The PANDA challenge: Prostate cANcer graDe Assessment using the Gleason grading system: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

The PANDA challenge: Prostate cANcer graDe Assessment using the Gleason grading system

Challenge acronym

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

PANDA

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.

With 1.1 million new diagnoses every year, prostate cancer (PCa) is the most common cancer in men in developed countries. The biopsy Gleason grading system is the most important prognostic marker for prostate cancer but suffers from significant inter-observer variability, limiting its usefulness for individual patients. The Gleason grade is determined by pathologists on hematoxylin and eosin (H&E) stained tissue specimens based on the architectural growth patterns of the tumor.

Automated deep learning systems have shown promise in accurately grading prostate cancer. Several studies have shown that these systems can achieve pathologist-level performance. A large multi-center evaluation on diagnostic data is still missing. In this challenge, we strive to improve on these works by publishing the most extensive multi-center dataset on Gleason grading as of yet. The training set consists of up to 11.000 whole-slide images of digitized H&E-stained biopsies originating from two centers. This is the largest public whole-slide image dataset available, roughly 8 times the size of the CAMELYON17 challenge. Furthermore, in contrast to previous challenges, we do not use small tissue micro-arrays, but full diagnostic biopsy images. Using a sizeable multi-center test set, graded by expert uropathologists, we will evaluate challenge submissions on their applicability to a clinically relevant task

Challenge keywords

List the primary keywords that characterize the challenge.

computational pathology, prostate cancer, Gleason grading, computer-aided diagnosis

Year

The challenge will take place in ...

2020

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

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 other challenges on the Kaggle platform we expect around 500-1000 participants in the online challenge. For the challenge workshop, we expect 75 - 100 participants.

Publication and future plans

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

We plan to publish a journal article on the challenge, the dataset, and its results, in collaboration with the top-10 ranked teams. The challenge will be left open after completion to allow for late submissions. The dataset and challenge can then be used as a benchmark for models on automated Gleason grading.

Space and hardware requirements

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

The challenge will be hosted through Kaggle and will finish in July. During the on-site event, challenge results will be presented in addition to five invited presentations by the top 10 competitors on their solutions.

Kaggle will provide resources for computation during the evaluation phase (inference only). In addition, participants will have the ability to apply for Google Cloud Credits through Kaggle to be used during the competition.

TASK: Automated Gleason grading of Prostate Biopsies

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.

See abstract of challenge.

Keywords

List the primary keywords that characterize the task.

computational pathology, prostate cancer, Gleason grading, computer-aided diagnosis

ORGANIZATION

Organizers

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

Wouter Bulten, Radboud University Medical Center, Nijmegen, The Netherlands
Geert Litjens, Radboud University Medical Center, Nijmegen, The Netherlands
Hans Pinckaers, Radboud University Medical Center, Nijmegen, The Netherlands
Peter Ström, Karolinska Institutet, Stockholm, Sweden
Martin Eklund, Karolinska Institutet, Stockholm, Sweden
Kimmo Kartasalo, Karolinska Institutet, Stockholm, Sweden
Maggie Demkin, Kaggle, USA
Sohier Dane, Kaggle, USA

b) Provide information on the primary contact person.

Wouter Bulten <wouter.bulten@radboudumc.nl>

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

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

One-time event with a fixed submission deadline in July. After the MICCAI Conference, the Challenge will be reopened for new submissions.

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.

Kaggle

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

https://www.kaggle.com/c/prostate-cancer-grade-assessment

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.

May not participate.

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

Awards will be determined beforehand and are based on the teams' positions on the private leaderboard.

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.

A public leaderboard will be available during the competition. For the final ranking, submitted methods will be evaluated on the private test set which is not accessible to participants. The final ranking of the leaderboard on the private test set will be made public before/during the MICCAI workshop. Five teams from the top 10 will be invited to present their solutions. The selection will mostly be based on ranking, but organizers can deviate to diversify in the methods presented.

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).

Our intent is to publish a challenge paper on the setup and results of the challenge. Ten teams are asked to contribute to this paper. The selection will mostly be based on ranking, but organizers can deviate to diversify in the methods presented. Of each team, two to three members are invited as authors. We ask challenge participants to respect an embargo until the challenge paper has been published. After the embargo is lifted, participants are free to publish their own results.

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.

Submissions should be submitted through the competition page on the Kaggle platform. Participants need to submit a Kaggle-kernel (similar to a Jupyter Notebook) that can run their method on Kaggle's servers. The data for the public and private leaderboard will be hidden, except for a few example cases of the public test set.

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.

A public leaderboard will be available which shows the performance of submissions on the public test set. Data used for the public leaderboard is independent of the data used for the private leaderboard.

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: Release of training and public test data.
Late July: Submission deadline.
Early October: Challenge event at MICCAI

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).

Radboudumc data: Data was anonymized for the challenge. The need for informed consent was waived by the local ethics review board (IRB number 2016-2275).

Karolinska data: Data will be anonymized for the challenge. The local Stockholm IRB approved the study (DNR 2012/572-31/1, 2012/438-31/3 and 2018/845–32).

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 method for determining the metric and ranking will be made available.

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

Upon acceptance of the challenge, we strive to obtain sponsorship to award prizes to the top 3 teams and reimburse travel costs for the workshop presenters. To be eligible to receive a prize, teams will need to make their code public under an Open Source Initiative-approved license (see www.opensource.org). In addition, participants will need to provide a write-up of their solution.

Conflicts of interest

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

Organizers from the Radboud University Medical Center are funded for this project through a grant of the Dutch Cancer Society (KUN 2015-7970).
Organizers from Karolinska Institutet are funded for this project through grants from the Swedish Research Council and the Swedish Cancer Society.

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

Screening, Diagnosis.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

Cohorts

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

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

Patients with a suspicion of prostate cancer who have been biopsied.

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

Patients with a suspicion of prostate cancer that were scheduled for a biopsy procedure. This data was collected retrospectively. The challenge cohort contains a larger percentage of positive cases then would be seen during daily practice, especially those from rarer high-grade cancers (i.e., Gleason 5). We explicitly included more high-grade cases to be able to evaluate challenge submissions on the full range of Gleason grades and to prevent the introduction of bias to lower grades.

Imaging modality(ies)

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

Whole-slide imaging / 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).

Each case is associated with a binary tumor label (Yes/No), the Gleason score (e.g., 3+4=7) and a biopsy-level grade group (e.g., 3). The test and the training set have the same type of labels.

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

All cases are from male patients who did not have any adjuvant treatment before the biopsy procedure. No further patient-specific information is given for the cases.

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.

Digitized whole-slide images of biopsies, stained using hematoxylin and eosin (H&E), taken from the prostate of patients.

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 algorithms need to determine whether cancer is present in the biopsies and if there is cancer, determine the grade. For each case, one label needs to be given by the algorithm: 0 - no cancer, 1-5 cancer & grade group (grade groups 1-5).

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)).

Agreement (measured through Cohen's kappa)

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).

Data was collected from two centers using different methods:

- Radboudumc: All slides were scanned using a 3DHistech Pannoramic Flash II 250 scanner.
- Karolinska: The slides were digitized using a Hamamatsu C9600-12 scanner and NDP.scan v. 2.5.86 software (Hamamatsu Photonics, Hamamatsu, Japan) and an Aperio ScanScope AT2 scanner and Aperio Image Library v. 12.0.15 software (Leica Biosystems, Wetzlar, Germany).

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

Radboudumc: All slides were scanned at 20x magnification (pixel resolution 0.24μm). To reduce the overall size of the dataset, and to achieve comparable resolution between datasets, the lowest magnification level will be removed resulting in images with a pixel resolution of 0.48μm. All images were converted to TIFF.

Karolinska: The pixel size at full-resolution (20X) was 0.45202 μm (Hamamatsu) or 0.5032 μm (Aperio). The resulting RGB images were stored at 8 bits per channel in NDPI (Hamamatsu) or SVS (Aperio) format. All images were converted to TIFF.

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.

Radboudumc: All cases were retrieved from the pathology archives of the Radboud University Medical Center. Patients with a pathologist report between 2012 and 2017 were eligible for inclusion.

Karolinska: All cases were retrieved from the STHLM3 study with participants from the Stockholm county, Sweden, during the years 2013-2015.

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).

Not applicable.

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.

For both the training and test set, a single case corresponds to a single biopsy (tissue specimen). Multiple cases can correspond to the same patient, but patients from the test set are independent of patients of the training set.

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

Training set: +/- 11.000 cases
Public test set: +/- 400 cases (with expert gradings)
Private test set: +/- 400 cases (with expert gradings)

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

We will publish all cases from two major studies on automated Gleason grading. Due to the wide range of growth patterns and tissue types present in prostate biopsies, a large training set is required to make pathologist-level algorithms. More information about these datasets can be found in the corresponding articles, both now also accepted for publication at the Lancet Oncology:

Karolinska: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30738-7/fulltext
Radboudumc: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30739-9/fulltext

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.

Cases were sampled based on the Gleason grade group.

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.

Radboudumc data: Case labels were retrieved from pathology reports, containing the original diagnosis for that case. Trained students read all reports, identified the biopsies, and assigned each biopsy with a label. Some minor label noise might exist in the training set, due to mistakes in the annotation process or due to inconclusive reports. The test set has been independently graded by three expert pathologists, specialized in uropathology. Through three rounds a consensus grade has been determined for every case in the test set.

Karolinska: All cases have been assessed by a single experienced pathologist. In addition to grading the cases, they were also annotated with a pen marker adjacent to cancer regions. The labels were collected from the pathology report into a database at the time of the study. Some minor label noise might exist due to this manual task. The test data were additionally graded by 3 experienced pathologists and a consensus grade has been derived (at least 2 of 3 agree in the grade). For these annotations, the approximately 200 cancer cases from STHLM3 were split in 100 + 100 slides. Two pathologists (plus the pathologists from the training/public set) graded one set each. Whenever there was a disagreement in the grade, another pathologist acted as a tiebreaker. If none of the 3 pathologists agreed the case was removed. Finally approximately 50 benign cases were added.

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.

Radboudumc data: For the training set, students were instructed to roughly outline each individual biopsy. Then, based on the description of the report, the students had to identify each biopsy and assign the label as present in the report. Students were instructed to flag cases if the report contained information regarding adjuvant treatment; these cases were later excluded from the study. If the pathologist report was inconclusive or lacked a description of individual biopsies, cases were flagged for a second review. If no match could be made in the second read, cases were excluded. For the test set, the expert pathologists graded the cases following the ISUP 2014 guidelines, as they would do during routine clinical practice.

Karolinska: By the STHLM3 study protocol there were no adjuvant treatment prior to biopsy. The pathologists were all from the ISUP board and followed the ISUP 2014 guidelines.

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.

Radboudumc data: The test set was graded independently by three pathologists who are subspecialized in uropathology. A final consensus score was determined in three rounds. The non-expert students that annotated the training set were all medical students with prior experience in annotating pathology cases.

Karolinska data: The test set was graded independently by 3 pathologists who are subspecialized in uropathology.

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.

Not applicable.

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.

Radboudumc: The original training dataset consisted of slides with multiple biopsies per slide. To make processing easier, each biopsy will be individually extracted and stored in separate files. Each file represents a single case and has one grade. The test cases are used as-is.

Karolinska: Each file represents a single case/slide and has one grade. Each slide typically consists of two sections from the same biopsy, but there is occasionally only one. In the case of cancer in the slide, one of the sections has a pen mark adjacent to the tissue where cancer is present.

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.

Radboudumc: The training set contains label noise. This label noise is introduced due to several reasons, including inconclusive pathologist reports, annotation errors, errors in the original diagnosis, disagreement between pathologists. To test the level of label noise, we let students annotate the test set with the same protocol as the training set. The labels, as determined by the students, were then compared to the consensus labels set by the experts. On grade group, the accuracy was 0.720 (quadratic weighted kappa 0.853). These values indicate a high agreement, but show the presence of label errors. Given the nature of Gleason grading and the problems of rater disagreement, handling this label noise is part of the challenge.
The test set was graded by three experts in consensus. We determined this as the best possible gold standard for this grading task. Still, due to the subjective nature of Gleason grading, some errors can still be present.

Karolinska: Similarly to the cases from Radboudumc, the Karolinska cases contain label noise due to the subjective nature of the Gleason grading system.

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

Not applicable.

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)

Quadratically weighted Cohen's kappa on Gleason grade group, with respect to the reference standard.

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

Gleason grading is a subjective task, with high inter- and intra-observer variability. Agreement between pathologists is often measured using Cohen's kappa. Moreover, most articles published on automated Gleason grading use Cohen's kappa as a value of model performance. Using this metric, challenge submissions can be more easily compared to others in the field.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Participants need to supply a case-level label. A kappa value is then computed for the full dataset by comparing the predicted labels with the reference standard. Participants are then ranked based on the kappa value (higher is better).

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

Participants need to label all cases from the test set.

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

The better the kappa value of the method, the closer the predictions of the method are to the reference standard.

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.

For the top-10 algorithms, we will compute confidence intervals on the private test set through bootstrapping. Similar as in [1,2], we will perform permutation tests to determine if algorithms are significantly better.

[1] https://www.nature.com/articles/s41746-019-0112-2

[2] https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30739-9/fulltext

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

Permutation tests were used in two similar studies on automated Gleason grading to assess algorithm performance.

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.

Not applicable.

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.

Articles describing the source of the data and methods for automated Gleason grading:

Karolinska: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30738-7/fulltext
Radboudumc: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30739-9/fulltext