Snke OS 3D Lung CT Segmentation Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Snke OS 3D Lung CT Segmentation Challenge

Challenge acronym

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

C19C-SnkeOsSeg

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.

Background: Since the outbreak of the global Covid19 pandemic, the number of confirmed COVID-19 cases has reached over 16 million globally [1, 2], affecting virtually every territory, and with a fatality rate ~2-3% among the cohort of PCR-positive cases. Given the high demand for effective diagnosis and treatment of cases, the WHO recently released a rapid advice guide in July 2020 [3], in which chest imaging is conditionally recommended for several purposes, e.g. to aid diagnosis in the absence/delay of PCR testing, to assess the need for ICU admission and to inform the therapeutic management of patients.

Purpose: In this challenge, we aim to aid radiologists and physicians through objective and quantitative computational assessment of chest imaging in the context of COVID-19. We provide access to a large dataset of 3D chest CT imaging of the lung, collected from several European and international radiological centers. We call the international research community to develop and test artificial intelligence algorithms on this dataset.

Dataset: We provide access to low-dose chest CT imaging volumes from a mixed cohort of COVID-19 and non-COVID-19 cases. The dataset contains 113 labeled/segmented cases (79 COVID-19, 34 non-COVID-19), and >100 unlabeled volumes. A particular scientific challenge will lie in the effective use of unlabeled data through semi- and self-supervised training techniques. Labels represent five lung lobes and two lesions types, consolidation and ground-glass opacities. Labels are provided in a multi-hot encoding to allow region overlaps (e.g. lesions within lung lobes). For local development, we provide a realistic toy dataset of 96 synthetic volumes with 4D labelmaps.

Infrastructure: To maintain privacy, the anonymized imaging data remains non-disclosed within a biobank. Participating teams can design their algorithms locally using the representative synthetic dataset. Once ready, teams can submit training and validation jobs on the real dataset through Eisen, a deep learning framework based on pyTorch. Models are trained in the cloud by sponsorship of AWS. We actively promote open science, and require all participating teams to provide their solutions open-source to the technical and medical research

community.

Participation: You can participate in two ways.

1. Hunters: Participate as a team with a maximum of 3 members as a competing team in the challenge. The incentive to the hunters: AWS cloud credits worth 7,500 EUR.
2. Rangers: Participate individually or in a team to help solve the Covid-19 challenge. You can submit tutorials, code or any educational material that is useful for the challenge. The incentive to the rangers: TBA.

Requirements: After the registration, there will be a "micro challenge" with the task of segmentation based on our synthetic toy dataset, for all teams in order to qualify for the main task.

Challenge keywords

List the primary keywords that characterize the challenge.

Covid19; CT/X-ray lung imaging; 3D segmentation; Clinical decision support systems

Year

The challenge will take place in ...

2020

TASK: Snke OS 3D Lung CT Segmentation Challenge

SUMMARY

Keywords

List the primary keywords that characterize the task.

Covid19; CT/X-ray lung imaging; 3D segmentation; Clinical decision support systems

ORGANIZATION

Organizers

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

15+ team members, 9 scientific advisors; detailed information under: https://www.covid19challenge.eu/#team;

b) Provide information on the primary contact person.

Simon Weidert (LMU, M3i; Munich, Germany; email: sw@m3i-muenchen.de)

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

One-time event with up to three sequentially announced sub-challenges, each with a fixed submission deadline.

Challenge venue and platform

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

Independent platform.

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

Custom platform (hosted by AWS): https://mxdb-compute.net/

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

https://www.covid19challenge.eu/

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.

Participants are not allowed to incorporate their own training data, e.g. via pre-trained networks or external datasets. A toy dataset with realistic volumes is provided by the challenge organizers for local model prototyping, model pre-training and/or fine-tuning of pre-trained models. A non-disclosed real-life dataset is available inside the challenge platform and accessible for training/validation/testing via the Eisen deep learning framework (www.eisen.ai).

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 from the organizing research groups (DSGZ, M3i) may participate but are not eligible for awards.

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

Hunters: The winning team of the segmentation challenge will receive a prize of 7,500 EUR of AWS computation credits, sponsored by Brainlab AG.

Rangers: The most enganged and helpful rangers will receive some incentives, TBA.

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 is hosted and maintained by the challenge organizers with team pseudonyms and performance metrics.

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).

The top 10 performing teams in each sub-challenge are invited to contribute towards a challenge publication. Each team is allowed to contribute two co-authors (chosen by the team). If a top 10 ranking team withdraws from co-authorship, the two author slots are offered to the 11th-ranking team etc., until max rank 20. Teams may publish their own results separately, the earliest allowed submission date is the challenge deadline / closing data of the leaderboard. There will be no further embargo time.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

For the pre-challenge, participants have to submit a Google Colab file.

Submission for the segmentation is done via Eisen in the submission platform https://mxdb-compute.net/. Detailed instructions are provided in the 'About' section. Incorporating different models, losses, optimizers etc. is possible in the framework, as long as they are implemented in pyTorch and derived from nn.module. A short

demo was provided in the live kickoff event: https://youtu.be/wGEQoRuSYBc?t=3247.

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.

Evaluation before the final submission is possible in several ways: 1) Unlimited evaluation on the publicly available synthetic toy dataset (an evaluation script will be made available to achieve the same evaluation metrics as featured in the leaderboard). 2) Training and validation on the Eisen/AWS-powered platform).

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

- Website goes live: March 25th 2020
- Data collection starts: March 30th 2020
- Release of toy dataset for local development: April 20th 2020, data update on August 3rd 2020
- Segmentation challenge announcement: August 3rd 2020
- Start of registration: August 3rd 2020
- Live Kickoff Event: August 17th 2020 (recording: https://youtu.be/wGEQoRuSYBc)
- Opening of job submission platform (incl. real training + validation data): August 21st 2020
- Submission data: TBA, approx. 2 months later, October 18th 2020
- Workshop day: TBA, approx. 2 weeks after submission
- On workshop day (TBA)

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).

For this challenge, we obtained ethical approval by the Ethics committee of the medical faculty of LMU Munich who supervises M3i's Digital Biobank.

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)

The data usage agreement is stated online and needs to be complied & signed when registering for the challenge (https://www.covid19challenge.eu/contact/):

1. The downloaded data or any data derived from these data are not redistributed under any circumstances (same for the link to this site or the data itself).
2. Data downloaded from this site may only be used for the purpose of preparing for the challenge and testing the system.
3. No attempt will be made to reconstruct the original datasets (reverse engineering and such) or to de-identify the datasets.
4. If there are doubts whether the intended use of the datasets is allowed or if you want to get permission to use the datasets in another context, please contact us (use the challenge's Discord server or the chat box on the website).

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 rankings and result visualization will be produced by the visualization toolkit challengeR (https://github.com/wiesenfa/challengeR).

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

All participating teams are required to provide their solutions as open-source code throughout the challenge. Each team must indicate a publicly accessible code repository (e.g. github). Rangers will get full access to the github repository containing common code.

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.

Due to the non-profit nature of the challenge and the non-disclosure of challenge data, there is no conflict of interest for participating teams. Challenge organizers have access to the test case labels throughout the challenge, and cannot compete towards awards due to the resulting conflict of interest.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Diagnosis, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation (lung lobes/GGO/CON in CT/X-ray).

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 suspected Covid19 infection with lung CT/X-ray imaging both (PCR positive and negative cases).

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

Patients with suspected Covid19 infection with lung CT/X-ray imaging both (PCR positive and negative cases).

Imaging modality(ies)

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

Low-dose lung CT.

Context information

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

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

CT labelmap volume:
1 = ILR (inferior lobe right)
2 = SLR (superior lobe right)
3 = MLR (middle lobe right)
4 = SLL (superior lobe left)
5 = ILL (inferior lobe left)
6 = GGO (ground glass opacities)
7 = CON (consolidation)

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

In the 3D segmentation challenge, no further data beyond CT volumes and multi-label volumes are provided.

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.

Chest imaging with a lung field-of-view.

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.

Segmentation of regions:
1 = ILR (inferior lobe right)
2 = SLR (superior lobe right)
3 = MLR (middle lobe right)
4 = SLL (superior lobe left)
5 = ILL (inferior lobe left)
6 = GGO (ground glass opacities)
7 = CON (consolidation)

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)).

Develop segmentation algorithms for five lung lobes and Covid19-related lesions in low-dose 3D lung CT volumes, with high sensitivity and specificity, to localize and quantify lesion extent.

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).

Different clinical CT scanners; clinical X-ray imaging units. The devices are specific to the data providing centers.

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

Low-dose chest CT. No further acqusition details available.

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.

Centers can decide whether they would like to be acknowledged. Officially contributing centers so far:

1. Professional Hospital Guaynabo (Puerto Rico) / Dr. Reynaldo Rosa (Director of Radiology), Dr Leonardo Valentin
2. Semmelweis University Budapest / Maurovich-Horvat Pál MD, PhD, MPH, Assistant Professor (Director of Radiology)
3. BG Unfallklinik Murnau / Prof. Dr. Marcus Treitl (Head of Radiology)

At least two further, non-disclosed contributing private and public radiological units.

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).

All data was acquired during clinical routine by board-certified radiologists and MTA clinical staff.

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.

In the segmentation challenge, a case comprises data from a single patient from a single visit or image acquisition setting, with 1) a single image (CT volume) and 2) a segmentation labelmap.

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

Volumes with segmentation annotations: 113 (79 Covid19 cases, 34 non-Covid19 cases).
Volumes without segmentation annotations: >100 cases.
Stratified split with training: 60%; Validation: 13%; Testing: 27%.

Synthetic dataset: 96 synthetic volumes with 4D labelmaps.

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

Tradeoff between data and labeling availability and sufficient data variability during training/validation/testing (based on experience).

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.

Non-annotated volumes (N>100) can be utilized during training for e.g. un-/self-supervised model training, but only with CT image data (i.e. no voxel-level segmentations).

The real dataset is not accessible to the participants due to data privacy constraints. To have a realistic dataset to work with locally, we provide a synthetic 'toy dataset', which is similar to real data, within deformation and resampling artifacts that could also be expected during typical augmentation methods.

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.

Manual image annotations, created by one annotator per volume (annotations created by a team of five annotators in total).

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.

All training, validation and test cases were annotated in the same manner, following segmentation guidelines. Educational videos were provided [4, 5] and articles were recommended [1, 6, 7, 8]. The following regions were annotated: ILR (inferior lobe right), SLR (superior lobe right), MLR (middle lobe right), SLL (superior lobe left), ILL (inferior lobe left), GGO (ground glass opacities), CON (consolidation).

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.

Medical students in fifth/final year of study, after successful passing of radiology and pneumology study modules (medical study program, LMU Munich, GER), and after training in 3D-Slicer based Covid19 lung and lesion

segmentation.

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.

Only single-rater groundtruth labelmaps are available (for all cases), so no aggregation is needed. Subsequent quality assurance was provided by by students and M3i staff (4-eyes principle).

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.

All volumes are geometrically centered (origin at volume center) and stored as Nifti files (.nii.gz). Slicer segmentation files (.seg.nrrd) are exported to 4D labelmap Nifti files (.nii.gz) to encode multilabel voxel annotations (e.g. a voxel can be both part of a lung lobe and a lesion).

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.

No error estimation was performed through inter-/intra-rater variability.

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

No other possible sources of error were investigated.

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)

Segmentation quality measures: Dice overlap coefficient (Dice) and Hausdorff surface distance (HD) for each lung lobe region and averaged, as well as separately for GGO/CON. Physiological parameters (lung/GGO/CON volumes, lesion volume fraction) are automatically computed from the segmentation labelmaps.

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

Dice and HD are common segmentation performance measures. Volume measures have a prognostic value for Covid19 progression.

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.

A case-based ranking is computed for the hunters for each metric and category. The detailed ranking scheme is

provided in Appendix A.

For the rangers, the participants entitled to the incentive will be selected by the organizing committee and a voting method based on the popularity of the code within the ranger and hunters group.

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

In a case-based ranking scheme, algorithms with missing values will be set to the worst possible rank for the specific case.

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

The ranking scheme for the hunters was chosen in order to ensure a proper missing value handling for the HD distance. As the metric doesn't provide an upper bound, it is not straightforward to determine a worst value. In a case-based ranking, this problem is easily solved by penalising a missing value with the worst rank for the specific case. Furthermore, rankings are computed for each metric and category to highlight important performance difference for different aspects and anatomical regions.

The ranking scheme for the rangers makes sure that the hunters themselves will be involved in the question which provided the most helpful code.

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.

Ranking variability and robustness will be examined by bootstraping analyses. Different visualization techniques will be used to highlight different aspects of the performances. Analyses will be performed within the challenge visualization toolkit challengeR (https://github.com/wiesenfa/challengeR).

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

Bootstrapping analyses have shown promising in an extensive study [9]. Furthermore, the importance to visualize and analyse challenge results was shown in [10].

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.

For the challenge publication, the top 5 performing algorithms will be evaluated as an ensemble method and benchmarked against the individual models (after the competition).

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.

[1] Bell, D.J. COVID-19. https://radiopaedia.org/articles/covid-19-4
[2] ACR. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection
[3] WHO - Radiation and health. Use of chest imaging in COVID-19. https://www.who.int/publications/i/item/use-of-chest-imaging-in-covid-19
[4] Organ segmentation 3.1 : Lungs, lobes, volumetry. https://www.youtube.com/watch?v=cplH8cvAB8s&t=1203s
[5] 3D slicer SARS COV-2 (COVID) 3D reconstruction. https://www.youtube.com/watch?v=Q34zooUk64E&t=150s
[6] Murphy, A. et al. Lobar consilidation. Radiopaedia. https://radiopaedia.org/articles/lobar-consolidation?lang=us
[7] Bell, D.J. et al. Ground glass opafication. Radiopaedia. https://radiopaedia.org/articles/ground-glass-opacification-3?lang=us
[8] Weerakkody, Y. et al. Pleural effusion (summary). Radiopaedia. https://radiopaedia.org/articles/pleural-effusion-summary
[9] Maier-Hein, L. et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications, 9(1), 5217. https://doi.org/10.1038/s41467-018-07619-7
[10] Wiesenfarth, M. et al. (2019). Methods and open-source toolkit for analyzing and visualizing challenge results. ArXiv:1910.05121 [Cs, Stat]. http://arxiv.org/abs/1910.05121

Appendix A

Case-based ranking scheme for the hunters

1. For each algorithm $a_l, l = 1, \dots, N$, a rank is separately determined each metric $m_i \in M = \{DSC, HD\}$ for each category $c_j \in C = \{ILR, SLR, MLR, SLL, ILL, GGO, CON\}$ and is referred to as metric-category-rank $r_{i,j}(a_l)$. It is computed as follows:

    a. Determine the performance $m_i(a_l, t_k)$ for all algorithms $a_l, l = 1, \dots, N$ for all test cases $t_k, k = 1, \dots, M$ where M is the number of test cases and N is the number of competing algorithms.

    b. Based on $m_i(a_l, t_k)$, compute a case-specific rank $r_{i,j,k}(a_l)$ which is specific to the metric i and the category j.
        i. If $m_i(a_l, t_k) = NA$, $r_{i,j,k}(a_l)$ is set to the worst possible rank.

    c. Compute the metric-category-rank $r_{i,j}(a_l)$ by aggregating over the case-specific ranks $r_{i,j,k}(a_l)$ using the mean.

2. Based on the metric-category-ranks, a rank over the categories will be computed for each metric and algorithm. The metric-rank $r_i(a_l)$ is computed by aggregating over the category-specific ranks $r_{i,j}(a_l)$ by the mean.

3. The final rank $r(a_l)$ is computed by aggregating over the metric-specific ranks $r_i(a_l)$ by the mean.