Brain Tumor Progression Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Brain Tumor Progression Challenge

Challenge acronym

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

BraTPRO

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.

Brain tumors are among the most researched diseases in the field of medical image computing. This is reflected by the popularity of challenges like the MICCAI BraTS [1], the MICCAI FeTS challenge [2], and the use of brain MRIs in challenges like MICCAI MOOD [3] for anomaly detection. Methods for semantic segmentation developed in this context show astonishing results, even comparable to intra- and inter-rater variability [4], with only marginal differences between the top performing participants. However, most of the challenges and further research focus only on single time-points. Longitudinal properties are usually only inferred from the single time-point segmentation results if needed (e.g. [5]). These longitudinal properties play a huge role in the field of brain tumor research, because they can be used for assessing treatment response. The RANO (Response Assessment in Neuro Oncology) working group [6] defines different types of response, namely complete response, partial response, stable disease and progressive disease. Progressive disease by contrast enhancing lesions is defined as an increase of the product of perpendicular diameters enhancing tumor lesions by at least 25% [6] or the appearance of any newly formed enhancing lesion [6]. The early detection of these kinds of progression in brain tumor patients is crucial for further treatment decisions, as well as assessing the response to drugs, e.g. in clinical studies.
The two kinds of progression can be extracted in a (semi)-automatic manner from the segmentations of the individual time points.

The gain in tumor volume can be extracted from automated segmentation on the individual time points, by translating the threshold on tumor growth to an increase of 40% in tumor volume [4,8]. These measurements are already optimized on the single scans.
Detection of newly formed lesions from automated segmentations is a significantly more sophisticated procedure. It involves registration between consecutive scans (for difficulties regarding longitudinal registration see [7]), the definition of "volume at risk" (i.e. volume where a new lesion might appear) and finally distinguishing newly formed lesions in this volume from lesion growth from existing lesions. A more in-depth description of this

procedure is given in [8].
This is further complicated by the optimization targets of common semantic segmentation methods. On the one hand, they do not optimize for the detection of individual lesions, posing the risk that a newly formed lesion is not detected by the network. On the other hand, they often do not take longitudinal information into account, ignoring important information.
Finally, automated segmentations do not properly cover non-measurable lesions, which play an important role in the qualitative response assessment.

An end-to-end approach for the detection of disease progression, circumventing manual interventions, and tumor segmentations would therefore be preferable. To this end, we propose the Brain Tumor Progression Challenge (BraTPRO) to address this gap in current research, in partnership with the RANO group. The challenge consists of two tasks, both tackling the classification of the different types of response according to RANO criteria (complete response, partial response, stable disease, progressive disease - see [6] for more details about the different types of response).

For the first task, participants will develop novel methods on our publicly available dataset with response classification annotations. They then have to submit their final method for training and evaluation on a private dataset on the organizers' computing infrastructure. The first task represents an idealized scenario of iid hidden training and test datasets.

In the second task, participants will develop and train their method with all data available to them - including our provided public dataset with response classification annotations and any other accessible data sources. Their final model will be submitted and inference will be executed on the organizers' hidden test set. As the hidden test set is proprietary, this represents a more realistic scenario as a shift between the participants chosen training dataset and ours may be present.

Splitting the challenge in these two tasks will allow us to investigate the contributions of method development on the one hand, as well as contributions due to dataset selection on the other hand.

The publicly available longitudinal dataset that is provided to the participants is the LUMIERE dataset [9]. It comprises 91 patients with a total of 616 scans and annotations according to RANO criteria.. Our hidden test dataset is the multicentric EORTC-26101 dataset. This dataset comprises 306 patients with glioblastoma, as previously reported in [8], for which we provide response classification annotations according to RANO criteria and ground truth segmentation masks. We split this dataset into 300 cases that are reserved for the final test phase and 6 cases that are used for validation.
In the first phase of the challenge participants are provided the LUMIERE dataset with response classification annotations and the automatically generated segmentations. During this phase participants can develop their methods and train the models depending on the task. Moreover, participants can submit their model or training algorithm once per day to the organizers for validation on the 6 validation cases.
In the final phase participants have to submit their method or trained model to the organizers. In Task 1 the submitted method will be trained on 200 of the 300 test cases of the EORTC-26101 dataset, with the remaining 100 cases being used for performance evaluation. In Task 2 the performance of the trained models will be evaluated as is on the same 100 test cases to allow a direct comparison between solutions created in Task 1 and

Task 2.
We hope that this challenge will raise awareness of the gap in current research related to longitudinal properties beyond the field of brain tumor research.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Brain Tumors, Disease Progression, RANO, Longitudinal Image Analysis

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

This challenge will run in coordination with the BraTS challenge.

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.

Due to the strong connection of this challenge to the BraTS challenge and the large interest of the community in brain tumor research, we estimate a number of participants of around 30-40 teams.

Publication and future plans

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

We plan to publish the results after the challenge, where we also plan to present a meta-analysis of the different submitted methods and the two tasks.

Space and hardware requirements

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

The algorithms are run on the participant's hardware for development in task 1 and training for task 2. Docker or MLCube images will be uploaded via the Synapse platform for validation runs and the final training/testing runs, which will be carried out on the organizers' computing infrastructure.
We will be able to support at least O(50) submissions for model training in task 1 over the given time frame (latest submission date - beginning of conference). If we exceed this limit, only the 50 best performing submissions on task 2 will be evaluated for task 1.

Hardware requirements in case of an in-person meeting: 1 projector, 2 microphones, loudspeakers.

TASK 1: RANO Classifier Development

SUMMARY

Abstract

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

This task focuses on method development. Participants will develop their method using only the provided dataset and then submit the model to be trained and evaluated on the organizers' data and computing infrastructure. Details about the datasets used in this task can be found in the respective dataset section.

Keywords

List the primary keywords that characterize the task.

Brain Tumors, Disease Progression, RANO, Longitudinal Image Analysis

ORGANIZATION

Organizers

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

Yannick Kirchhoff, Tassilo Wald, Balint Kovacs, Maximilian Zenk, Klaus Maier-Hein
German Cancer Research Center (DKFZ), Heidelberg, Division of Medical Image Computing, Germany

Philipp Vollmuth
Department of Neuroradiology, Heidelberg University Hospital, Germany

Jens Kleesiek, Jan Egger
Institute for AI in Medicine (IKIM), University Hospital Essen, Germany

Yannick Suter, Mauricio Reyes
ARTORG Center, University Bern, Switzerland

André Ferreira
Center Algoritmi, University of Minho, Braga, Portugal
Institute for AI in Medicine (IKIM), University Hospital Essen, Germany

Spyridon Bakas (Indiana University, Indianapolis, IN, USA)
Raymond Y Huang (MGB, Boston, MA, USA)
Javier Villanueva-Meyer (University of California San Francisco, San Francisco, CA, USA)
AI-RANO Group leads

b) Provide information on the primary contact person.

Yannick Kirchhoff
yannick.kirchhoff@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

One time event with fixed submission deadline.
However, after the challenge we plan to enable further testing of methods on the test set (with a separate leaderboard). In case of a successful challenge, we plan to repeat this challenge at future MICCAIs.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

N/A

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

synapse.org

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

Project SynID: syn53752772

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.

Methods will be trained on the organizers' private dataset. There are no special rules for method development, however, using the provided public dataset is recommended.

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 not participate

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

Prizes are yet to be determined.

e) Define the policy for result announcement.

Examples:

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

The top 3 performing methods will be announced publicly at the conference if they don't opt-out and the teams will be invited to present their method. All teams are free to decide if they want to show up on the public leaderboard.

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).

Teams will be invited to nominate members as co-authors if they:
- follow all the rules of the challenge
- open-source their algorithm
The organizers reserve the right to exclude teams or members from the author list in case of violation of these rules.
Submissions of teams which decide to not nominate anyone as co-authors will still be used in the publication.
We do not limit the number of co-authors per team.
The participating teams are encouraged to publish their methods separately and we will not enforce an embargo time.

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.

Docker/MLCube container via the Synapse platform. Link for submission and detailed instructions will be provided later.

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

There will only be one possible submission for the final training and testing, which also determines the final score. However, we offer the possibility of submissions for the validation set which is taken from the same data-distribution as the training/test set. Submissions for the validation set for this task should mainly serve testing the training/inference code and are limited to one submission per day.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

1. April 2024: Release of development cases
1. June 2024: Opening of evaluation pipeline for validation
1. August 2024: Opening of evaluation pipeline for final submission + registration deadline
1. September 2024: Deadline for submission of dockers and submission of short papers with report of method and preliminary results
8. September 2024: Deadline to open source submitted algorithm

Ethics approval

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

The development set relies on the already publicly available dataset, for which the cantonal ethics committee of Bern (Switzerland) approved the studies and waived written informed consent.

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)

Public development set: CC BY-NC
Training and test set are not published

Code availability

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

The evaluation code for the challenge will be made publicly available on GitHub.

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

Participating teams have to open-source their code under a license which allows public use (preferably CC-BY license) in order to be eligible to win the challenge.

Conflicts of interest

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

The challenge may obtain dedicated sponsoring or funding, which will not have influence on challenge design, evaluation and results.
Only members of MIC@DKFZ and the clinical partners responsible for data acquisition and labeling have access to the test 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

Decision support, Longitudinal study, Research, Diagnosis

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization

- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification

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.

Patients with a previous medical record of glioma and multiple consecutive MRI scans, pre- and/or posttreatment. At least one follow-up examination.

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

Patients with a previous medical record of glioma and multiple consecutive MRI scans, pre- and/or posttreatment. At least one follow-up examination.

Imaging modality(ies)

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

MRI (T1, cT1, T2, FLAIR)

Context information

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

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

Development: The development dataset contains many additional information in addition to the MRI scans like MR acquisition parameters and time between scans. Notably automatically generated segmentations using two tools are given for each scan.
Training: Statistics to image parameters on the training, e.g. spacing and size are given at the beginning of the challenge. Ground Truth segmentation masks will be available to participants for training.
Testing: None

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

Development: The development set contains metadata on the patients regarding for example patient sex, overall survival and age
Training and Testing: None

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 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.

Brain tumors

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)).

Precise classification of response according to RANO

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).

Development data: The development data, the LUMIERE dataset [9], stems from the Bern University Hospital (Inselspital), the pre-operative scans were acquired between 2008 and 2013, follow-ups were recorded until 2017. 95% of the 2487 provided MRI images have been acquired on Siemens scanners, 3% on Philips scanners (Philips Medical Systems/Philips Healthcare), and 2% on scanners from GE Medical Systems. Information on the respective scanner is available for all scans.
Training and Testing data: n/a

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

Development data: Image acquisition parameters are given for all individual scans.
Training and Testing data: n/a

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.

Development data: The development data was acquired at the Bern University Hospital (Inselspital).
Training and Testing data: Multi-center dataset, more precise information on included centers can be shared with reviewers if they agree not to participate in the challenge.

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).

Development data: n/a
Training and Testing data: n/a

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.

Cases consist of two consecutive brain MRI scans of a single patient. The provided sequences are unenhanced and contrast-enhanced T1-weighted, T2-weighted and T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Development data: 91 patients with a total of 616 scans. Some scans do not contain all sequences.
Training data: 200 patients with 2 scans each resulting in a total of 400 scans
Validation data: 6 patients with 2 scans each resulting in a total of 12 scans

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

The development data is the full publicly available dataset.
The training, validation and testing data is part of a private dataset. The validation set is important to test the training and inference code of submitted docker images.

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.

There is a (possible) distribution shift between the development dataset on the one hand (data from a single center) and the training and testing cases on the other hand (data from several centers and scanners). In addition to the acquisition shift there might also be a shift in the class distributions as well as a population shift.

Information on these will however not be made available for participants.

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.

Development dataset: Manual image annotation by one annotator
Training and Testing dataset: Manual image annotation by two annotators

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.

Development dataset: No specific instructions
Training and Testing dataset: No specific instructions

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.

Development dataset: expert neuroradiologist with 14 years experience
Training and Testing dataset: n/a

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.

Development dataset: n/a
Training and Testing dataset: Consensus discussion

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.

Development data: Scans are converted to Nifti file format and skull-stripped using the HD-BET [13] tool.
Training and Testing data: Scans are converted to Nifti file format, skull-stripped using the HD-BET [13] tool and registered to the T1 volumes. Scans are not resampled to a common spacing in order to keep it close to the clinical workflow.

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 annotations are done by experienced radiologists. However, even though the RANO categorisation aims to make assessment objective, there is still a potential source of error related to ambiguous cases.

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)

Balanced Accuracy (BA), F1-score, TPR, TNR, AP, AUROC

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

The metrics were chosen following the guidelines from the Metrics Reloaded Framework [14] for an imbalanced dataset (BA, F1-score, AP) with further metrics added for additional insights (TPR, TNR, AUROC).

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.

AP and F1-score represent multi-threshold and counting metrics, respectively. The used ranking scheme will ensure that winning methods need to perform well on a fixed cutoff as well as on moving thresholds and outperform other methods on all classes.

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

As participants submit docker images for testing there should not occur any missing results. If an algorithm should still fail to produce a result for a specific case, we will assign the worst possible result to this case, i.e. a wrong label with lowest score for the true class.

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

Submissions will be ranked using F1-score and AP for all classes separately. The final ranking is based on the respective ranks by averaging over classes and metrics.

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.

Variability of rankings will be assessed by bootstrapping methods.

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

Bootstrapping is among the suitable methods for the assessment of ranking variability according to [15].

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

We plan to further analyze patterns in the predictions to find hard cases in the dataset and investigate the effect of training the methods on unseen training data and the corresponding distribution shift between development and training data.
In addition, we will investigate ranking variability using bootstrapping.

TASK 2: Data-driven RANO classification

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.

In this task, participants will develop and train their methods using any data available to them. This may include public as well as private datasets. Details about the datasets used in this task can be found in the respective dataset section.

Keywords

List the primary keywords that characterize the task.

Brain Tumors, Disease Progression, RANO, Longitudinal Image Analysis

ORGANIZATION

Organizers

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

Yannick Kirchhoff, Tassilo Wald, Balint Kovacs, Maximilian Zenk, Klaus Maier-Hein
German Cancer Research Center (DKFZ), Heidelberg, Division of Medical Image Computing, Germany

Philipp Vollmuth
Department of Neuroradiology, Heidelberg University Hospital, Germany

Jens Kleesiek, Jan Egger
Institute for AI in Medicine (IKIM), University Hospital Essen, Germany

Yannick Suter, Mauricio Reyes
ARTORG Center, University Bern, Switzerland

André Ferreira
Center Algoritmi, University of Minho, Braga, Portugal
Institute for AI in Medicine (IKIM), University Hospital Essen, Germany

Spyridon Bakas (Indiana University, Indianapolis, IN, USA)
Raymond Y Huang (MGB, Boston, MA, USA)
Javier Villanueva-Meyer (University of California San Francisco, San Francisco, CA, USA)
AI-RANO Group leads

b) Provide information on the primary contact person.

Yannick Kirchhoff
yannick.kirchhoff@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

One time event with fixed submission deadline.
However, after the challenge we plan to enable further testing of methods on the test set (with a separate leaderboard). In case of a successful challenge, we plan to repeat this challenge at future MICCAIs.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

N/A

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

synapse.org

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

Project SynID: syn53752772

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.

Public and private data is allowed in addition to the provided public dataset. All data sources must be reported.

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 not participate

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

Prizes are yet to be determined.

e) Define the policy for result announcement.

Examples:

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

The top 3 performing methods will be announced publicly at the conference if they don't opt-out and the teams will be invited to present their method. All teams are free to decide if they want to show up on the public leaderboard.

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).

Teams will be invited to nominate members as co-authors if they:
- follow all the rules of the challenge
- open-source their algorithm
The organizers reserve the right to exclude teams or members from the author list in case of violation of these rules.
Submissions of teams which decide to not nominate anyone as co-authors will still be used in the publication.
We do not limit the number of co-authors per team.
The participating teams are encouraged to publish their methods separately and we will not enforce an embargo time.

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.

Docker/MLCube container via the Synapse platform. Link for submission and detailed instructions will be provided later.

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

There will only be one possible submission for the test set, which determines the final score. However, we offer the possibility of submissions for the validation set which is taken from the same data-distribution as the test set. This will allow participants to judge the effect of the distribution shift between training and test set, as well as the performance of their methods.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

1. April 2024: Release of training cases
1. June 2024: Opening of evaluation pipeline for validation
1. August 2024: Opening of evaluation pipeline for final testing + registration deadline
1. September 2024: Deadline for submission of dockers and submission of short papers with report of method and preliminary results
8. September 2024: Deadline to open source submitted algorithm

Ethics approval

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

The training set relies on the already publicly available dataset, for which the cantonal ethics committee of Bern (Switzerland) approved the studies and waived written informed consent.

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)

Public training set: CC BY-NC
Test set is not published

Code availability

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

The evaluation code for the challenge will be made publicly available on GitHub.

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

Participating teams have to open-source their code under a license which allows public use (preferably CC-BY license) in order to be eligible to win the challenge.

Conflicts of interest

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

The challenge may obtain dedicated sponsoring or funding, which will not have influence on challenge design, evaluation and results.
Only members of MIC@DKFZ and the clinical partners responsible for data acquisition and labeling have access to the test 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

Decision support, Longitudinal study, Research, Diagnosis

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling

- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification

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.

Patients with a previous medical record of glioma and multiple consecutive MRI scans, pre- and/or posttreatment. At least one follow-up examination.

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

Patients with a previous medical record of glioma and multiple consecutive MRI scans, pre- and/or posttreatment. At least one follow-up examination.

Imaging modality(ies)

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

MRI (T1, cT1, T2, FLAIR)

Context information

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

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

Training: The training dataset contains many additional information in addition to the MRI scans like MR acquisition parameters and time between scans. Notably automatically generated segmentations using two tools are given for each scan.
Testing: None

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

Training: The training set contains metadata on the patients regarding for example patient sex, overall survival and age
Testing: None

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 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.

Brain tumors

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)).

Precise classification of response according to RANO

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).

Training data: The training data, the LUMIERE dataset [9], stems from the Bern University Hospital (Inselspital), the pre-operative scans were acquired between 2008 and 2013, follow-ups were recorded until 2017. 95% of the 2487 provided MRI images have been acquired on Siemens scanners, 3% on Philips scanners (Philips Medical Systems/Philips Healthcare), and 2% on scanners from GE Medical Systems. Information on the respective scanner is available for all scans.
Testing data: n/a

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

Training data: Image acquisition parameters are given for all individual scans.
Testing data: n/a

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.

Training data: The training data was acquired at the Bern University Hospital (Inselspital).
Testing data: Multi-center dataset, more precise information on included centers can be shared with reviewers if

they agree not to participate in the challenge.

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).

Training data: n/a
Testing data: n/a

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.

Cases consist of two consecutive brain MRI scans of a single patient. The provided sequences are unenhanced and contrast-enhanced T1-weighted, T2-weighted and T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Training data: 91 patients with a total of 616 scans. Some scans do not contain all sequences.
Validation data: 6 patients with 2 scans each resulting in a total of 12 scans
Testing data: 100 patients with 2 scans each resulting in a total of 200 scans

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

The training data is the full publicly available dataset.
The validation and testing data is part of a private dataset. The validation set is important to gauge the performance of trained models under the distribution shift.

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.

There is a (possible) distribution shift between the development dataset on the one hand (data from a single center) testing cases on the other hand (data from several centers and scanners). In addition to the acquisition shift there might also be a shift in the class distributions as well as a population shift. Information on these will however not be made available for participants.

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.

Training dataset: Manual image annotation by one annotator
Testing dataset: Manual image annotation by two annotators

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.

Training dataset: No specific instructions
Testing dataset: No specific instructions

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.

Training dataset: expert neuroradiologist with 14 years experience
Testing dataset: n/a

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.

Training dataset: n/a
Testing dataset: Consensus discussion

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.

Training data: Scans are converted to Nifti file format and skull-stripped using the HD-BET [13] tool.
Testing data: Scans are converted to Nifti file format, skull-stripped using the HD-BET [13] tool and registered to the T1 volumes. Scans are not resampled to a common spacing in order to keep it close to the clinical workflow.

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 annotations are done by experienced radiologists. However, even though the RANO categorisation aims to make assessment objective, there is still a potential source of error related to ambiguous cases.

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)

Balanced Accuracy (BA), F1-score, TPR, TNR, AP, AUROC

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

The metrics were chosen following the guidelines from the Metrics Reloaded Framework [14] for an imbalanced dataset (BA, F1-score, AP) with further metrics added for additional insights (TPR, TNR, AUROC).

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.

AP and F1-score represent multi-threshold and counting metrics, respectively. The used ranking scheme will ensure that winning methods need to perform well on a fixed cutoff as well as on moving thresholds and outperform other methods on all classes.

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

As participants submit docker images for testing there should not occur any missing results. If an algorithm should still fail to produce a result for a specific case, we will assign the worst possible result to this case, i.e. a wrong label with lowest score for the true class.

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

Submissions will be ranked using F1-score and AP for all classes separately. The final ranking is based on the respective ranks by averaging over classes and metrics.

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.

Variability of rankings will be assessed by bootstrapping methods.

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

Bootstrapping is among the suitable methods for the assessment of ranking variability according to [15].

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or

- ranking variability.

We plan to further analyze patterns in the predictions to find hard cases in the dataset and investigate the effect of the distribution shift between training and testing data.
In addition, we will investigate ranking variability using bootstrapping.

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] Bakas, S. et al. "The International Brain Tumor Segmentation (BraTS) Cluster of Challenges"
doi: 10.5281/zenodo.7837973
[2] Bakas, S. et al. "The Federated Tumor Segmentation (FeTS) Challenge 2022"
doi: 10.5281/zenodo.6362408
[3] Zimmerer, D. et al. "Medical Out-of-Distribution Analysis Challenge 2023"
doi: 10.5281/zenodo.7845019
[4] Menze, B. et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)" IEEE Transactions On Medical Imaging 34, 1993-2024 (2015)
doi: 10.1109/TMI.2014.2377694
[5] Menze, B. et al. "Proceedings of MICCAI-BRATS 2016"
https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2016_proceedings.pdf
[6] Wen PY, Macdonald DR, Reardon DA, et al. "Updated response assessment criteria for high-grade gliomas: Response Assessment in Neuro-Oncology Working Group." Journal of Clinical Oncology 28:1963-1972 (2010)
[7] Baheti, B. et al. "The Brain Tumor Sequence Registration (BraTS-Reg) Challenge"
doi: 10.5281/zenodo.6362419
[8] Kickingereder, P., Isensee, F. et al. "Automated quantitative tumor response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study." The Lancet Oncology 20, 728-740 (2019)
https://doi.org/10.1016/S1470-2045(19)30098-1
[9] Suter, Y., Knecht, U., Valenzuela, W., Notter, M., Hewer, E., Schucht, P., Wiest, R. and Reyes, M., 2022. "The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation." Scientific data, 9(1), p.768.
[10] Isensee, F., Jaeger, P.F., Kohl, S.A.A. et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature Methods 18, 203-211 (2021).
https://doi.org/10.1038/s41592-020-01008-z
[11] HD-GLIO-AUTO: https://github.com/NeuroAI-HD/HD-GLIO-AUTO
[12] DeepBraTumIA: https://www.nitrc.org/projects/deepbratumia/
[13] Isensee, F., Schell, M., Tursunova, I. et al. "Automated brain extraction of multi-sequence MRI using artificial neural networks." Human Brain Mapping 40, 4952-4964 (2019)
https://doi.org/10.1002/hbm.24750
[14] Maier-Hein, L., Reinke, A. et al. "Metrics reloaded: Pitfalls and recommendations for image analysis validation" ArXiv:2206.01653 [Cs] (2022)
https://doi.org/10.48550/arxiv.2206.01653
[15] Maier-Hein, L., Eisenmann, M., Reinke, A. et al. "Why rankings of biomedical image analysis competitions

should be interpreted with care" Nature Communications 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7

Further comments

Further comments from the organizers.

N/A