ToothFairy2 Challenge: Multi-Structure Segmentation in CBCT Volumes: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

ToothFairy2 Challenge: Multi-Structure Segmentation in CBCT Volumes

Challenge acronym

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

ToothFairy2

Challenge abstract

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

The use of Cone Beam Computed Tomography (CBCT) is increasing not only in dentistry, but in the whole field of head and neck surgery. The main advantages of CBCTs are related to the short acquisition time and to the low radiation dose, while keeping an excellent visualization of anatomical structures, especially of hard tissues. In this regard, during the previous year challenge (ToothFairy) we have tackled the segmentation of the Inferior Alveolar Canal (IAC), a noble structure that lies within the mandible, whose identification and preservation represent a primary objective of many surgical interventions. In 2024 challenge, we aims at increasing the number of anatomical structures to be considered in the segmentation, thus including the mandible, the teeth, the maxillary bone, and the pharynx. Their framing is cross-disciplinary, as they are involved in all the head and neck surgical specialties, as well as in clinical and anesthesiological daily practice. In this regards, deep learning models can support medical personnel in the surgical planning procedures by providing an automatic voxel-level segmentation.

The challenge we propose aims at pushing the development of deep learning frameworks to segment anatomical structures in CBCTs by incrementally extending the amount of publicly available 3D-annotated CBCT scans and providing the first public-available fully annotated dataset. With respect to ToothFairy this new edition appears as an innovative and multidisciplinary one, expanding the field of view and the tasks of interest of the previous challenge in the perspective of a ever increasing cross-disciplinarity and clinical application.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

CBCTs, Segmentation, 3D Volumes, Maxillofacial

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.

N/A

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.

We are expecting about 30 participating teams.

This is the second edition of the challenge. The first edition was about the segmentation of the Inferior Alveolar Canal (IAC) from CBCT volumes and it received submissions (algorithms and associated description/paper) from ~20 different teams. During the onsite event additional people (not belonging to any of the challenge participating teams) showed interest in the event.

Potential participants are listed below:

Yusheng Liu, Rui Xin, Tao Yang, Lisheng Wang [1]
Haoshen Wang, Chenfan Xu, Zhiming Cui [2]
Marek Wodzinski, Henning Müller, Yannick Kirchhoff, Maximilian R. Rokuss, Klaus Maier-Hein [3]
Jaehwan Han, Wan Kim, Tae-Hoon Yong, Byungsun Choi [4]
Tomasz Szczepański, Michal K.Grzeszczyk, Przemyslaw Korzeniowski [5]
Vicent Caselles Ballester [6]

Affiliations:
[1] Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, China
[2] School of Biomedical Engineering, ShanghaiTech University, China
[3] German Cancer Research Center (DKFZ), Heidelberg, Division of Medical Image Computing, Germany
[4] Osstem Implant Laboratory, Korea
[5] Sano Centre for Computational Medicine, Poland
[6] Universitat Oberta de Catalunya, Spain

Publication and future plans

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

All the members of the team qualify as authors of the challenge submission. A paper resuming challenge results and including description of each proposal will be published after the challenge: at most two authors for each group may be eligible as authors of this paper.

All the participants are allowed to submit their own results in any venue (conferences, workshops, etc) with embargo restriction: 6 months after the MICCAI2024 event.

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.

We require standard technical equipment, including a projector, microphones, and a stable internet connection.

TASK 1: Multi-class Segmentation in CBCT Volumes

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.

The use of Cone Beam Computed Tomography (CBCT) is increasing not only in dentistry, but in the whole field of head and neck surgery. The main advantages of CBCTs are related to the short acquisition time and to the low radiation dose, while keeping an excellent visualization of anatomical structures, especially of hard tissues. In this regard, during the previous year challenge (ToothFairy) we have tackled the segmentation of the Inferior Alveolar Canal (IAC), a noble structure that lies within the mandible, whose identification and preservation represent a primary objective of many surgical interventions. In 2024 challenge, we aims at increasing the number of anatomical structures to be considered in the segmentation, thus including the mandible, the teeth, the maxillary bone, and the pharynx. Their framing is cross-disciplinary, as they are involved in all the head and neck surgical specialties, as well as in clinical and anesthesiological daily practice. In this regards, deep learning models can support medical personnel in the surgical planning procedures by providing an automatic voxel-level segmentation.

The challenge we propose aims at pushing the development of deep learning frameworks to segment anatomical structures in CBCTs by incrementally extending the amount of publicly available 3D-annotated CBCT scans and providing the first public-available fully annotated dataset. With respect to ToothFairy this new edition appears as an innovative and multidisciplinary one, expanding the field of view and the tasks of interest of the previous challenge in the perspective of a ever increasing cross-disciplinarity and clinical application.

Keywords

List the primary keywords that characterize the task.

CBCTs, Segmentation, 3D Volumes, Maxillofacial

ORGANIZATION

Organizers

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

Prof. Federico Bolelli (University of Modena and Reggio Emilia)
Luca Lumetti (University of Modena and Reggio Emilia)
Shankeeth Vinayahalingam (Radboud University)
Mattia Di Bartolomeo (Sapienza University of Rome)
Niels van Nistelrooij (Radboud University)
Kevin Marchesini (University of Modena and Reggio Emilia)
Prof. Alexandre Anesi (University of Modena and Reggio Emilia)

Prof. Costantino Grana (University of Modena and Reggio Emilia)

b) Provide information on the primary contact person.

Federico Bolelli (federico.bolelli@unimore.it)

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

The proposed challenge is a repeated event with an annual fixed submission deadline. Indeed, this is the second edition of the ToothFairy Challenge, namely ToothFairy 2.

While the first challenge edition focused on the segmentation of the Inferior Alveolar Canal (IAC), in the second edition more anatomical structures will be considered: mandible, teeth, maxillary bone, and the pharynx.

All the schedule details can be found in the "Challenge Schedule" section.

Challenge venue and platform

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

The challenge is associated with MICCAI conference. The on-site event of the challenge will be a thematic half-day event organized in conjunction with two other challenges concerning similar topics.

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

The platform used to run the platform challenge will be grand-challenge.org. Previous challenge edition is available at https://toothfairy.grand-challenge.org/.

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

A specific website will be implemented and available at www.toothfairychallenge.eu. Right now, the link points to the previous edition of the challenge.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Only fully automated methods are accepted as the submission is in the form of docker containers. It is not possible to submit manual annotations or interactive methods.

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.

It is allowed to use additional datasets and/or pre-trained networks, as long as it is clearly stated in the submission and they are publicly available.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Members of the organizers' institutes may participate in the challenge but are not eligible for awards and will be clearly identified in the leaderboards.

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

The challenge will award first three classified institutions/research group. For each institution/research group only the last submission will be considered in the final ranking.

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.

The top three performing methods will be announced publicly and they must share their code on a public repository (e.g. GitHub). If the submission authors will not publicly share the code with appropriate README for replicating the experiments, the prize(s) will shift to next classified institution(s).

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).

All the members of the team qualify as authors of the challenge submission. A paper resuming challenge results and including description of each proposal will be published after the challenge: at most two authors for each group may be eligible as authors of this paper.

All the participants are allowed to submit their own results in any venue (conferences, workshops, etc.) with embargo restriction: 6 months after the MICCAI2024 event.

Since we are also proposing a related workshop at MICCAI2024, Large Medical Data Segmentation (LaMDaS), in case of workshop acceptance an exception will be given to the 6 months embargo. This means that any of the challenge participants will be allowed to submit a paper describing the algorithmic solutions employed in the challenge to the LaMDaS workshop. The embargo will still be valid for any other workshop/conference venue.

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.

Since the test dataset cannot be released, all the participants would submit a docker container or a GitHub repository, following the template that we will provide in case of challenge acceptance. The submission instructions will be detailed at https://toothfairy2.grand-challenge.org/how-to-submit/ in case of challenge acceptance.

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.

Teams can have multiple submissions on the preliminary test set (a subset of three volumes taken from the entire test set) to verify the quality of their methods and the compliance with submission policies. However, the number of submissions per day per team is limited to 1 in order to restrict test data hacking. If a team overcome submission limits, they will be disqualified.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

The challenge schedule is defined as follows:

- Release of training cases: 1st April 2024;
- Release of test cases: test cases will not be publicly released;
- Submission date(s): from 1st July 2024 to 31st July 2024 included;
- Associated workshop days: 6-10 October 2024;
- Release date of the results: 15th September 2024.

Submissions performed between the 1st and the 31st of July 2024 will be evaluated on a subset of the test. This way, all the participants will have feedback on the correctness of the submitted docker/code and access to a temporary leaderboard.

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).

Data composing the training set have already received the approval of the ethical committee and they can be shared for research purposes only. More specifically, the ethical approval for this study was obtained from Comitato Etico dell'Area Vasta Emilia Nord (Approval Number 1374/2020/OSS/ESTMO SIRER ID 1275 - NAI-CBCT-D).

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 can be used for research purposes and for participating in this challenge. Any publication using our data must explicitly reference this challenge and cite [1,2]. The license that is applied to the data is CC BY-SA (Attribution-ShareAlike).

[1] Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., ... & Grana, C. (2022). Deep segmentation of the mandibular canal: a new 3D annotated dataset of CBCT volumes. IEEE Access, 10, 11500-11510.

[2] Cipriano, M., Allegretti, S., Bolelli, F., Pollastri, F., & Grana, C. (2022). Improving segmentation of the inferior alveolar nerve through deep label propagation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21137-21146).

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The code used to evaluate the submissions, as well as the whole source code of the website will be publicly available on GitHub. The link pointing to the aforementioned repository will be placed in the challenge website.

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

As previously mentioned, the top three performing methods must share their code on a public repository (e.g. GitHub).

To be considered for the post challenge publication, teams must make their algorithm publicly available (e.g., on a GitHub repository) as well as the network's parameters. An appropriate license may be selected. Anyway, all the participants are warmly invited to do so, thus guarantee reproducible research and results.

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.

We are in contact with different companies and institutions that are eligible to become sponsors of the challenge, including See Through S.r.l, NewTom, and others. Members of the companies that will be sponsors of the challenge can participate with submission, but will not be eligible for prizes. As mentioned, the test set will not be publicly released, but we plan to release GT masks after the challenge.

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

The main fields of application targeted by the algorithms regards diagnosis and intervention planning about dental operations.

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.

Final biomedical applications will acquire data using standard CBCT scanning protocol.

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

Target cohort includes patients older than 16 years who performed a CBCT for monitoring or surgery planning. Only CBCTs that were unreadable or where there was a massive alteration of the normal anatomy (mainly due to previous surgical resective procedures) have been excluded, thus allowing a wide variability close to a real clinical setting. The target cohort and the challenge cohort (should) match since the challenge dataset has been acquired in a real-case scenario.

Imaging modality(ies)

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

Any kind of image technique that can be used to automatically segment 3D volumes is accepted.

Context information

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

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

Along with the images, a ground-truth segmentation is provided in the form of an image, i.e., a 3D volume having the same spatial size of the image data with 0 values corresponding to background and an unique integer identifier for each different anatomical part.

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

No patient clinical data is 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.

Lower and upper jaw of the patients, acquired by means of Cone Beam Computer Tomography.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm must target the different anatomical structures present in the jaws, by providing a volume where each voxel is identified with an integer value corresponding to the detected structure.

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)).

The aim is to provide a reliable tool that can be integrated in the daily clinical practice. A correct and three-dimensional detection of each anatomical structure significantly improves the planning and execution of different maxillofacial surgical procedures.

In this scenario, with the term "reliable" we refer to an automatic system that is precise, i.e. the fraction of relevant instances among the retrieved instances is high, and sensitive, i.e. The set of relevant instances retrieved is close to the set of all relevant instances present in the data. In this regard, the Dice score, which can also be viewed as the F1-Score, is a combination of precision and sensitivity.

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).

All the data of the training set has been obtained using a NewTom/NTVGiMK4, 3 mA, 110 kV, 0.3 mm cubic voxels. Instead, for what concerns test data they are obtained using a standard CBCT scanning protocol (i-CAT, 3D Imaging System, Imaging Sciences International Inc, Hatfield, PA, USA) in "Extended Field" modus (FOV: 16 cm diameter/22 cm height; scan time: 2 x 20s; voxel size: 0.4 mm).

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

Pixel spacing and intra-slice distance is always 0.3 millimeters. Data volumes are already converted to the Hounsfield Unit (HU) and their values range between -1000 and 5264.

Volume shapes range from (148, 265, 312) to (178, 423, 463) for the Z, Y and X axes respectively.

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.

The 3D CBCT volumes composing training dataset have been acquired by the Affidea center located in Modena, Italy. Affidea is a leading pan-European healthcare group specialized in the provision of advanced diagnostics, specialist outpatient services, laboratory analyses, physiotherapy and rehabilitation, cancer diagnosis and treatment. It counts 312 different centers in 15 different countries, with about 11,000 professionals.

The test set is from the department of oral and maxillofacial surgery in Radboud University Nijmegen Medical Centre

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).

Both data acquisition and data annotation tasks were performed by medical experts with many (>5) years of experience.

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.

Both training and test cases are CBCT volumes of the jaws. Although every test case entails a thorough scan of the entire targeted anatomical structures, the training set exhibits variations in the field of view. Specifically, within a subset of the training cases, there may be instances where a portion of the upper jaw as well as the upper teeth are partially absent. Anyway, each case is associated with a 3D annotation where each anatomical structure is identified with a unique integer value.

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

The training dataset is composed of more than 450 volumes as training set and 50 volumes used as test set. All the training volumes will be provided with a complete annotation. While we provide an "official" split setting for the public data, each challenge participant is free to decide how to split training data to create an internal validation set.

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

The test set comprises a number of cases that is (more than) enough to evaluate the performance of algorithms under different perspectives. Having more (and more) data as training set is always helpful, but expensive and time consuming (extremely so). The current amount of training data selected for the challenge represents the biggest training set available in literature at the moment of writing this proposal.

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.

As mentioned, the test will not be released for ethical issues. We chose to use a set of data coming from different institutions to keep the training set as big as possible. When selecting the test cases we ensured that the clinical distributions (i.e., age, sex, etc.) and the annotators are aligned with those of the training set.

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.

All the annotations have been obtained by means of a specifically developed software that allows experts to annotate each image data. Annotators are doctors with more than 5 years of experience. The total number of annotators involved is 5.

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.

The annotators are medical experts and the annotation tool has been developed following their instructions/requirements thus maximizing the annotation accuracy. For this reason, no instructions are needed for the annotators.

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.

Every case has been annotated by medical experts with years (>5) of experience in the maxillofacial field.

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

No multiple annotations are available in our scenario, meaning that each CBCT volume is annotated by a single expert only.

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.

No pre-processing has been applied to the data, which is provided in a "raw" format.

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.

Volumes have been manually annotated by medical experts, but no multiple annotations are available for the same CBCT.

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

No other relevant source of error has been identified.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

The metric used to rank the submitted proposals is the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95). Please refer to the ranking method(s) section for more details.

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

Together with IoU (Intersection over Union), DSC represents the goal standard for the evaluation of segmentation algorithms. Since their meaning is practically the same, there is no reason to preserve both the metrics. The average Hausdorff Distance is a widely used performance measure employed in medical image segmentation as it provides a distance between two points sets, which is extremely relevant in clinical applications.

In order to select the most appropriate metrics, Metrics Reloaded (https://metrics-reloaded.dkfz.de/) recommendations have been employed. Among the suggested metrics there was the Dice score, and the Normalized Surface Distance (NSD). Compared to the Hausdorff Distance (HD), the NSD is less sensitive to the outliers. We opted for the HD95 because, as an example, when measuring the distance between the inferior alveolar nerve and the teeth roots, it is mandatory to have the upper bound error and not an average one. This allows practitioners to carefully plan surgical operations for teeth removal while preventing any risk of damaging the nerve.

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 schema involves the following steps:

- For each class and for each volume, calculate the Dice score (DSC) and the HD95. Compute also the maximum used memory (Mem), and the total execution time (Time) for all cases;
- Average the DSC and the HD95 obtained for each class across all volumes, obtaining a DSCc and a HD95c for

each class c.
- Rank all the DSCc, HD95c, maximum used memory, and running time, independently;
- Average the rankings obtained at point 2 for each DSCc and HD95c to produce the final rank;
- If two or more final ranks obtained at point 4 are equal, compute the average of the rankings obtained at point 2 for Mem and Time to break ties;
- If two or more ranks are still equal, it is a tie: the prize will be evenly split.

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

If an algorithm does not produce any result for a specific CBCT volume we will consider the predicted segmentation as a volume of 0s, meaning that the Dice score for that image will be equal to 0 for each class and the HD95 will be the maximum unsigned integer value representable using 64bit.

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

There is no interdependence between cases. Each CBCT (both in train and test) belong to a different patient and have been annotated by only one expert. The selected metrics will provide homogeneous numbers on different patients that can be later averaged to provide the final rank. Different metrics (i.e., DICE and Hausdorff distance) calculated separately are later aggregated to produce the final ranking.

To ensure robustness in the final ranking, the recommendations provided by Maier et al. [1*] have been followed. In general, there are two contrasting approaches to aggregate metrics across the test cases. The first approach, known as metric-based aggregation, involves initially aggregating metric values across all test cases (e.g., using mean or median), and then ranking the algorithms on the aggregated value. The second approach is the case-based aggregation, which starts by calculating a rank for each test case, and then the final rank is determined by aggregating the ranks of the test cases. According to the paper by Maier et al. [1*] the single-metric rankings (specifically for DSC and HD95 in our case) demonstrate higher statistical robustness when employing metric-based aggregation and using the mean instead of the median for aggregation.

[1*] Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications 9, 5217

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.

We will perform bootstrapping to assess ranking uncertainty.

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

NA

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.

In the challenge summary paper we plan to discuss and evaluate ranking variability, inter-algorithm variability, and the combination of submitted algorithms via ensembling.

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.

[1] Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., ... & Grana, C. (2022). Deep segmentation of the mandibular canal: a new 3D annotated dataset of CBCT volumes. IEEE Access, 10, 11500-11510. DOI: 10.1109/ACCESS.2022.3144840
[2] Cipriano, M., Allegretti, S., Bolelli, F., Pollastri, F., & Grana, C. (2022). Improving segmentation of the inferior alveolar nerve through deep label propagation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21137-21146). DOI: 10.1109/CVPR52688.2022.02046

Further comments

Further comments from the organizers.

No