Computational Precision Medicine Radiology-Pathology challenge on Brain Tumor Classification 2020: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Computational Precision Medicine Radiology-Pathology challenge on Brain Tumor Classification 2020

Challenge acronym

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

CPM-RadPath

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.

The goal of CPM-RadPath 2020 is to assess automated brain tumor (glioma) classification algorithms, when data from both radiology (MRI) and histopathology (digital pathology) imaging are used. The algorithmic performance will be evaluated based on a retrospective cohort of three types of gliomas, i.e., glioblastoma, oligodendroglioma, and astrocytoma. The significance of CPM-RadPath 2020 challenge is the integrated use of two different types of imaging, at different spatial resolutions, both of which are key in routine clinical diagnosis and management of brain tumor patients. The selection and number of features from each imaging type will be left to the participants. However, they will be required to use at least one feature from each imaging data type in their algorithms, but the decision about how information from the two imaging types is integrated is left to the participants.

This challenge will make use of 388 cases, multiparametric (mpMRI) and histopathology, collected from the same patients. The challenge will be conducted in three phases:
i. Training (70% of cases) – labels revealed,
ii. Validation (20%) – labels hidden, best of 3 submissions placed on the leaderboard,
iii. Test (10%) – labels hidden, single submission of a docker or a singularity container for the final ranking.

Products of CPM-RadPath 2020 challenge include: (1) publication of short papers from all participants who complete the validation phase, (2) submission of a journal manuscript reporting a summarized meta-analysis of the challenge outcomes and findings, co-authored by the CPM-RadPath 2020 organizers and members of participating teams, (3) public dissemination of dockerized solutions from top teams through the CPM-RadPath DockerHub site (per participants' prior agreement), and (4) designation of CPM-RadPath challenge dataset in The Cancer Imaging Archive and the Imaging Data Commons of the National Cancer Institute.

Challenge keywords

List the primary keywords that characterize the challenge.

cancer, brain tumor, tumor classification, diagnosis, radiology, digital pathology

Year

The challenge will take place in ...

2020

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

Brain-Lesion (BrainLes) Workshop

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.

Based on past CPM challenges, and our plan to start the challenge early to increase participation, we expect about 40 teams to participate in the CPM-RadPath 2020 challenge.

Publication and future plans

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

As in previous year, we plan to coordinate the publication of papers, reviewed and selected by the CPM organizers, in the Springer series on Lecture Notes in Computer Science (per prior agreement with the MICCAI BrainLes organizers). In addition, we plan on the publication of a manuscript summarizing the challenge results and collective findings from CPM-RadPath 2019 and 2020, by the challenge organizers in a peer-reviewed journal. This manuscript will have selected participants from top ranking teams as co-authors. Furthermore, we plan to place a dockerized tool from the top teams in the CPM-RadPath DockerHub (per prior agreement with the participants), shortly after MICCAI 2020. Finally, following the publication of the summarizing journal manuscript, we will provide the CPM-RadPath challenge dataset in The Cancer Imaging Archive and the Imaging Data Commons of the National Cancer Institute for public access.

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.

1 computer,
1 (or 2) projectors,
2 microphones.

TASK: Brain tumor classification using MRI scans and digital pathology slides

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.

Radiology and pathology are two important and complementary data modalities at different scales, routinely used for cancer diagnosis, progression assessment, and treatment response. Our challenge provides a unique benchmark dataset from both public and private sources, as well as a unique platform for evaluating machine learning methods focused on integrated use of radiology and pathology imaging data types.

Keywords

List the primary keywords that characterize the task.

cancer, brain tumor, tumor classification, diagnosis, radiology, digital pathology

ORGANIZATION

Organizers

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

Keyvan Farahani,
Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health

Tahsin Kurc,
Stony Brook Cancer Center

Spyridon Bakas,
University of Pennsylvania

Benjamin Aaron Bearce,
Massachusetts General Hospital

Jayashree Kalpathy-Cramer,
MGH, Harvard University

John Freymann,
Fredrick National Lab for Cancer Research

Joel Saltz,
Stony Brook Cancer Center

Eric Stahlberg,

Fredrick National Lab for Cancer Research

George Zaki,
Fredrick National Lab for Cancer Research

Expert consultants:

MacLean P Nasrallah,
University of Pennsylvania - Neuropathology

Russell Taki Shinohara,
University of Pennsylvania - Biostatistics

b) Provide information on the primary contact person.

Keyvan Farahani,
Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health
cpm2020@cbica.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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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.

http://miccai2019.eastus.cloudapp.azure.com/

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

https://www.med.upenn.edu/cbica/cpm2020.html

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 not eligible for awards and not listed in leaderboard.

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

Top 3 ranking teams will be recognized as follows:
1. NCI CPM Certificate of Merit, team presentation at MICCAI,
2. Publication of short paper in the BrainLes-CPM proceedings in LNCS,
3. Co-authorship on the peer-reviewed publication of the summarizing journal manuscript reporting on both CPM-RadPath 2019 and 2020.
In addition, similarly with last year's configuration, we are currently coordinating with NVIDIA for the possibility of sponsoring a GPU as an award to the top-performing team.

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.

During the validation phase (2nd phase) or the challenge, a live Leaderboard will announce the ranking of participants. At the end of the challenge, upon completion of the test phase (3rd phase), the top 3-ranked performers will be announced on the challenge's website.

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).

There will be two, possibly three, opportunities for publications.
1. Participants who complete the validation phase of the challenge may have their short papers included in the BrainLes-CPM LNCS proceedings,
2. At least two members of the top three performing teams may be included as co-authors in the peer-reviewed journal manuscript for CPM-RadPath 2019-2020,
3. Participants may pursue publication of their own methods and results separately, subject to a 9-month embargo after MICCAI 2020

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.

Results will be submitted in the online platform used for ranking. In addition, participants will be asked to consider offering a containerized submission of their algorithms for further evaluation on additional hidden data. One or both of the following docker submission methods will be pursued: (1) CPM platform (Azure), and (2) NCI Cloud One (AWS).

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.

Participants may pre-evaluate their algorithms during the training phase, and the validation phase (limited to best of three submissions).

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

April 15: Opening of CPM website and registration
April 22: Training phase begins
July 8: Validation phase begins
July 22: Validation phase ends
July 29: CPM short papers due
Aug 4: Test phase begins - Docker submission from top 6-ranked teams due
Aug 11: Announcement of final CPM results
Aug 18: Public dissemination of dockerized solutions by top 6 teams in DockerHub

These dates are tentative and subject to change considering the recent outbreak of the COVID-19 virus.

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 CPM-RadPath challenge will use data from The Cancer Genome Atlas (TCGA)/The Cancer Imaging Archive (TCIA) and the University of Pennsylvania. The TCGA/TCIA data were collected from multiple institutions under a Central IRB approval. The data from the University of Pennsylvania were acquired under UPenn IRB approval (available upon request).

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY.

Code availability

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

The CPM-RadPath evaluation site is based on CodaLab, an open source challenge evaluation platform. Last year's evaluation platform could be accessed at: http://miccai2019.eastus.cloudapp.azure.com/

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

The participants are required to send the output of their algorithms to the evaluation platform for the scoring to occur during all training, validation and testing phases. Furthermore, for verification purposes during the final testing phase, but also to enable reusability and public availability of the developed algorithms, the participants will be requested to submit their algorithm in the form of a docker container.
Dockerized code from the top ranked (possibly the six top-ranked) teams will be made available on the CPM-RadPath DockerHub site, subject to the participants' approval.

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.

Organizers have no conflict of interest in CPM-RadPath 2020. In the event of CPM selection by NVIDIA for a GPU card award, this information will be disclosed on the CPM-RadPath website and at the challenge session at MICCAI 2020.

Only CPM organizers will have access to the labels used in the validation and test phases of the challenge.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research, Assistance, Intervention planning, Treatment planning, Surgery, Training, Diagnosis, Decision support, Prognosis.

Additional points: Multi-disciplinary radiology and pathology data science

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 suffering from one of three types (classes) of brain gliomas.

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

Patients suffering from one of three types (classes) of brain gliomas.

Imaging modality(ies)

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

Multi-parametric MRI (mp-MRI) scans and digital pathology H&E slides.

Context information

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

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

Multi-parametric MRI (mp-MRI) scans and digital pathology H&E slides.

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

Cancer patients (male and female) presented with one of three types of brain gliomas.

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 in MRI scans, and tumor histological specimen shown in digital pathology 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 tumor

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy of classification using two data types.

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).

All CPM-RadPath data were acquired at 16 International institutions using mp-MRI and digital pathology scanners from several leading manufacturers. MRI scans were acquired on 1T-3T scanners.

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

mp-MRI (T1, T1-Gd, T2- T2-FLAIR) and corresponding digital histopathology images of the tissue that was resected during the first operation on the patient.

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 Cancer Genome Atlas (TCGA) studies were performed at many leading institutions in the United States, under a strict and carefully developed protocol, approved by the National Cancer Institute. Additional cases were acquired at the University of Pennsylvania, as part of clinical practice.

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).

All data were acquired by, or under the supervision of, board-certified neuroradiologists, neurosurgeons, and neuropathologists, with a minimum of 4 years of experience, using state-of-the-art imaging, surgical, and pathology instruments.

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.

All cases included in the CPM-RadPath 2020 challenge represent mp-MRI and digital pathology images of human brain tumors and are labeled with the tumor type, confirmed by pathology.

A case refers to all information acquired from one particular patient in TCGA or a similar clinical study. This information always includes the image information as specified in data sources and may include context information. All training, validation, and testing cases were annotated with the brain tumor type.

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

We will use a minimum of 388 cases in CPM-RadPath 2020 challenge, distributed as follows
a. Training: 270 cases
b. Validation: 78
c. Testing: 40
There is a strong possibility of adding an additional 100 cases from a private collection, by the start of the

challenge, which will be distributed proportionally across the three phases of the challenge.

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

Total number of cases was based on all of the qualifying cases in the TCGA Glioma collections and additional private cases available from UPenn. Accordingly, the specific number of cases in each phase was determined to optimize the proportion of cases relative to the total, according to real world distributions.

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

The CPM-RadPath 2020 dataset consists of multi-institutional paired radiology scans and digitized histopathology images of brain gliomas, obtained from the same patients, as well as their diagnostic classification label. Taking into consideration the latest classification of CNS tumors, the classes used in the CPM-RadPath challenge are:
- A = Lower grade astrocytoma, IDH-mutant (Grade II or III)
- O = Oligodendroglioma, IDH-mutant, 1p/19q codeleted (Grade II or III)
- G = Glioblastoma and Diffuse astrocytic glioma with molecular features of glioblastoma, IDH-wildtype (Grade IV).

Class distribution was chosen in correspondance to the distribution in the TCGA study.

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.

Sixteen board-certified neuropathologists with substantial expertise in brain tumors, and a minimum of 4 years of professional experience classified the cases.

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.

All cases were from real world clinical cases and annotations were obtained as part of the standard of care, following the TCGA protocol. Here an annotation is defined as the classification of a patient (subject) into one of the three classes, A, O, G, as described above.

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.

All cases were revisited and annotated by board-certified neuropathologists having a minimum of 4 years of professional experience, while considering the latest WHO classification scheme.

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.

Annotations were not merged but, in many cases (e.g., academic institutions), they reflect concurrence by multiple readers per institution, e.g., concurrence between a neuropathology fellow and the attending neuropathologist.

Data pre-processing method(s)

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

All imaging and pathology data in this challenge were de-identified and contain no protected health information (PHI).

The radiology data of the CPM-RadPath challenge describes multi-institutional routine clinically acquired pre-operative multi-parametric MRI (mpMRI) scans of brain gliomas. Specifically, the radiology scans used in this challenge are available as NIfTI files (.nii.gz) and correspond to mpMRI images comprising a) native (T1) and b) post-contrast T1-weighted (T1Gd), c) T2-weighted (T2), and d) T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes. All brain scans were acquired with different clinical protocols and various scanners (1T-3T) from multiple institutions. The provided data are distributed after their pre-processing, i.e. co-registered to the same anatomical template, interpolated to the same resolution (1 cubic mm), and skull-stripped.
The histopathology data of the CPM-RadPath challenge contain one digitized whole slide tissue image for each patient, captured from Hematoxylin and Eosin (H&E) stained tissue specimens. The tissue specimens were scanned at 20x or 40x magnifications. Note that there may be color and intensity variations among the images because of batch effects and other image acquisition artifacts. The images are stored in tiled tiff format. The participants can use the OpenSlide library (https://openslide.org) to read the images, or any other library of their preference.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

A potential source of error may be misclassification of tumor type at the time of surgery. However, all classifications were confirmed in post-mortem analysis.

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

Another source of error might be a mismatch between imaging (MRI) and digital pathology data for a given patient. However, the possibility of such error was mitigated largely through the use of common identifiers among MRI and pathology data corresponding to the same subject.

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)

1. F1-Score,
2. Cohen's Kappa,
3. Balanced Accuracy.

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

These metrics were chosen after extensive consultation with biostatisticians and data scientists in the organizing team (RTS, JKC, and TK) and because they are widely used in evaluating performance of algorithms and manual raters.

- As a multi-class classifier, the F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

- Cohen's Kappa addresses classification by chance. A kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance. A limitation of kappa is that it is affected by the prevalence of the finding under observation.

- Balanced Accuracy is calculated as the average of the proportion corrects of each class individually.

We chose the aforementioned statistical evaluation metrics after taking into consideration 1) the imbalanced distribution of data across the 3 classes (i.e., the 3 tumor types – Glioblastoma, Oligodendroglioma, and Astrocytoma), and notably 2) avoiding metrics that depend on the proportion of the data in each class.

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.

Participants will be ranked for each metric, and then the average if these ranks will be computed towards the final ranking of the participants.

Since there are 40 cases in the testing cohort, and each case will be evaluated using the 3 metrics described above, each participant will end up having 40 x 3 = 120 rankings. The final ranking score will be the average of all these rankings normalized by the number of teams.

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

Missing results will be considered as incorrect classifications.

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

This approach will allow us to evaluate the pair-wise statistical significance across the ranked teams.

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.

Standard statistical software packages and libraries for computing the above metrics will be used in the statistical analysis of the results and rankings.

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

We chose the aforementioned evaluation metrics after taking into consideration of 1) the imbalanced distribution of data across the 3 classes (i.e., the 3 tumor types (Glioblastoma, Oligodendroglioma, and Astrocytoma), and 2) avoiding metrics that depend on the proportion of the data in each class.

Hence, the three statistical methods, described above, were chosen after consultation with biostatisticians and data scientists in the organizing team (RTS, JKC, and TK) and because they are widely used in evaluating performance of algorithms and manual raters.

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.

In case of a tie among teams, the challenge organizers will make the final determination.

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, Akbari H, Sotiras A, Biello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Segmentation labels and radiomic features for pre-operative MRI scans of TCGA-GBM and TCGA-LGG collections. Nat. Sci Data 2017; 4:170117; doi: 10.1038/sdata.2017.117.

2. Maier-Hein L, et. al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun., 9(1):5217, 2018. doi: 10.1038/s41467-019-08563-w.

3. Maier-Hein L, Reinke A, Kozubek M, Martel A, Arbel T, Eisenmann M, Hanbuary A, Jannin P, Muller H, Ongour S, Saez-Rodriguez J, van Ginneken B, Kopp-Schneider A, Landman. BIAS – Transparent reporting of biomedical image analysis challenges. arXiv:1910.04071v3

Further comments

Further comments from the organizers.

1. The impact of this challenge is largely in using two distinctly different imaging types (i.e., MRI and digital pathology) to drive the development of ML algorithms for automated classification of brain tumors. To the best of our knowledge there are hardly any other imaging challenges that combine data in this manner. This is important because not only it incentivizes innovation in ML design, but also the task related to key areas of medicine, namely radiology and pathology. Two disciplines that for the past decade have been considering merging their practices, but only at the operational and reporting levels, not in information processing.

2. We will certainly be willing to shorten our challenge session to 1.5-2 hrs, held as merged or tandem sessions, with BraTS 2020.

3. The scope of byproducts of this challenge, including publications of (a) methods, (b) top-ranking algorithms as dockers, and (c) the datasets in a public repository, represent a broad range of information dissemination, helping to reach the full potential of a biomedical imaging challenge, in line with a recent publication by L.Maier-Hein, et al on Transparent reporting of biomedical image analysis challenges (arXiv:1910.04071v3).