Kidney Pathology Image Segmentation Challenge 2024: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Kidney Pathology Image Segmentation Challenge 2024

Challenge acronym

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

KPIs

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.

Chronic kidney disease (CKD) poses a significant health risk, causing more deaths annually than breast and prostate cancer combined. Impacting over 10% of the worldwide population, it affects upwards of 800 million individuals. Kidney biopsy, encompassing both open and percutaneous methods, is the gold standard for diagnosing and guiding the treatment of CKD.

In pathological image analysis, particularly in kidney disease, tissue segmentation is of paramount importance. The rise of deep learning has been transformative in kidney pathology image segmentation, yet it has also exposed a lack of comprehensive benchmarks for developing and evaluating these techniques. A major hurdle has been the scarcity of large-scale disease data in existing public datasets for kidney pathology segmentation, as they predominantly comprise samples from normal patients. This is mainly because the tissue samples from humans are typically obtained through needle biopsies, yielding only small tissue samples. Consequently, there's a pressing need to release extensive kidney pathology digital data spanning various CKD disease models.

In our challenge, we've expanded the dataset from CKD disease models by utilizing preclinical animal models, particularly whole kidney sections from diseased rodents. The primary rationale for using rodent data is the morphological similarity between rodent and human kidney pathologies, making them a prevalent choice in pre-clinical medical research and drug discovery. Secondly, whole kidney sections can be sourced from comprehensive disease models, providing an abundance of tissue in each whole slide image (WSI). This is a significant advantage, as a single WSI from these models can encompass more tissue content than what would be achievable from thousands of needle biopsies in human disease models, an approach that is often impractical.

The Kidney Pathology Image Segmentation (KPIs) challenge encompasses a broad spectrum of kidney disease models, including normal and multiple specific CKD conditions, derived from preclinical rodent models. As a pioneering effort in the MICCAI community, the challenge features an extensive collection of 10,000 normal and diseased glomeruli from over 60 Periodic acid Schiff (PAS) stained whole slide images. Each image includes nephrons, with each nephron containing a glomerulus, and a small cluster of blood vessels. The objective for participants is to develop algorithms that can precisely segment glomeruli at a pixel level. To the best of our knowledge, this represents the first MICCAI challenge focused exclusively on segmenting functional units in kidney pathology across various CKD disease models.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

segmentation, digital pathology, deep learning, glomeruli, CKD

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.

Medical Optical Imaging and Virtual Microscopy Image Analysis (MOVI) workshop

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

Over 20 teams.
This is difficult to estimate because this is the first time the challenge has been run. However, based on pathology segmentation challenges in related topics we expect to attract at least 20 teams to submit finalized entries to the challenge.

Publication and future plans

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

The highest-achieving teams in the challenge will have the opportunity to collaborate on a joint challenge paper. This paper, aiming for submission to a prestigious journal like IEEE Transactions on Medical Imaging, will feature concise summaries of their methodologies. The selection of teams for this collaboration will depend on their performance results, ensuring a diverse representation of methods for scientific interest. Additionally, teams will have the freedom to independently publish more detailed accounts of their methods in other forums.

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 on grand-challenge.org. All algorithms will be evaluated under the same hardware configuration. For the on-site event, the basic technical equipment to support presentations (e.g., projectors, computers, monitors, loud speakers, microphones) will be needed.

TASK 1: Diseased glomeruli 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.

Provide a summary of the challenge's purpose. This should include a general introduction to 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 extends beyond mere detection to encompass the identification of functional tissue units (FTUs) across varying tissue preparation methods. The kidney's primary functional unit, the nephron, includes both a coiled renal tubule and an intricate network of peritubular capillaries. The focus of this competition is to develop and implement a detector that is not only successful but also robust in identifying glomeruli FTUs, a critical component of the nephron.

Participants are expected to create an algorithm that can adapt to different CKD disease models and tissue conditions, ensuring accurate identification of the glomeruli in various states. This includes the ability to distinguish the glomeruli from surrounding tissue components under diverse preparation scenarios, thus demonstrating the versatility and precision of your approach. Additionally, the competition seeks solutions that are innovative in handling potential challenges such as variations in glomeruli size, shape, and structural integrity due to disease states or preparation techniques.

The evaluation criteria will not only include the accuracy of glomeruli segmentation but also the efficiency and adaptability of the algorithm in processing images from different sources and conditions. The ultimate goal is to pave the way for advancements in automated pathological analysis, enhancing both the speed and reliability of kidney disease diagnosis and research. The contribution could be a significant step towards achieving this goal, impacting both the field of medical imaging and the future of kidney disease treatment and management.

Keywords

List the primary keywords that characterize the task.

glomerulus, segmentation, disease, glomeruli

ORGANIZATION

Organizers

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

Yuankai Huo - Department of Computer Science, School of Engineering, Vanderbilt University, Nashville, USA
Haichun Yang - Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, USA

Mengmeng Yin - Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, USA
Shilin Zhao - Department of Biostatistics, Vanderbilt University Medical Center, Nashville, USA
Yu Wang - Department of Biostatistics, Vanderbilt University Medical Center, Nashville, USA
Ruining Deng - Department of Computer Science, School of Engineering, Vanderbilt University, Nashville, USA
Juming Xiong - Department of Electrical and Computer Engineering, School of Engineering, Vanderbilt University, Nashville, USA
Yucheng Tang - NVIDIA Corp, Bethesda, USA
Bennett A Landman - - Department of Electrical and Computer Engineering, School of Engineering, Vanderbilt University, Nashville, USA

b) Provide information on the primary contact person.

Yuankai Huo, yuankai.huo@vanderbilt.edu

Life cycle type

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

This contest is designed to be an open call challenge.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

N/A

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org (application currently in progress)

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

The challenge website will be set up after proposal acceptance.

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.

We are actively seeking sponsorship and anticipate being able to provide cash prizes or hardware (e.g., GPU cards) for the top 3 teams.

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.

All performance results will be made public via the challenge website.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Publication and Future Plans above. There is no embargo period.

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.

Algorithms will be accepted as Docker containers according to the technical requirements of https://grand-challenge.org.
In the testing phase, since the test dataset cannot be released, participants are required to provide their algorithms in the form of a Docker container. This container should be submitted to our designated Docker repository. Access to this repository will be granted following approval. This approach is chosen to ensure a more comprehensive evaluation of the algorithm's reproducibility in different environments, thereby confirming its robustness and reliability in practical applications.

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.

Our plan is to make a validation set available in May, giving participants the opportunity to refine their methods. When it comes to the testing phase, the rules differ slightly. Only the final submission of the Docker container by each participant will be considered for official evaluation. This last submitted version will be the one used to calculate the competition 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

Expected Website Opens: March 28, 2024
Expected Training Set Released: March 30, 2024
Expected Validation Set Released: May 1, 2024
Expected Model Submission Deadline: Aug 1, 2024
Expected Announcement of the Winner and Speaker Invitation: Sep 15, 2024
(Subject to change depending on MICCAI 2024 Deadlines)

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).

All data contributions to this study were approved by the health research institutes and research ethics committees of corresponding medical institutions.

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 evaluation and ranking code will be available after the end of the challenge.

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

Participants of the test phase of the competition will be asked to submit their models in the form of docker during the test phase. Detailed instructions for creating docker containers will be provided.

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.

This challenge is supported by NIH funding.
Yuankai Huo and clinical evaluators 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

Diagnosis, Education, Decision support, Assistance, Screening, Intervention planning, Research.

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.

Patients seeking a diagnosis of different CKD diseases and rodents for CKD research

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

The challenge cohort is composed of 60 whole slide images from 20 sacrificed rodents with three CKD disease models and normal kidney conditions.
Whole kidney WSIs are captured from 4 groups of mouse models:
1. Normal group: normal mice, sacrificed at the age of 8 weeks.
2. 5/6Nx group: mice underwent 5/6 nephrectomy, sacrificed at 12 weeks after nephrectomy (age of 20 weeks).
3. DN group: eNOS-/-/ lepr(db/db) double-knockout mice, sacrificed at age of 18 weeks.
4. NEP25 group: transgenic mice that express human CD25 selectively in podocytes (NEP25), sacrificed at 3 weeksafter immunotoxin-induced glomerular injury (age of 11 weeks).

Imaging modality(ies)

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

Whole mouse kidney with histology slides and PAS staining using whole slide image (WSI)

Context information

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

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

kidney biopsy sections under a bright field with 40x magnification

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

Whole kidney WSIs are captured from 4 groups of mouse models:
1. Normal group: normal mice, sacrificed at the age of 8 weeks.
2. 5/6Nx group: mice underwent 5/6 nephrectomy, sacrificed at 12 weeks after nephrectomy (age of 20 weeks).
3. DN group: eNOS-/-/ lepr(db/db) double-knockout mice, sacrificed at age of 18 weeks.
4. NEP25 group: transgenic mice that express human CD25 selectively in podocytes (NEP25), sacrificed at 3 weeksafter immunotoxin-induced glomerular injury (age of 11 weeks).

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 image data were acquired by experienced pathologists from experimental mice.

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 glomerular regions in the kidney pathology images

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)).

Dice Similarity Coefficient (DSC), F1 score

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).

Slide scanner from DigitCell Inc, KF-PRO-040

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

The Leica SCN400 Slide Scanner was used for the high-resolution imaging of whole slides in brightfield - 25 mm x 75 mm standard microscope slides. These high-capacity robotic autoloading scanners provide a high-resolution method to digitally archive histological slides.

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.

Vanderbilt University Medical Center, Nashville, USA

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).

Pathologists with at least 10 years of experience acquired the data using clinically defined protocols.

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 describes a 512x512 pixels image patch from a single kidney biopsy tissue, and each case has a pixel-wise annotated label map corresponding to glomeruli.

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

There is a total of 20,000 image cases (patches) from 60 whole slide images. We made a 60%/10%/30% split to get 12,000 training cases, 2,000 validation cases, and 6,000 testing cases.

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

The overall quantity of cases in our dataset is limited by the funding available for collection and annotation at this stage.
We have opted for a split of 60% for training and 40% for validate/test data, further divided into 60%/10%/30% respectively for a few reasons. Primarily, allocating 60% of the data for training aligns with standard practices in the machine learning community. This division is designed to provide a substantial amount of data for the initial development and tuning of algorithms, while still reserving a significant portion for validation and testing to ensure comprehensive evaluation and robustness of the models.

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 data are acquired from four CKD models, where training, validation, and test cases share a similar proportion of the four CKD models.

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.

Our approach involves the expertise of two experienced human annotators. Initially, every case will be annotated by the first pathologist, who brings over 10 years of experience to the task. Subsequently, to ensure the highest quality and accuracy of the annotations, a second pathologist with over 20 years of experience will conduct a thorough quality assurance review on all the annotated images. This two-step process is designed to provide a reliable and expertly validated dataset.

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.

Follow the paper Yao, Tianyuan, et al. "Glo-In-One: holistic glomerular detection, segmentation, and lesion characterization with large-scale web image mining." Journal of Medical Imaging 9.5 (2022): 052408-052408.

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.

QuPath v0.5.0 software is used to annotate all data

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.

Our approach involves the expertise of two experienced human annotators. Initially, every case will be annotated by the first pathologist, who brings over 10 years of experience to the task. Subsequently, to ensure the highest quality and accuracy of the annotations, a second pathologist with over 20 years of experience will conduct a thorough quality assurance review on all the annotated images.

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 original 60 whole slide images will be randomly segmented into 20,000 image patches. Each whole kidney section will contain about 10,000x20,000 pixels, which could be easily segmented to over 300 cases per WSI.

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 inter- and intra-annotator variability would be possible error resources. However, we will let the two annotators to perform an additional manual segmentation on all testing cases to compute inter- and intra-annotator variability.

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

N/A

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)

Dice Similarity Coefficient (DSC), F1 score

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

We selected the Dice Similarity Coefficient (DSC) as our main validation metric due to several key factors. Its simplicity makes it easy to understand and implement, while its popularity ensures widespread acceptance and familiarity within the community. Additionally, DSC is known for its stability in ranking, providing consistent and reliable comparisons between different methods. Most importantly, it is an effective tool for evaluating the accuracy of predictions, making it a suitable choice for our validation purposes. F1 is a supplementary metric to measure the performance at the instance level.

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 ranking method we're using is a standard approach, one that has been successfully implemented in previous challenges with positive outcomes. We are confident that this scheme is both transparent and fair to all participants, ensuring an equitable and clear evaluation process.

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

Any submission that is not complete will not be evaluated and will not count as a valid validation submission. During the final testing phase, predictions for the test set will be generated locally using the Docker containers provided by each team. As a result, we anticipate no missing results in the test set submissions.

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

For the evaluation process, we will calculate both the mean and standard deviation for each label's Dice Similarity Coefficient (DSC). The team with the highest DSC value will achieve the top rank. These metrics will be computed across all test cases, and the overall performance of each participating team will be determined by averaging their metrics. Teams will then be ranked according to these average metrics to establish their standings in the competition.

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.

The variability of the rankings will be characterized using bootstrap methods, which will be performed in Python.

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

The bootstrap method, known for its simplicity and non-parametric nature, has been effectively utilized in numerous past challenges. This approach operates on minimal assumptions, making it a reliable and straightforward choice for various applications.

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.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., ... & Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications, 9(1), 5217.

Wiesenfarth, M., Reinke, A., Landman, B. A., Eisenmann, M., Saiz, L. A., Cardoso, M. J., ... & Kopp-Schneider, A. (2021). Methods and open-source toolkit for analyzing and visualizing challenge results. Scientific reports, 11(1), 2369.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A. L., Arbel, T., Eisenmann, M., ... & Landman, B. A. (2020). BIAS: Transparent reporting of biomedical image analysis challenges. Medical image analysis, 66, 101796.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., ... & Kopp-Schneider, A. (2019).

Author Correction: Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications, 10(1), Article-number.

Armato, S. G., Drukker, K., & Hadjiiski, L. (2023). AI in medical imaging grand challenges: translation from competition to research benefit and patient care. The British Journal of Radiology, 96(1150), 20221152.

Eisenmann, M., Reinke, A., Weru, V., Tizabi, M. D., Isensee, F., Adler, T. J., ... & Maier-Hein, L. (2023). Why is the winner the best?. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp.

19955-19966).

Further comments

Further comments from the organizers.

N/A