MItosis DOmain Generalization Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

MItosis DOmain Generalization Challenge

Challenge acronym

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

MIDOG Challenge 2021

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.

Digital pathology is a fast-growing field that has seen strong scientific advances in recent years. Especially the accurate diagnosis and prognosis of tumors is a topic of particular interest, as documented by recent challenges on MICCAI, ICPR and elsewhere. In this context, the detection of cells undergoing division (mitotic figures) in histological tumor images receives high attention. The density of mitotic figures (mitotic count) is known to be a relevant and practical measure for tumor proliferation (growth fraction) and is thus highly relevant for tumor prognostication. It is one of the most relevant parameters used within histopathology grading and can have an influence on the decision of appropriate therapy. The current gold standard method is visual assessment by a trained pathologist. The morphology of mitotic figures strongly overlaps with similarly looking imposters. Therefore, the mitotic count is notorious for having a high inter-rater variability, which can severely impact prognostic accuracy and reproducibility. This calls for computer-augmented methods to help pathology experts in rendering a highly consistent and accurate assessment.

While previous competitions focused on the detection of mitotic figures in small sections of H&E-stained tumor tissue (e.g., the ICPR 2012 and AMIDA challenges), more recent studies moved closer to a realistic diagnostic workflow by targeting the prediction of tumor behavior on microscopy whole slide images (WSI) (e.g., the CAMELYON16/17 challenges, the PANDA challenge and MICCAI-TUPAC 16 challenge). Recent studies have shown that given a sufficient amount of high quality and high quantity annotations for tumor specimens, current deep learning-based approaches can yield performance comparable to well-trained human experts for mitotic figure identification. However, this performance severely degrades with the variability of images, caused by the tissue preparation and image acquisition. This so-called domain-shift is inevitable to some degree even for the same highly standardized laboratory due to differences in tissue handling and manual steps in specimen preparation; however is especially notable between different laboratories. Probably the most important source of a domain shift is the whole slide image acquisition due to highly variable color representation and other image parameters between different types of whole slide scanners. Therefore, naively applying machine learning algorithms

developed using images from one laboratory may lead to low performance at another laboratory. For this challenge, we digitized microscopy slides from 300 cases of suspected breast cancer at different laboratories, using six different scanner types. The target of this challenge is thus to develop strategies that lead to machine learning solutions that are invariant to this domain-shift, and work equally well, regardless of the scanner that was used for the image digitization.

The topic of this challenge is thus highly relevant in order to promote machine learning-based algorithms to a routine and widespread diagnostic use across laboratories. It will be the first challenge in the field of histopathology to compare methods of domain adaptation on a competitive, large-scale dataset from multiple scanners of different manufacturers. We also expect the results to deliver insights into domain generalization approaches on microscopy images in general.

Challenge keywords

List the primary keywords that characterize the challenge.

mitosis detection, mitotic figure, domain adaptation, domain generalization, breast cancer

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.

none.

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 predecessors (AMIDA, TUPAC), we are hoping for at least 20-25 participants due to the high relevance of the topic.

Publication and future plans

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

We plan to publish the results and insights into the successful methods in a peer-reviewed, high-impact journal (e.g., Medical Image Analysis).

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.

We plan to use grand-challenge.org for the submission of docker containers. The evaluation of the containers on the final test set will be done at the organizers' institutions.

TASK: Mitotic figure detection

SUMMARY

Keywords

List the primary keywords that characterize the task.

mitotic figure, domain generalization, domain adaptation, breast cancer

ORGANIZATION

Organizers

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

- Marc Aubreville, Technische Hochschule Ingolstadt, Germany
- Christof Bertram, Institute of Pathology, University of Veterinary Medicine, Vienna, Austria
- Mitko Veta, Medical Image Analysis Group, TU Eindhoven, The Netherlands
- Robert Klopfleisch, Institute of Veterinary Pathology, Freie Universität Berlin, Germany
- Nikolas Stathonikos, Pathology Department, UMC Utrecht, The Netherlands
- Katharina Breininger, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
- Natalie ter Hoeve, Pathology Department, UMC Utrecht, The Netherlands
- Francesco Ciompi, Computational Pathology Group, Radboud UMC Nijmegen, The Netherlands
- Andreas Maier, Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

b) Provide information on the primary contact person.

Marc Aubreville (marc.aubreville@thi.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.

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.

We intend to use grand-challenge.org

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

https://imi.thi.de/midog/

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 not participate.

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

Currently there are no awards. We are currently reaching out to potential sponsors in this regard.

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.

Results will be announced during the challenge workshop (i.e., after the submission deadline)

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).

We aim to publish a summary of the challenge in a peer-reviewed journal (e.g. Medical Image Analysis). Participants are requested to publish a description of their method and results on arxiv.org together with their submission. The first and last author of that paper will qualify as authors in the summary paper. Participating teams are free to publish their own results in a separate publication.

Participants may publish papers including their official performance on the challenge data set, given proper reference of the challenge. There is no 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.

Participants have to submit their algorithms in the form of a Docker container, which will be applied to the test set by the organizing team to evaluate the performance on the test set. An exemplary docker container using a

reference algorithm will be provided by the challenge organizers. Fine-grained instructions will be published on the challenge website. Once the preliminary test set becomes available, there will also be an automatic process (run on the submitted containers) that evaluates the methods on the preliminary test set, which allows participants to check their results for validity.

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 will be providing the possibility to test on a preliminary test set, approximately 2 weeks before the submission deadline. Participants are allowed 100 submissions of docker containers to be evaluated on this preliminary set. The last submitted container is used for final evaluation. Results on the hold-out test set will be provided per scanner for the participants.

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

- March 15th, 2021: Go-Live of the challenge, registration open for participants
- April 1st, 2021: Availability of training data and dataset description
- August 14, 2021: Deadline for registration of participants
- August 15, 2021: Availability of preliminary test set
- August 31, 2021: Deadline for docker container submission and for two-page arxiv abstract submission
- Sept 27-Oct 1st, 2021: Announcement of results at MICCAI 2021

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 have filed for approval with the IRB of UMC Utrecht for this study (inquiry number TCBio 20-776). Due to less frequent meeting, the inquiry is still pending, but we expect positive feedback.

Data usage agreement

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

Examples:

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

CC BY NC ND.

Additional comments: Training data is licensed as CC BY-NC-ND (Attribution-NonCommercial-NoDerivs), i.e. everyone (also non-participants of the challenge) are free to use the training data set in their respective work, given attribution in the publication.

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.

Access to the evaluation software will be made available through github. Evaluation software may be updated during or after the challenge participation phase if factual errors or bugs are found.

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

We will provide links to the docker containers the participants submitted, given their consent. Participants will be encouraged to give permission for this and make their code publicly available.

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.

Grants of the organizers:
C.A. Bertram: Dres. Jutta & Georg Bruns-Stiftung für innovative Veterinärmedizin
Access to the test case labels will only be available to the organizers only, and on a need-to-know basis to perform the evaluation.

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, Prognosis, Decision support, Diagnosis, Research, Education.

Task category(ies)

State the task category(ies).

Examples:

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

Detection.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort consists of patients with confirmed breast cancer and with biopsies taken for histopathologic assessment.

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

The challenge cohort consists of 300 cases of breast cancer, with a tissue area of 2mm2 each, collected at the UMC Utrecht (The Netherlands). The cases are selected sequentially from 2017 and 2018 according to the following inclusion criteria:

- Breast cancer excision
- There is a pathology report including the mitotic count
- Patients did not opt out for their data to be used in research projects

Imaging modality(ies)

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

Histopathology images, hematoxylin and eosin-stained, whole slide image acquisition by six different types of whole slide image scanners.

Context information

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

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

We do not provide additional information to the imageset.

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

As this is not relevant for the task, we do not provide meta data about the patient.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The data consists of image patches, selected from whole slide image scans of human breast cancer tissue, acquired with different scanners of different manufacturers and at different hospitals.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm target is the detection of mitotic figures, i.e. cells undergoing cell division. The algorithm shall generalize to multiple scanners, out of which two are not disclosed to the participants.

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, Robustness, Precision.

Additional points: Find a mitotic figure detection algorithm that has high domain-invariance to the scanner used for the image acquisition. Algorithmic performance will be measured in the overall F1 score over all images and

scanners of the test set.

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 images are acquired using six different scanners:
SCANNER No. 1: Hamamatsu XR nanozoomer 2.0
SCANNER No. 2: Hamamatsu S360 (0.5 NA)
SCANNER No. 3: Aperio CS2
SCANNER No. 4: Leica Aperio GT450
SCANNER No. 5+6: Are not being disclosed at this point in time, since it is part of the challenge to adapt to an unknown scanner.

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

Acquisition parameters:
SCANNER No. 1: resolution: 0.23 microns / px at 40x
SCANNER No. 2: resolution: 0.23 microns / px at 40x, Numerical aperture of lens: 0.5
SCANNER No. 3: resolution: 0.25 microns / px at 40x, Numerical aperture of lens: 0.75
SCANNER No. 4: resolution: 0.26 microns / px at 40x, Custom optics by Leica Microsystems for native 40x scanning with 1 mm FOV (Field of View).
SCANNER No. 5+6: not disclosed to keep scanners unknown, but similar resolution to scanners 1-4 and also with 40x magnification.

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.

All data is provided from the histopathologic archive of UMC Utrecht.

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).

The regular diagnostic workflow for H&E-stained histopathology images from UMC Utrecht was applied. We expect no relevant dependency on the technician or the personnel performing tissue excision.

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.

Training and test cases both are given as a section of a whole slide microscopy images. Cases do not necessarily represent individual patients, but individual tumor cases. This holds true for the training as well as for the test set. We ensure that patients from the training set will not be part of the test set and vice versa.

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

Training set: 150 annotated cases from four scanners (Scanner No. 1, Scanner No. 2, and Scanner No. 3, 50 cases each). Additionally, 50 cases from a third scanner (No. 4) will be provided without annotations for unsupervised domain adaptation approaches.

Preliminary test set: The preliminary test set consists of 20 annotated cases from four scanners (5 each from scanners No. 1 and No. 4, 5 each from two from undisclosed scanners No.5 and No.6).

The test set consists of 80 annotated cases from four scanners (No. 1, No. 4, No. 5 and No.6 as in the preliminary test set, 20 each).

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

The challenge dataset represents a good trade-off between capturing the naturally occurring variability of whole slide histopathology images and time invested for the annotation. The selected proportion of training, validation (preliminary test) and test cases allows for a realistic estimation of the performance between preliminary test and test set. At the same time, it allows to test how robust the methods are with respect to domain shift across scanners. Additionally, it represents the largest annotated mitotic figure data set on human breast cancer, exceeding the currently largest (TUPAC) by a factor of two in terms of cases with mitotic figure annotations.

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 training, validation- and test split was performed completely at random for the scanners that are part of the test and the training set. The distribution of grades is normalized across the dataset. We thus assume that there is no statistically significant difference between the sets besides the one caused by different scanners being used.

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.

The ground truth is derived as the result of a two-stage annotation process [1]. Initially, a pathologist will manually screen all the images for mitotic figures twice. Additionally, the pathologist will also annotate mitotic figure

mimics, i.e. cells with similar visual appearance that the pathologist does not assume to be mitotic figures. After this initial process, the annotations are thus of two classes: the positive (mitotic figures) and negative (non-mitotic figures) cell annotations. As a second step, the annotations will be class-blinded and handed to a second pathologist, who then blindly has to assign another label to each annotated object. In case of agreement, the object is directly accepted as being a mitotic figure or a non-mitotic cell. For each disagreed case (one label positive for mitotic figure, another label negative), a third expert will render the final vote, not knowing, however, the individual labels of the first two experts.

As second stage, to minimize the risk of mitotic figures being missed in the initial process, we train a state-of-the-art object detection network (typically a RetinaNet-derivative), and derive a list of additional (previously not annotated) mitotic figure candidates using a cross-validation scheme. We use low cutoffs to achieve high sensitivity and low specificity. All candidate cells will then be judged by the two pathology experts, and, in case of disagreement, also by a third expert.

[1] https://www.nature.com/articles/s41597-019-0290-4

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.

Mitotic figure identification is a routine task for all of the expert annotators. No special instructions were given. The annotators are familiar with the annotation procedure, as it has been used for multiple data sets already.

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 subjects involved with the annotation process are professional pathologists with 3+ years of experience in identification of mitotic figures.

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.

See above. We define annotations for mitotic figures as belonging to the same cell if their centers are within 7.5 microns (approx. 30 px) of each other.

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.

We defined a region of interest from each whole slide image and then extracted TIFF images containing only the size of 2mm2, as would be used in a typical diagnostic workflow. We performed no additional preprocessing.

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 following errors are plausible to occur:

- Missed annotation: Especially for atypical mitotic figures, the machine-learning aided setup could fail to provide a remedy for this. In previous work, we estimated the error of potentially missed mitotic figures to be below 5 percent, due to the highly diligent and algorithmically aided annotation process
- Classification error: If both pathologists independently and erroneously assign the wrong label (either mitotic figure or non-mitotic figure), it is taken as a consensus label. We previously used a clustering approach [1] to allow for re-assessment of potentially misassigned labels. The potential error was identified to be in the area of < 3% [1].
- Bias in the perception of mitotic figures: The experts will likely have a subjective bias as to what to account as a mitotic figure and what not. However, the three-expert consensus should mitigate this issue to a degree. Wilm et al. have recently shown that involving more than three experts will likely not increase the overall annotation quality significantly. [2]

[1] https://www.nature.com/articles/s41597-019-0290-4
[2] https://arxiv.org/abs/2012.02495

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

Besides annotation errors we do not expect other sources of error.

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)

A detected object is considered to be a true positive (TP) if the Euclidean distance to a ground truth location is less than 7.5 microns (approx. 30 pixels, depending on the scanner resolution).
This value corresponds approximately to the average size of mitotic figures in the data set, and provides a reasonable tolerance for misalignment of the ground truth location and the detection. All detections not within 7.5µm of a ground truth annotation are counted as false positives (FP). All ground truth objects without detection within a proximity of 7.5 microns are considered false negatives (FN).

The data set does not contain multiple annotations for the same mitotic figure. We expect participants to also run non-maximum suppression to ensure a mitotic figure object is only detected once. We will count detections matching an annotation only once as true positive, multiple annotations thus lead to false positives.

We calculate the number of True Positives, False Positives, and False Negatives, and from that the total overall F1 score as the primary outcome metric.
In the same way, the F1 score for each scanner will be calculated as a secondary outcome metric.

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

We chose F1 because it combines precision and recall in a single metric. It is also the single most often used metric to compare mitotic figure performance, and has been used for the ICPR-MITOS challenges as well as the AMIDA13 and TUPAC16 challenges (mitotic figure detection subtask). For the medical application, which commonly involves estimating the density of mitotic figures, an overestimation is equally undesirable as an under-estimation. This is

why a symmetric measure like F1 is suitable.

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.

We calculate the F1 score (F1 = 2*TP / (2 * TP + FN + FP)) on the complete test set (i.e., for all four scanners) as the primary metric used for ranking.

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

Since docker containers need to be submitted, the challenge organizers will make sure the algorithms are run on every case of the test set.

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

The goal of the challenge is domain generalization, and over-estimation is equally bad as under-estimation, which justifies the use of the F1 score. We calculate the F1 score from the totality of two entirely unknown scanners, one scanner where only images were known a priori, and one scanner that was part of the training set. This gives an emphasis on the generalization of the method.

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.

We will be using bootstrapping on the test set, and calculate the 95% CI for the evaluation metrics.

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

Bootstrapping is used to assess the variability of results given a different set of cases. This aims at estimating the error for a new test set drawn from the same distribution.
This allows us to report a 95% confidence interval for the F1 score the participants submitted.

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 will perform an in-depth analysis of inter-algorithm performance, depending on the actual scanner and the case. We will report F1 score, precision and recall for each individual case and for each participant. This allows us to showcase strengths and weaknesses of the participating approaches.

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.