Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge: Structured description of the challenge design

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

CHALLENGE ORGANIZATION

Title

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

Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge

Challenge acronym

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

M&Ms

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.

In the recent years, many machine/deep learning models have been proposed to accurately segment cardiac structures in magnetic resonance imaging [1-4]. However, when these models are tested on unseen datasets acquired from distinct MRI scanners or clinical centres the segmentation accuracy can be greatly reduced [5,6]. This makes it difficult for these tools to be applied consistently across multiple clinical centres, especially when subjects are scanned using different MRI protocols or machines. The M&MS challenge is the first international competition to date on cardiac image segmentation combining data from different centres, vendors, diseases and countries at the same time. It will evaluate the generalisation ability of machine/deep learning and cross-domain transfer learning techniques for cardiac image segmentation, by testing these on a cohort of 350 cardiac MRI studies comprising healthy, hypertrophic and dilated hearts, acquired in three different countries (Spain, Germany and Canada) and by using four distinct MRI vendors (Siemens, Philipps, General Electric and Canon). The challenge will be supported by the H2020 euCanSHare project (www.eucanshare.eu), which is building a multi-centre big data platform for cardiovascular personalised medicine research. The three top teams will receive prizes of 500, 300 and 200 Euros, respectively.

Challenge keywords

List the primary keywords that characterize the challenge.

Cardiac image segmentation – Multiple MRI machines – Machine/deep learning

Year

The challenge will take place in ...

2020

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

STACOM (Statistical Atlases and Computational Modelling of the Heart)

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.

Between 10 and 20 teams, based on numbers of previous segmentation challenges.

Publication and future plans

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

A challenge journal paper will be prepared and submitted. We propose two co-authors per participating centre (e.g. junior and senior author). Participants are free to publish their results separately.

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.

All challenge information will be hosted on a website created for that purpose managed by Universitat de Barcelona. The website URL will be available in the next months.

TASK: Cardiac image 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.

The purpose of the challenge is to accurately segment the cardiac anatomy in short axis cine magnetic resonance images independently of the machine vendor. For this purpose, the segmentation outcome from the different models will be tested for four different vendors. The participants will have access to images from two different vendors and the final test will have two additional vendors not seen during the training phase.

Keywords

List the primary keywords that characterize the task.

Cardiac image segmentation – Multiple MRI machines – Machine/deep learning

ORGANIZATION

Organizers

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

Víctor M. Campello (Universitat de Barcelona, Spain)
José F. Rodríguez Palomares (Vall d'Hebron Hospital, Barcelona, Spain)
Andrea Guala (Vall d'Hebron Hospital, Barcelona, Spain)
Mahir Karakas (Clinical University Hamburg, Germany)
Matthias Friedrich (McGill University Health Centre, Montreal, Canada)
Karim Lekadir (Universitat de Barcelona, Spain)

b) Provide information on the primary contact person.

Contact: Víctor M. Campello (victor.campello@ub.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.

Examples:

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

One time event.

Challenge venue and platform

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

MICCAI.

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

The platform will be a website created for that purpose managed by Universitat de Barcelona.

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

https://www.ub.edu/mnms/

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.

No additional data 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.

The top three teams will receive prizes of 500, 300 and 200 Euros, respectively.

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 selected solutions will be announced publicly.

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).

A challenge journal paper will be prepared and submitted. We propose two co-authors per participating centre (e.g. junior and senior author). Participants are free to publish their results separately.

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.

Participants will submit their models using Docker containers where test data will be processed.

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.

Each team will be able to submit the segmentation results to the organization before the deadline a maximum of five times. The last submission will be considered for the final ranking.

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

Release of training data: 1st May 2020
Registration period: 1st May - 31st May 2020
Last submission: 15th July 2020
Workshop day: 3rd October 2020
Results release: 3rd October 2020

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 ethical approvals will be obtained from the participating clinical centres. All datasets are fully anonymized.

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.

Additional comments: The data will be distributed under the license CC BY and the challenge participants will need to commit to not disseminate the data during the challenge duration.
However, the studies from McGill will not be included in this data distribution. These will only be used during the test phase as demanded by the institution due to legal and ethical issues.

Additional comments: The data will be distributed under the license CC BY and the challenge participants will need to commit to not disseminate the data during the challenge duration.
However, the studies from McGill will not be included in this data distribution. These will only be used during the test phase as demanded by the institution due to legal and ethical issues.

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 for computing the performance scores used by the organizers is available in the public repository for medpy library: https://github.com/loli/medpy/

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

Participants will be highly encouraged to post their code openly in a public repository.

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.

No conflicts of interest are declared. Test labels will only be handled by the organizers and the clinical collaborators, and will be released publicly after the workshop.
The challenge will be supported by the H2020 euCanSHare project (www.eucanshare.eu), which is building a multi-centre big data platform for cardiovascular personalised medicine research.

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, CAD, Research, Prognosis, Screening.

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.

The target cohort will be patients scanned in different centres with different vendors and having different cardiomyopathies.

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 patients with hypertrophic and dilated cardiomyopathies as well as healthy subjects that were scanned in clinical centres of three different countries (Spain, Germany and Canada) using four different magnetic resonance vendors (Siemens, General Electric, Philips and Canon).

Imaging modality(ies)

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

Magnetic resonance imaging (MRI)

Context information

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

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

Images will be stratified by vendor and centre using anonymized indices.

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

Sex and disease type will be provided for each patient.

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.

Short-axis view of cardiac magnetic resonance imaging.

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 target will be the cardiac anatomic structures corresponding to the left and right ventricle blood pools and the myocardium.

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)).

Robustness.

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).

Four different scanners from four different vendors (Siemens, General Electric, Philips and Canon) were used during the acquisition. The field strength is 1.5T for all of them.

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

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.

Since the data will be publicly available and for anonymization reasons, the exact centre and scanner will not be matched. No information about the study centre will be provided.
100 studies were acquired at Hospital Vall d'Hebron (Spain).
50 studies were acquired at Hospital Universitari Dexeus (Spain).
50 studies were acquired at McGill University Health Centre (Canada).
50 studies were acquired at Clínica Sagrada Familia (Spain).

50 studies were acquired at Universitätsklinikum Hamburg-Eppendorf (Germany).
50 studies were acquired at Hospital de la Santa Creu i Sant Pau (Spain).

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).

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 represents a patient scan, that is, the set of two-dimensional short-axis images that cover the heart from the apex to the base. There will be cases from two different vendors in the training case.

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

Training: 150
Testing: 200

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

The training and test set numbers are determined by the amount of data collected for each vendor (100 cases each, except for only 50 cases for Canon). 25% of images from "seen"/training vendors will be reserved for testing (25+25 cases), as well as the whole set of images from the two extra unseen vendors (100+50 extra cases).

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 training set will contain 75 annotated images for each of two different vendors.
The participants will be able to select a validation set from the training cases.
The test cases will correspond to new images from the two vendors provided in the training sample and two extra unseen vendors.

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.

Each case will be segmented by an experienced clinician from each institution. This manual annotation will be used as ground truth.

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.

Annotators were asked to annotate the images using the Circle cvi42 software.

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.

Each centre will have its own expert clinician for their annotations, with experiences ranging from 2 to more than 10 years.

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.

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 cases will be transformed from dicom to nifti format and no further pre-processing will be applied to images.

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.

Most relevant sources of error are the delineation of the epicardial contour on dark regions with possible overestimation of myocardial wall thickness. Moreover, depending on the annotator, there could be differences in the delineation of the endocardial contour and the amount of trabeculae that is included in the left ventricle mask. Finally, the delineation of the slice where the mitral valve appears could bring disagreement between the annotators.

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

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 following metrics will be computed: Dice similarity coefficient, Jaccard index, surface distance and Hausdorff distance. All metrics will be used to compute the ranking.

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

The reason for this metric selection is to account for changes in the boundary apart from global changes, that are

already taken into account when computing the Dice and Jaccard coefficients.

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.

Mean and standard deviation will be computed centre- and vendor-wise for each of the anatomical regions in the heart (left and right ventricles and myocardium). The results will be combined with the following weighted sum: v1/6 + v2/6 + v3/3 + v4/3, where v1 and v2 correspond to results for vendors used during training, and v3 and v4 correspond to results associated to the new vendors. The weights were selected so that the new vendors have twice the weight of already seen vendors.
The details for computing the final performance rank are the following: (a) the weighted average metric for each region (LV, RV and MYO) is computed, (b) a min-max normalization is computed across subjects, (c) the final normalized metrics are averaged to extract one unique value between 0 and 1, (d) this final value is used to rank the participants.
*An excel file was provided for clarification on this method.

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

Missing results in the submission will get a zero for Dice and Jaccard coefficients and the equivalent worst value for Hausdorff and surface distances (i.e. when normalized in the range [0,1], being 1 the worst and 0 the best value, missing results will get a 1).

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

The ranking is computed in the aforementioned way to motivate participants to think about generalization of their models and accurate segmentation on new unseen vendors. Nevertheless, segmentation results on vendors used for training will still be part of the final score in the ranking.

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.

Welch's t-tests will be performed to assess significance in the difference of mean segmentation accuracy between scanners, centres and diseases.

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

Welch's t-test is chosen because it allows to test two populations with possibly different variances and unequal number of samples.

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.

The intra-class correlation coefficient will be computed for all the test set to identify those with larger disagreement when segmented with automatic methods.

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] Bernard, Olivier, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?." IEEE transactions on medical imaging 37.11 (2018): 2514-2525.
[2] Tran, Phi Vu. "A fully convolutional neural network for cardiac segmentation in short-axis MRI." arXiv preprint arXiv:1604.00494 (2016).
[3] Isensee, Fabian, et al. "Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features." International workshop on statistical atlases and computational models of the heart. Springer, Cham, 2017.
[4] Zotti, Clément, et al. "GridNet with automatic shape prior registration for automatic MRI cardiac segmentation." International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, Cham, 2017.
[5] Tao, Qian, et al. "Deep learning–based method for fully automatic quantification of left ventricle function from Page 1 of 12Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge cine MR images: a multivendor, multicenter study." Radiology 290.1 (2018): 81-88.
[6] Dangi, Shusil, Ziv Yaniv, and Cristian Linte. "A Distance Map Regularized CNN for Cardiac Cine MR Image Segmentation." arXiv preprint arXiv:1901.01238 (2019).
[7] Zhuang, Xiahai et al. “Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge.” Medical Image Analysis (2019).

Further comments

Further comments from the organizers.

Another challenge, with multiple centres and scanners (MM-WHS challenge, STACOM 2017) involved MRI data sets that were all acquired in the same country and city (London, UK, three countries in our case) and two types of scanner only (four scanners in our cases) [7].