Automated Gleason Grading Challenge 2022: Structured description of the challenge design

Remark: This challenge have been slightly modified. All changes are highlighted in red.

CHALLENGE ORGANIZATION

Title

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

Automated Gleason Grading Challenge 2022

Challenge acronym

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

AGGC22

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.

Prostate cancer is characterized by an abnormal growth of cells in the prostate glands. It is also the second most common cancer among men worldwide which kills 1 in 40 men annually. The progression of it is determined according to the Gleason grading system, which also serves as a guide to decide the appropriate treatment a patient should receive.
In diagnosis, tissue samples are first obtained during prostate biopsies and examined via visual inspection by pathologists. Unfortunately, manual examination can be prone to inter-observer variability even between expert pathologists. This may result in missing a cancer diagnosis or unnecessary treatment due to over-grading. Furthermore, such methods are also time consuming as differentiation between malignant and benign biopsy samples are often in extensive amounts.
Therefore, this competition aims to utilize the potential of automated deep learning systems in the diagnosis of prostate cancer, with the ultimate goal of improving prognosis and quality of life of patients. Research studies have shown support for such artificial intelligence methods in achieving pathologist-level performance, as well as improvements in speed, accuracy, and consistency of the results. This competition will also seek to define an assessment standard for machine learning-based algorithms for reported results, as well as choosing the most effective solution for future clinical trials.
In this challenge, we publish a H&E-stained whole slide image dataset of prostatectomy and biopsy specimens with pixel-level annotations performed by experienced pathologists and Gleason Score. Additionally, we also provide a set of images scanned by multiple scanners to assess the algorithm performance of handling variations caused by image digitalization. The submitted algorithm should be accurate to detect different Gleason Patterns and also generalized to process images scanned by different scanners. To the best of our knowledge, this is the first challenge in the field of digital pathology that investigate the variations caused by image scanning.

Challenge keywords

List the primary keywords that characterize the challenge.

Prostate Cancer; Digital Pathology; Deep Learning

Year

The challenge will take place in ...

2022

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.

30 teams

Publication and future plans

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

We plan to coordinate a publication of challenge results with the top five participating teams.

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.

This is an online challenge. We will run the challenge on grand-challenge.org.

TASK: Gleason Pattern Detection and Gleason Score Prediction on Prostate Histopathological Images

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.

Prostate cancer is characterized by an abnormal growth of cells in the prostate glands. It is also the second most common cancer among men worldwide which kills 1 in 40 men annually. The progression of it is determined according to the Gleason grading system, which also serves as a guide to decide the appropriate treatment a patient should receive.
In diagnosis, tissue samples are first obtained during prostate biopsies and examined via visual inspection by pathologists. Unfortunately, manual examination can be prone to inter-observer variability even between expert pathologists. This may result in missing a cancer diagnosis or unnecessary treatment due to overgrading. Furthermore, such methods are also time consuming as differentiation between malignant and benign biopsy samples are often in extensive amounts.
Therefore, this competition aims to utilize the potential of automated deep learning systems in the diagnosis of prostate cancer, with the ultimate goal of improving prognosis and quality of life of patients. Research studies have shown support for such artificial intelligence methods in achieving pathologist-level performance, as well as improvements in speed, accuracy, and consistency of the results. This competition will also seek to define an assessment standard for machine learning-based algorithms for reported results, as well as choosing the most effective solution for future clinical trials.
In this challenge, we publish a H&E-stained whole slide image dataset of prostatectomy and biopsy specimens with In this challenge, we publish a H&E-stained whole slide image dataset of prostatectomy and biopsy specimens with pixel-level annotations performed by experienced pathologists and Gleason Score. Additionally, we also provide a set of images scanned by multiple scanners to assess the algorithm performance of handling variations caused by image digitalization. The submitted algorithm should be accurate to detect different Gleason Patterns and also generalized to process images scanned by different scanners. To the best of our knowledge, this is the first challenge in the field of digital pathology that investigate the variations caused by image scanning.

Keywords

List the primary keywords that characterize the task.

Prostate Cancer; Digital Pathology; Deep Learning

ORGANIZATION

Organizers

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

Dr Tan Soo Yong
Head and Associate Professor, Department of Pathology, National University of Singapore.
Head and Senior Consultant, Department of Pathology, National University Hospital
Group Chief of Pathology, National University Health System.

Director, Advanced Molecular Pathology Laboratory, Institute of Molecular and Cell Biology, A*STAR.

Dr Weimiao Yu
Principal Investigator, Computational Digital Pathology Lab (CDPL), Bioinformatics Institute, A*STAR.
Head, Computational & Molecular Pathology Lab (CMPL), Institute of Molecular and Cell Biology, A*STAR.

Dr Lee Hwee Kuan
Senior Principal Investigator, Computer Vision and Pattern Discovery for Bioimages Group, Bioinformatics Institute, A*STAR.

Dr Loo Lit Hsin
Senior Principal Investigator, Complex Cellular Phenotype Analysis Group, Bioinformatics Institute, A*STAR.

Dr Jun Xu
Professor, Institute for AI in Medicine (AIM), Nanjing University of Information Science and Technology (NUIST).

Dr Ong Kok Haur
Research fellow, Computational Digital Pathology Lab (CDPL), Bioinformatics Institute, A*STAR.

Ms. Xinmi Huo
Research officer, Computational Digital Pathology Lab (CDPL), Bioinformatics Institute, A*STAR.

Mr. Haoda Lu
Non-graduating Ph.D student, Institute for AI in Medicine (AIM), Nanjing University of Information Science and Technology (NUIST).
Intern, Computational Digital Pathology Lab (CDPL), Bioinformatics Institute, A*STAR.

b) Provide information on the primary contact person.

Xinmi HUO (huo_xinmi@bii.a-star.edu.sg)

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

Open call challenge.

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.

grand-challenge.org

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

https://aggc22.grand-challenge.org/

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.

Publicly available data is 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.

Members of the organizers' institutes may participate in the challenge and are eligible for awards.

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

TBD

e) Define the policy for result announcement.

Examples:

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

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

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 will publish a paper to describe the challenge as well as the methods and results of the top 5 participating teams. 2-3 members from the selected teams will be invited to contribute to the manuscript and qualified as authors. The participating teams may publish their own results separately after the challenge paper is published.

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.

Detail will be announced later through our website.

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.

Multiple submissions are allowed but only last run is officially counted to compute challenge results

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

The release date of the training data: Apr 1, 2022
The registration period: Apr 1 – Apr 15, 2021
The release date of the public test data: May 30, 2022
The submission period: Jun 24 – July 22, 2022
The release date of the results: Aug 29, 2022

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).

Ethical approval from the National Health Group (NHG) Singapore has been granted.

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 SA.

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 code to produce ranking will be made available.

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

The top 3 teams will need to 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.

Only the organizers will have access to the test case labels.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Decision support, Research, Assistance, Diagnosis, CAD.

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.

Digitized prostatectomy or biopsy specimens of patients scanned by multiple scanners.

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

Subset 1: Digitized prostatectomy specimens of patients who underwent the surgery. The images were scanned by Akoya Biosciences Vectra Polaris only.

Subset 2: Digitized core needle biopsy specimens of patients who underwent biopsy. The images were scanned by Akoya Biosciences Vectra Polaris only.

Subset 3: Digitized prostatectomy specimens of patients who underwent the surgery. The images were scanned by Akoya Biosciences Vectra Polaris and scanners from Olympus, Zeiss, Leica, KFBio, Philips respectively.

Imaging modality(ies)

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

H&E stained bright-field microscopy

Context information

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

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

In training set,
Each image comes with a set of binary masks of annotations ("Gleason Pattern3", "Gleason Pattern4","Gleason Pattern5","Normal","Stroma") performed by pathologists. The number of binary masks varies from case to case. The size of the pixel is 0.5 µm/pixel for the images pathologists annotated.
The original annotations were performed on images scanned by Akoya BioSciencens. Since the images scanned by different scanners are not aligned, we perform image registration to transform the original annotation masks so that they are aligned with images scanned by different scanners.

In test set,
Subset 1 and 2: No information will be given.
Subset 3: Name of the scanner is specified in the file name of the image

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

No such information will be given.

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.

Target cohort: Digitized prostatectomy or biopsy specimens of patients scanned by any scanners.

Challenge cohort:
Subset 1: Digitized prostatectomy specimens of patients who underwent the surgery. The images were scanned by Akoya Biosciences Vectra Polaris only.

Subset 2: Digitized core needle biopsy specimens of patients who underwent biopsy. The images were scanned by Akoya Biosciences Vectra Polaris only.

Subset 3: Digitized prostatectomy specimens of patients who underwent the surgery. The images were scanned by Akoya Biosciences Vectra Polaris and scanners from Olympus, Zeiss, Leica, KFBio, Philips respectively.

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.

Detect different patterns in the images and generate binary masks for each class (Stroma, Normal, Gleason Pattern 3, Gleason Pattern 4, Gleason Pattern 5)

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.

Additional points: Find automated Gleason Grading algorithm with high accuracy across images scanned by different scanners.

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).

Scanners from Akoya Biosciences, Olympus, Zeiss, Leica, KFBio, Philips

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

Pixel size of all images as well as the binary masks are 0.5µm/pixel (20x).

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.

Glass slides are provided by National University Hospital, Singapore. Image scanning is done in multiple centers including Institute of Molecular and Cell Biology (IMCB), Olympus, Philips, Zeiss, Leica.

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).

Trained pathologists prepare the glass slides according to a predefined research protocol. Research officers scan the glass slides using different scanners.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case refers to an image of either a prostatectomy specimen or a core need biopsy scanned by any scanner.

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

Subset 1:
Training set: 105 cases of prostatectomy specimens scanned by Akoya Biosciences Vectra Polaris
Test set: 45 cases of prostatectomy specimens scanned by Akoya Biosciences Vectra Polaris

Subset 2:
Training set: 37 cases of biopsy samples scanned by Akoya Biosciences Vectra Polaris
Test set: 16 cases of biopsy samples scanned by Akoya Biosciences Vectra Polaris

Subset 3:
Training set: each prostatectomy specimen is scanned by multiple scanners (Akoya Biosciences Vectra Polaris and scanners from Olympus, Zeiss, Leica, KFBio, Philips). Each scanner scanned 26 cases except 25 for Philips and 15 for Zeiss. In total, 26*4+25*1 +15*1 = 144 cases.

Test set: each prostatectomy specimen is scanned by multiple scanners (Akoya Biosciences Vectra Polaris and

scanners from Olympus, Zeiss, Leica, KFBio, Philips). Each scanner scanned 12 cases except 7 for Zeiss. In total, 12*5+7*1 = 67 cases.

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

70%-30% train-test splitting is a rule of thumbs in Machine Learning.

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.

Even though the amount of annotated area of each class varies from case to case, the train-test ratio of annotations area of each class is also around 70%-30%.

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.

Pixel-level annotations were performed by three pathologists from NUH. Each image was annotated by one pathologist only. They used A!HistoNotes, a cloud-based annotation platform to annotate the WSIs.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Training was provided prior to the annotation process to ensure that the pathologists are skillful at A!HistoNotes.

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.

The annotation process was performed by Dr. Hue Swee Shan Susan (more than 15 years of medical experience), Dr. Lau Kah Weng (more than 10 years of medical experience), and Dr Tan Char Loo (more than 10 years of medical experience).

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.

none

Data pre-processing method(s)

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

The annotation masks are extracted from the json files downloaded from A!HistoNotes, our annotation platform.

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.

Gleason Grading is subjective in a way, therefore, the pixel-level annotations performed by one pathologist might be disagreed by other pathologists. What's more, the Gleason Score retrieved from hospital's database refers to a patient's case which might consists of multiple specimens. However, we only got one specimen for each patient. Therefore, the Gleason Score of that particular specimen might be different.

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

Freehand drawing might not be precise.

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)

Weighted-average F1-score of detection results with respect to Ground Truth pixel-level annotations: calculate the precision and recall of detected binary mask for each class. The ground-truth annotations are not comprehensive. Pathologists might not annotate all the relevant regions. Therefore, in the generated index image, only the region that is within the Ground Truth annotated region are valid for assessment.

Weighted-average F1-score = 0.25 * F1-score_G3 + 0.25 * F1-score_G4 +0.25 * F1-score_G5 +0.125 * F1-score_Normal +0.125 * F1-score_Stroma, where:
F1-score=2×Precision×Recall/(Precision+ Recall)
Precision=TP/(TP+FP)
Recall=TP/(TP+FN)

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

The algorithm is expected to assist pathologists to locate tumor area, assess tumor burden. Therefore, we should measure the performance of Gleason Pattern detection using weighted-average F1-score.

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.

Ranking by final score.
Final_score = 0.6* weighted F1-score_subset_1 + 0.2* weighted F1-score _subset_2 + 0.2* weighted F1-score _subset_3

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

Participants need to submit results for all test cases.

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

We assess the model performance of Gleason Pattern detection. Since subset 2 and 3 are much smaller than

subset 1, the performances on subset 2 and 3 are less indicative. Therefore, we give lower weight for these two subsets.

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.

Matlab or Python will be used for data analysis.

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

MATLAB and Python are powerful tools for data analysis.

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.

none

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.

Further comments

Further comments from the organizers.