Cross-Organ and Cross-Scanner Adenocarcinoma Segmentation: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Cross-Organ and Cross-Scanner Adenocarcinoma Segmentation

Challenge acronym

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

COSAS

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 field of digital pathology has witnessed remarkable advancements, particularly around accurate tumor diagnosis and quantitative assessment. Recent challenges hosted at conferences like MICCAI, ISBI, and others underscore the escalating interest in this field. Gland segmentation, particularly in adenoma regions, has emerged as a central focus due to the widespread occurrence of adenocarcinomas in bodily tissues like the breast, colon, and lung. Key challenges such as CAMELYON16/17, GlaS, DigestPath, have played a pivotal role in propelling segmentation algorithms forward in digital pathology.

Despite these advancements, the efficacy of current algorithms encounters a significant challenge due to the inherent diversity present in digital pathology images and tissues. The variances arise from diverse organs, tissue preparation methods, and image acquisition processes, resulting in what is termed as domain-shift. This phenomenon markedly impacts the performance of machine learning algorithms when transitioning from one organ or laboratory to another, even in instances where tissues exhibit similar morphological features and adhere to standardized tissue preparation norms. Consequently, research on domain adaptation and domain generalization for pathological images has gained momentum, exemplified by competitions like MIDOG 21/22.

In this challenge, we have digitally captured over 800 patch pathology images spanning seven distinct adenocarcinomas, utilizing six different types of scanners. The primary objective is to formulate strategies that empower machine learning solutions to be robust against domain-shift, ensuring consistent performance across diverse organs and scanners employed in image digitization.

This challenge holds profound significance in advancing the development of machine learning-based algorithms for routine diagnostic applications in laboratories. Particularly noteworthy, it marks the inaugural challenge in

histopathology, offering a platform for the comparative evaluation of domain adaptation methods on a competitive, extensive dataset featuring different organs and scanners from various manufacturers. Anticipated outcomes encompass valuable insights into domain generalization approaches applicable to pathology images at large.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

histopathology, adenocarcinoma, segmentation, domain generalization

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.

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.

50.
In 2018, we hosted the DigestPath challenge in MICCAI 2018. The event attracted more than 200 participant teams, and 32 teams submitted at least one docker container for the final evaluation. We expect around 50 teams to participate in the Challenge.

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.

As with DigestPath, we aim to use grand-challenge.org for the evaluation. For this, participants have to upload docker containers with their algorithm to the platform, where these will be run.

TASK 1: Cross-Organ Adenocarcinoma Segmentation

SUMMARY

Abstract

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

This task centers around assessing the generalization capacity of machine learning algorithms in adenocarcinoma segmentation tasks across various organs. It marks the inaugural competition task specifically designed to evaluate the generalization performance of algorithms in the context of multi-organ segmentation. The training set will consist of images depicting three distinct adenocarcinomas along with their corresponding annotations. The test set, on the other hand, will encompass images of five different adenocarcinomas. All the data will originate from the same laboratory and undergo digitization using an identical scanner. Participants are restricted to utilizing only the images and labels provided by the challenge organizers. Furthermore, pre-trained networks are permissible only if they have been trained on conventional, non-medical image datasets such as ImageNet or MS COCO.

Keywords

List the primary keywords that characterize the task.

cross-ogran, adenocarcinoma, segmentation, domain generalization

ORGANIZATION

Organizers

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

Qian Da, Pathology Department, Ruijin Hospital, China
Chaofu Wang, Pathology Department, Ruijin Hospital, China
Yanfei Zuo, Histo Pathology Diagnostic Center, China
Yan Guo, Histo Pathology Diagnostic Center, China
Guofu Jiang, Histo Pathology Diagnostic Center, China
Linlin Shen, Shenzhen University, China
Xiaoling Luo, Shenzhen University, China
Meidan Ding, Shenzhen University, China
Jingxin Liu, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China
Xianxu Hou, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China
Jionglong Su, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China
Angelos Stefanidis, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China

b) Provide information on the primary contact person.

Jingxin Liu (Jingxin.Liu@outlook.com, Jingxin.Liu@xjtlu.edu.cn)

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).

N/A

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 of top 3 performing methods 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 as a short paper (3 pages) together with their submission. The first and last authors 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 also 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.

Algorithm container submission (type 2) on Grand Challenge.

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 offer the opportunity for participants to conduct tests on a preliminary test set, approximately 2 weeks before the submission deadline. Each participant is permitted to submit 1 docker container per day for evaluation on this preliminary set. Any attempts to circumvent this restriction may result in exclusion from the challenge. Following the preliminary evaluation phase, participants must submit a final container for assessment on the final test set.

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

- May 6, 2024: Go-Live of the challenge, registration open for participants

- Jun 17, 2024: Availability of training data and dataset description

- August 18, 2024: Deadline for registration of participants

- August 19, 2021: Availability of preliminary test set

- Sep 2, 2024: Deadline for docker container submission

- Sep 6, 2024: Deadline for three-page arxiv abstract submission

- Oct 6 - Oct 10, 2024: Announcement of results at MICCAI 2024

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 Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. 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

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 metric calculation code will be made available at the same time as the training data.

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. 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.

The organizers declare no conflict of interest. Access to the test case labels will 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

Diagnosis, Prognosis, Research, Training, Education.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction

- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation

Cohorts

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

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

The target cohort consists of patients with confirmed 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 task cohort consists of about 390 cases of 6 different adenocarcinomas collected at the Shanghai Jiao Tong University School of Medicine affiliated with Ruijin Hospital.

Imaging modality(ies)

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

Histopathology images, hematoxylin and eosin-stained, whole slide image digitalization by a single type of whole slide image scanner.

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 along with the image set.

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 dataset comprises image patches extracted from whole slide image scans of six distinct human cancer tissues (as listed below), all obtained using the same whole slide image scanner.

A. Gastric adenocarcinoma: Tubular adenocarcinoma
B. Ampullary adenocarcinoma: Tubular adenocarcinoma
C. Pancreatic ductal adenocarcinoma
D. colorectal adenocarcinoma: adenoma-like adenocarcinoma
E. gallbladder adenocarcinoma: Adenocarcinoma, intestinal type
F. Duodenal adenocarcinoma

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 segmentation of gland regions. The algorithm shall generalize to 6 different adenocarcinoma types, out of which three 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)).

Accuracy, Robustness.

Find a segmentation algorithm that has high domain-invariance to different adenocarcinoma types. Algorithmic performance will be measured in the Dice Similarity Coefficient (DSC) and mean Intersection of Union (mean IoU) over all images 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).

TEKSQRAY SQS-600P

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

x20 objective 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 histopathology archive of the Shanghai Jiao Tong University School of Medicine affiliated with Ruijin Hospital. For this task, we keep the morphological variance between different organs as the only domain shift reason.

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 Ruijin Hospital 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 an image patch cropped from a whole slide image. 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 whole slide images 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: 210 annotated cases (3 organs: A, D, E; 70 cases each).

Preliminary test set: The preliminary test set consists of 80 annotated cases of 4 types of adenocarcinomas, 2 types have appeared in the training set, and the remaining 2 types are unseen. (4 organs: B, C, D, E; 20 cases each).

The test set: 100 annotated cases of 5 types of adenocarcinomas, 2 types have appeared in the training set, and the remaining 3 types are unseen. (5 organs: A, B, C, D, F; 20 cases each).

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

Training set: a relatively large number of data for training a reliable model.

Preliminary test set: a relatively small number of data for model validation and to ensure the fairness of the challenge.

The test set: a relatively large number of data for a fair final leaderboard.

At the same time, the proportion allows to test how robust the methods are with respect to domain shift across

different organs.

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 images are randomly divided into training, validation, and test sets, to ensure that tumor distributions conform to real-world distributions.

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.

Four experienced pathologists manually annotate normal glands and adenocarcinoma regions into two classes at the pixel level. Then, two expert pathologists review the annotations.

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.

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 2+ years of experience in the identification of gland and tumor regions.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

No aggregation was applied.

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 PNG image patches. 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 possible to occur:

1. Classification error: Annotation errors may occur when the annotating pathologist and two reviewing pathologists independently misclassify tumor areas as normal or perform opposite identifications. However, the probability of such errors transpiring is highly unlikely.

2. Inaccurate boundaries: Pixel-level region annotation often suffers from inaccurate boundaries. However, this error is acceptable and does not affect the training, testing and evaluation of the model.

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)

The Dice Similarity Coefficient (DSC) and mean Intersection of Union (mean IoU) will be used to assess the segmentation performance.

In the preliminary test phase, to allow quick feedback to the participating teams, a weighted average of two metrics will be returned and displayed on the leaderboard. In the final testing phase, the results will be evaluated based on the ranking of all submissions across multiple metrics.

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

DSC and mIoU are the classical validation metrics for semantic segmentation and medical image segmentation tasks, which provide a fair, sensitive, and stable ranking to assess the segmentation accuracy.

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.

The objective of the challenge is domain-generalized semantic segmentation, two evaluation metrics of mean DSC and mean IoU will be used. We compute the weighted mean of DSC and meanIoU across complete test images of 6 distinct adenocarcinomas, emphasizing the method's generalization capabilities.
Specifically, the Preliminary Test Set has a weight of 0.4, while the Test Set weight is 0.6; the result on known organs has a weight of 0.4, while the weight of unknown organs' result is 0.6.
Each team's submission will be ranked by the following evaluation metrics separately first. The average rank of the two evaluation metrics of each team will be used as the overall rank of each team.

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.

As a domain generalisation challenge task, the weight on unseen organs is appropriately increased in the ranking. Since the participants had enough time to fine-tune their models on the preliminary test set, we appropriately increased the weight of the final test set.

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.

When data is missing or reporting errors, organizers need to help participants to complete the missing results in the first place. We will calculate the 95% CI for the evaluation metrics.

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

The challenge organizers will make sure the algorithms are run on every case of the test set. 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.

N/A

TASK 2: Cross-Scanner Adenocarcinoma Segmentation

SUMMARY

Abstract

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

This task focuses on evaluating the generalization capabilities of machine learning algorithms in adenocarcinoma segmentation across diverse whole slide image scanners. It represents the inaugural competition task specifically designed to assess the generalization ability in the context of a multi-scanner segmentation task. The training set will consist of images featuring invasive breast carcinoma of no special type with annotations, and captured by three different scanners. In the test set, images of the same adenocarcinoma will be provided, but captured by five different types of scanners. All data originates from the same laboratory, and participants are limited to using only the images and labels provided by the challenge organizers. Additionally, the use of pre-trained networks is allowed only if they have been trained on standard general-purpose (non-medical) image datasets such as ImageNet or MS COCO.

Keywords

List the primary keywords that characterize the task.

cross-scanner, adenocarcinoma, segmentation, domain generalization

ORGANIZATION

Organizers

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

Qian Da, Pathology Department, Ruijin Hospital, China
Chaofu Wang, Pathology Department, Ruijin Hospital, China
Yanfei Zuo, Histo Pathology Diagnostic Center, China
Yan Guo, Histo Pathology Diagnostic Center, China
Guofu Jiang, Histo Pathology Diagnostic Center, China
Linlin Shen, Shenzhen University, China
Xiaoling Luo, Shenzhen University, China
Meidan Ding, Shenzhen University, China
Jingxin Liu, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China
Xianxu Hou, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China
Jionglong Su, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China
Angelos Stefanidis, school of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, China

b) Provide information on the primary contact person.

Jingxin Liu (Jingxin.Liu@outlook.com, Jingxin.Liu@xjtlu.edu.cn)

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).

N/A

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 of top 3 performing methods 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 as a short paper (3 pages) together with their submission. The first and last authors 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 also 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.

Algorithm container submission (type 2) on Grand Challenge.

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 offer the opportunity for participants to conduct tests on a preliminary test set, approximately 2 weeks before the submission deadline. Each participant is permitted to submit 1 docker container per day for evaluation on this preliminary set. Any attempts to circumvent this restriction may result in exclusion from the challenge. Following the preliminary evaluation phase, participants must submit a final container for assessment on the final test set.

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

- May 6, 2024: Go-Live of the challenge, registration open for participants

- Jun 17, 2024: Availability of training data and dataset description

- August 18, 2024: Deadline for registration of participants

- August 19, 2021: Availability of preliminary test set

- Sep 2, 2024: Deadline for docker container submission

- Sep 6, 2024: Deadline for three-page arxiv abstract submission

- Oct 6 - Oct 10, 2024: Announcement of results at MICCAI 2024

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 Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. 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

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 metric calculation code will be made available at the same time as the training data.

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. 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.

The organizers declare no conflict of interest. Access to the test case labels will 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

Diagnosis, Prognosis, Research, Training, Education.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction

- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation

Cohorts

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

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

The target cohort consists of patients with confirmed 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 task cohort consists of about 390 cases of invasive breast carcinoma of no special type collected at the Shanghai Jiao Tong University School of Medicine affiliated with Ruijin Hospital.

Imaging modality(ies)

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

Histopathology images, hematoxylin and eosin-stained, whole slide image digitalization by 6 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 along with the image set.

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 dataset comprises image patches extracted from whole slide image scans of invasive breast carcinoma tissues, acquired with six different scanners (as listed below) of different manufacturers.

A. TEKSQRAY SQS-600P
B. KFBIO KF-PRO-400
C. LEICA Aperio GT450
D. MOTIC EasyScan One
E. 3DHISTECH pannoramic 250
F. 3DHISTECH pannoramic 1000

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 segmentation of gland regions. The algorithm shall generalize to invasive breast carcinoma and multiple scanners, out of which three of those scanners 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)).

Accuracy, Robustness.

Find a segmentation algorithm that has high domain-invariance to different scanner types. Algorithmic performance will be measured in the Dice Similarity Coefficient (DSC) and mean Intersection of Union (mean IoU) over all images 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).

A. TEKSQRAY SQS-600P
B. KFBIO KF-PRO-400
C. LEICA Aperio GT450
D. MOTIC EasyScan One
E. 3DHISTECH pannoramic 250
F. 3DHISTECH pannoramic 1000

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

x20 objective magnification for all 6 types of scanners

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 histopathology archive of the Shanghai Jiao Tong University School of Medicine affiliated with Ruijin Hospital. For this task, we keep the digitalization with different WSI scanners as the only domain shift reason.

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 Ruijin Hospital 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 an image patch cropped from a whole slide image. 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 whole slide images 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: 210 annotated cases (3 types of scanners; A, B, C; 70 cases each).

Preliminary test set: The preliminary test set consists of 80 annotated cases scanned by 4 types of scanners, 2 types have appeared in the training set, and the remaining 2 types are unseen. (4 types of scanners: A, B, D, E; 20 cases each).

The test set: 100 annotated cases of 5 types of scanners, 2 types have appeared in the training set, and the remaining 3 types are unseen. (5 types of scanners: A, C, D, E, F;20 cases each).

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

Training set: a relatively large number of data for training a reliable model.

Preliminary test set: a relatively small number of data for model validation and to ensure the fairness of the challenge.

The test set: a relatively large number of data for a fair final leaderboard.

At the same time, the proportion allows to test how robust the methods are with respect to domain shift across different scanners.

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 images are randomly divided into training, validation, and test sets, to ensure that tumor distributions conform to real-world distributions.

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.

Four experienced pathologists manually annotate normal glands and adenocarcinoma regions into two classes at the pixel level. Then, two expert pathologists review the annotations.

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.

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 2+ years of experience in the identification of gland and tumor regions.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

No aggregation was applied.

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 PNG image patches. 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 possible to occur:

1. Classification error: Annotation errors may occur when = the annotating pathologist and two reviewing pathologists independently misclassify tumor areas as normal or perform opposite identifications. However, the probability of such errors transpiring is highly unlikely.

2. Inaccurate boundaries: Pixel-level region annotation often suffers from inaccurate boundaries. However, this error is acceptable and does not affect the training, testing and evaluation of the model.

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)

The Dice Similarity Coefficient (DSC) and mean Intersection of Union (mean IoU) will be used to assess the segmentation performance.

In the preliminary test phase, to allow quick feedback to the participating teams, a weighted average of two metrics will be returned and displayed on the leaderboard. In the final testing phase, the results will be evaluated based on the ranking of all submissions across multiple metrics.

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

DSC and mIoU are the classical validation metrics for semantic segmentation and medical image segmentation tasks, which provide a fair, sensitive, and stable ranking to assess the segmentation accuracy.

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.

The objective of the challenge is domain-generalized semantic segmentation, two evaluation metrics of mean DSC and mean IoU will be used. We compute the weighted mean of DSC and meanIoU across complete test images from 6 different types of scanners, emphasizing the method's generalization capabilities.
Specifically, the Preliminary Test Set has a weight of 0.4, while the Test Set weight is 0.6; The result on known scanners has a weight of 0.4, while the weight of unknown scanners' result is 0.6.

Each team's submission will be ranked by the following evaluation metrics separately first. The average rank of the two evaluation metrics of each team will be used as the overall rank of each team.

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.

As a domain generalisation challenge task, the weight on unseen scanners is appropriately increased in the ranking. Since the participants had enough time to fine-tune their models on the preliminary test set, we appropriately increased the weight of the final test set.

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.

When data is missing or reporting errors, organizers need to help participants to complete the missing results in the first place. We will calculate the 95% CI for the evaluation metrics.

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

The challenge organizers will make sure the algorithms are run on every case of the test set. 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.

N/A

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

Further comments

Further comments from the organizers.

N/A