numbers from 2020 are:
Training data: 369 patients

Validation data: 125 patients
Testing data: 166 patients

For further data augmentation we already 1) working with clinical experts from our institutions to manually annotate scans of existing publicly available datasets (e.g., ECOG-ACRIN clinical trial available data available through TCIA), and 2) provide more multi-parametric MRI scans of gliomas from our affiliated institutions.

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 from at least 2 experienced neuroradiologists.

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 2021 challenge is the BraTS 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.

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 >12 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 2021 challenge is identical with the one evaluated and followed by the BraTS challenge. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format [10], 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 [5]) and interpolating to the same resolution as this atlas (1 mm^3).
The exact registration process comprises the following steps:
1. N4 Bias field correction (notably the application of N4 bias field correction is a temporary step. Taking into consideration we have previously [4] 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 in the LPS-oriented scans towards the final co-registered output images used in the challenge).
2. Rigid Registration of T1, T2, T2-FLAIR to the T1-Gd scan, and obtain the corresponding transformation matrix.
3. Rigid Registration of T1-Gd scan to the SRI-24 atlas [5], and obtain the corresponding transformation matrix.
4. 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 [11]. 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.

[4] 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] T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.
[10] 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.
[11] 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 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),
Budget time provided to the participants, defined as the product of bytes sent/received multiplied by the number of federated rounds.

Sensitivity,
Specificity,

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 describes what is typically resected during a surgical procedure.
iii) the whole tumor as it defines the whole extent of the tumor, 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 havings too much weight,
iii) The budget will capture the communication cost for each algorithm
v) 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 [6].

[6] 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.

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 (Drs Chen and Shinohara), 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 2020, 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 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

TASK: Federated Evaluation of Glioma Segmentation Methods In The Wild

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 discrepancy between AI systems' performance in research environments and real-life applications is one of the key challenges in our field. The reason behind this "AI chasm" can be related to limited training data that impede the generalization ability of models, as they do not reflect the variety of real-world datasets acquired from different imaging devices and sequences ("in the wild"). Federated setups, in which clinicians may contribute data to a challenge without having to publicly release them, allow to extend the size and diversity of typical test datasets substantially, thus constituting an important step towards the evaluation of model robustness in the wild.

In this task, we are seeking to demonstrate that it is feasible to scale up the concept of challenges by implementing a challenge within a real-world federated evaluation environment. Specifically, during training, the participants will be provided the multi-institutional BraTS dataset including information on data origin and acquisition settings. They can explore the effects of data partitioning and distribution shifts between contributing sites, with the aim of finding tumor segmentation algorithms that are able to generalize to data acquired at institutions that did not contribute to the training dataset ("domain generalization"). Note that training on pooled data will also be allowed. After training, all participating algorithms will be evaluated in a distributed way on datasets from various institutions, such that the test data are always retained within their owners' servers.

To facilitate this, we will leverage the collaborators of the first real-world medical federation, as described in www.fets.ai. While federated evaluation schemes circumvent many of the common obstacles with sharing medical data (such as data-privacy issues), they also come with their own set of challenges and particularities. This "real-world semantic segmentation challenge" is set up to address these obstacles. Hopefully, it will provide a blueprint for similar endeavors in the future, where successful challenges could enter a "phase 2" with a federated setup and thus move one step closer towards the real-life use case. From a methodical perspective, the main goal of this task is to identify segmentation algorithms that are robust to unknown and realistic distribution shifts between training and test data.

Keywords

List the primary keywords that characterize the task.

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

ORGANIZATION

Organizers

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

Organizing team:

Maximilian Zenk – [Lead Organizer - Contact Person]

Div. Medical Image Computing (MIC), German Cancer Research Center (DKFZ)

Klaus Maier-Hein, Ph.D.
Div. Medical Image Computing (MIC), German Cancer Research Center (DKFZ)

Spyridon Bakas, Ph.D.
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Micah Sheller
Intel Labs

Sarthak Pati, M.Sc.
CBICA, University of Pennsylvania, Philadelphia, PA, USA

Brandon Edwards, Ph.D.
Intel Labs

G Anthony Reina, M.D.
Intel Internet of Things Group

Ujjwal Baid, Ph.D.
CBICA, University of Pennsylvania, Philadelphia, PA, USA

Yong Chen, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Russell (Taki) Shinohara, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Jason Martin
Intel Labs

Bjoern Menze, Ph.D.
University of Zurich, Switzerland

Shadi Albarqouni, Ph.D.
Helmholtz AI & Technical University of Munich, Germany

David Zimmerer
Div. Medical Image Computing (MIC), German Cancer Research Center (DKFZ)

Annika Reinke
Div. Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ)

Lena Maier-Hein, Ph.D.

Div. Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ)

Jens Kleesiek, Ph.D., MD
Translational Image-guided Oncology, Institute for AI in Medicine (IKIM), University Hospital Essen

Clinical Evaluators and Annotation Approvers:

Michel Bilello, MD, Ph.D.,
University of Pennsylvania, Philadelphia, PA, USA

Suyash Mohan, MD, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Data Contributors:

John B. Freymann & Justin S. Kirby - on behalf of The Cancer Imaging Archive (TCIA), Cancer Imaging Program, NCI, National Institutes of Health (NIH), USA

Christos Davatzikos, Ph.D.,
CBICA, University of Pennsylvania, Philadelphia, PA, USA

Hassan Fathallah-Shaykh, M.D., Ph.D.,
University of Alabama at Birmingham, AL, USA

Roland Wiest, M.D.,
University of Bern, Switzerland

Andras Jakab, M.D., Ph.D.,
University of Debrecen, Hungary

Rivka R. Colen, M.D.
University of Pittsburgh Medical Center

Aikaterini Kotrotsou, Ph.D.,
MD Anderson Cancer Center, TX, USA

Daniel Marcus, Ph.D., & Mikhail Milchenko, Ph.D., & Arash Nazeri, M.D.,
Washington University School of Medicine in St.Louis, MO, USA

Marc-Andre Weber, M.D.,
Heidelberg University, Germany

Abhishek Mahajan, M.D. & Ujjwal Baid, Ph.D.,
Tata Memorial Center, Mumbai, India

Philipp Vollmuth, M.D.
Heidelberg University Clinics

b) Provide information on the primary contact person.

Maximilian Zenk – [Lead Organizer - Contact Person]
Div. Medical Image Computing (MIC), German Cancer Research Center (DKFZ)
m.zenk@dkfz-heidelberg.de

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).

MICCAI.

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

In consistency with the BraTS 2017-2020 challenge, we will be using the University of Pennsylvania's Image Processing Portal (ipp.cbica.upenn.edu) for running the challenge.

Implementation of exact evaluation metrics will be made publicly available in favor of transparency, through FeTS.

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

www.fets.ai/miccai2020 - (Website will be publicly visible after the challenge approval)

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 will not be eligible for awards.

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 leaderboard. Note that any potential participating Intel employee (outside the immediate groups of the organizers) should not be eligible for any awards.

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

Intel will sponsor a $5K award for the top 3 teams.

Additionally, the top 3 algorithm will be incorporated in the FeTS platform and utilized in the first ever real world federation across >55 institutions (www.fets.ai) to create the largest FL model focusing on the segmentation of brain glioma. The authors of the top-ranked participating team will also be co-authors of the manuscript that will summarize the results of this analysis.

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 extend their papers to 12-14 pages, and hence 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 2020, making comparative assessment with the summary results of the BraTS 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. Participants need to submit:
1) their inference algorithm in the form of an FeTS/OpenFL component, and

2) the weights of the selected model.
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. Notably, Docker or other containerization technology will probably not be available due to IT restrictions in the federation of participating clinical sites. Towards this end, participants should implement their approach in compliance with the FeTS/OpenFL framework to enable ranking in these independent institutions, while ensuring safety during execution and avoiding potential malware inclusions in the participating algorithms.

During the training and validation phases, the participants will have the chance to test the functionality of their submission through both the FeTS platform (which will provide the implementation of the evaluation metrics), as well as via the online evaluation platform (ipp.cbica.upenn.edu).

The FeTS/OpenFL component requirements, specifications, and examples will be described in: www.fets.ai/miccai2020/openflcomponents

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 in June, 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 (ipp.cbica.upenn.edu) 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

- Registration dates: From challenge's approval until submission deadline of short papers reporting method and preliminary results (see below).
- 01 May 2020: Expected release of training data
- 15 June 2020: Expected release of validation data
- 16 July 2020: Submission of 1) short papers reporting method & preliminary results and 2) source code & aggregated model.
- 19 July - 27 August 2020: Evaluation on testing data (by the organizers - only for participants with submitted papers).
- 04 September 2020: Contacting top performing methods for preparing slides for oral presentation.
- 27 September 2020: Announcement of final top 3 ranked teams: Challenge at MICCAI
- 30 October 2020: Camera-ready submission of extended papers for inclusion in the associated workshop proceedings

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 MICCAI BraTS challenge data, that are currently coordinated for release 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, 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 algorithm as a FeTS/OpenFL component implemented in PyTorch together with their final submitted results, during the validation phase. Specific instructions for creating FeTS/OpenFL components will be provided at www.fets.ai/miccai2020/openflcomponents

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.

Monetary awards are expected by Intel.
Only individual institutions 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

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

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 multi-parametric MRI acquisition protocol including i) native 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 multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-

weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Imaging modality(ies)

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

MRI

Context information

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

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

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

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

Age, Gender.

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 multi-parametric MRI scans.

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)).

Specificity, Sensitivity.

Additional points: Find highly accurate brain tumor segmentation algorithm for MRI images, that works robustly "in the wild".

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 [4]. Since then, multiple institutions have contributed data to the create the current BraTS dataset and these are listed in the current BraTS arxiv paper [3]. We are currently in coordination with TCIA to make the complete BraTS dataset permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[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
[4] 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

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 these 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 [4]. Since then multiple institutions have contributed data to the create the current BraTS dataset and these are listed in the current BraTS arxiv paper [3]. We are currently in coordination with TCIA to make the complete BraTS dataset permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[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
[4] 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

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.

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)

Additional data will be used during the testing phase from the numerous institutions involved in the federation described at: www.fets.ai

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) native 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 2021 training, validation, and testing data will be leveraging the BraTS data, augmented with metadata related to the institution the data originates from and acquisition protocols. We plan to use the BraTS 2021 training data, but since these are still in process due to the BraTS challenge being an RSNA challenge in 2021, we report the numbers for the BraTS 2020 dataset below, which is the minimum amount of data to be expected. The

exact numbers from 2020 are:
Training data: 369 patients
Validation data: 125 patients
Testing data (for the actual testing phase of this task): The actual federated evaluation will be performed on a subset of the numerous institutions participating in the FeTS federation as described at: www.fets.ai. After coordinating with these institutions, we expect that each of the independent institutions will contribute at least 20 cases, with some of the institutions of the FeTS federation will contribute larger amounts of test cases.
In addition, we intend to use the BraTS testing data (166 patients from the 2020 challenge) for complementary analyses, and not for the actual challenge.

For further data augmentation we are already 1) working with clinical experts from our institutions to manually annotate scans of existing publicly available datasets (e.g., ECOG-ACRIN clinical trial available data available through TCIA), and 2) provide more multi-parametric MRI scans of gliomas from our affiliated institutions.

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 from at least 2 experienced neuroradiologists.

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 2021 challenge is the BraTS 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.

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 >12 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 2021 challenge is identical with the one evaluated and followed by the BraTS challenge. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format [10], 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 [5]) and interpolating to the same resolution as this atlas (1 mm^3).
The exact registration process comprises the following steps:
1. N4 Bias field correction (notably the application of N4 bias field correction is a temporary step. Taking into consideration we have previously [4] 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 in the LPS-oriented scans towards the final co-registered output images used in the challenge).
2. Rigid Registration of T1, T2, T2-FLAIR to the T1-Gd scan, and obtain the corresponding transformation matrix.
3. Rigid Registration of T1-Gd scan to the SRI-24 atlas [5], and obtain the corresponding transformation matrix.
4. 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 [11]. 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.

[4] 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] T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.
[10] 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.
[11] 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.

The inter-annotator variability of the brain tumor segmentation task has been studied in the context of the BraTS challenge [3]. In our federated setup, the additional challenge will be to cope with the multi-centric setting with different annotators. Inter-annotator variability will hence be a possible source of errors despite the shared annotation protocol. In our attempt to control the quality of annotations and correct inconsistencies, we have been in teleconferences with each independent site, using screen sharing, to virtually confirm that the annotation protocol is followed.. Furthermore, we plan to identify outlier institutions for closer investigation by analysing the performance of a baseline segmentation model.

[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) 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),

Sensitivity,
Specificity,
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 describes what is typically resected during a surgical procedure.
iii) the whole tumor as it defines the whole extent of the tumor, 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 havings too much weight,
v) 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.

Only the external FeTS testing institutions (that are not part of the training data) are used for the ranking. In a first step, the significance ranking used in the medical segmentation decathlon [7] is adopted, which performs a statistical test between all pairs of algorithms to determine whether one works significantly better than the other. This produces six rankings (two metrics times three evaluated tumor regions) for each testing institution separately. The resulting per-institution rankings are averaged in a second step to obtain the final ranking.

[7] A. L. Simpson et al., “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” arXiv:1902.09063 [cs, eess], Feb. 2019. Available: http://arxiv.org/abs/1902.09063.

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.

The goal of this challenge is to compare the performance of the participating algorithms on different institutions (i.e., data distributions). Since the number of samples per institution varies widely, a pooled analysis of all test cases would bias the ranking towards institutions with many patients. Hence we compute rankings for each testing institution using the test-based ranking method from [7], which takes into account statistical significance. For the final ranking, we used the same rank aggregation method as in previous BraTS challenges.

[7] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," arXiv:1902.09063 [cs, eess], Feb. 2019. Available: http://arxiv.org/abs/1902.09063.

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 2020, 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. These will be combined by averaging ranks for the measures, and statistical significance will be quantified by permuting the relative ranks for each segmentation measure and tissue class per subject of the testing data. Ranking stability will further be investigated via bootstrapping and hypothesis testing (using the challengeR library [8]).

[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.
[8] M. Wiesenfarth, A. Reinke, B. A. Landman, M. J. Cardoso, L. Maier-Hein, and A. Kopp-Schneider, "Methods and open-source toolkit for analyzing and visualizing challenge results," ArXiv191005121 Cs Stat, Dec. 2019, Accessed: Dec. 02, 2020. [Online]. Available: http://arxiv.org/abs/1910.05121.

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. Bootstrapping was identified as an appropriate approach to investigate ranking variability in [9].

[9] 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.

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.

Based on the pooled results and meta-data (like scanner type, acquisition settings or anonymized patient information), we plan to analyse which meta-features are correlated with algorithm performance and to identify outlier institutions. We will also compare the performances on the out-of-distribution FeTS test set and on the in-distribution BraTS test set to assess robustness. Thus, we hope to gain insights into the main reasons for algorithms succeeding or failing to generalize to unseen data distributions.

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] M.J.Sheller, et al. "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data." Scientific reports. 10:1-12, 2020.
DOI: 10.1038/s41598-020-69250-1

[2] B. H. Menze, 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

[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

[4] S. Bakas, 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] T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.

[6] Duan R, et al. PALM: Patient-centered Treatment Ranking via Large-scale Multivariate Network Meta-analysis. medRxiv. 2020.

[7] A. L. Simpson et al., “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” arXiv:1902.09063

[8] M. Wiesenfarth, et al. “Methods and open-source toolkit for analyzing and visualizing challenge results,” arXiv:1910.05121.

[9] L. Maier-Hein, et al., “Why rankings of biomedical image analysis competitions should be interpreted with care,” Nat. Commun., 9(1):1–13, 2018.
DOI: 10.1038/s41467-018-07619-7

[10] R.Cox, et al. "A (Sort of) new image data format standard: NIfTI-1: WE 150", Neuroimage, 22, 2004.

[11] S.Thakur, et al. "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