Diabetic Foot Ulcers Grand Challenge 2022: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Diabetic Foot Ulcers Grand Challenge 2022

Challenge acronym

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

DFUC2022

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.

Diabetes is a global epidemic affecting approximately 425 million people. This figure is expected to rise to 629 million people by 2045. Diabetic Foot Ulcers (DFU) are a serious condition that frequently results from the disease. The rapid rise of the condition over the last few decades is a major challenge for healthcare systems around the world. Cases of DFU frequently lead to more serious conditions such as infection and ischaemia that can significantly prolong treatment and often result in limb amputation, with more serious cases leading to death. In an effort to improve patient care and reduce the strain on healthcare systems, recent research has focused on the creation of detection algorithms that could be used as part of a mobile app that patients could use themselves (or a carer/partner) to monitor their condition and to detect the appearance of DFU [2-4]. To this end, the collaborative work between Manchester Metropolitan University, Lancashire Teaching Hospitals and the Manchester University NHS Foundation Trust has created an international repository of up to 11,000 DFU images for the purpose of supporting more advanced methods of DFU research. Analysis of ulcer regions and surrounding skin is an important aspect in DFU management [1]. Manual delineation of ulcers and periwound regions are very time-consuming and challenging for podiatrists. With joint effort from the lead scientists of the UK, US, India and New Zealand, this challenge will solicit original works in automated DFU segmentation and promote interactions between researchers and interdisciplinary collaborations.

Challenge keywords

List the primary keywords that characterize the challenge.

diabetic foot ulcers, deep learning, segmentation

Year

The challenge will take place in ...

2022

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none

Duration

How long does the challenge take?

Half day.

Expected number of participants

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

The expected number of participants is 60. In DFUC 2020 (DFU detection), we received 31 submissions from 11 teams. This is expected to grow to 50 participants in DFUC 2021 (DFU classification). Since the release of the DFUC 2020 dataset in April 2020, we received 55 requests from 26 countries. Many of the researchers were seeking for dataset for segmentation tasks, which is the main driver for organising DFUC 2022 (DFU segmentation).

Publication and future plans

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

We will write at least one journal paper to summarize the challenge results. We will continue to make the dataset available for research use. Our future plan is to continue to collect and delineate the ulcer and periwound regions.

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 will be hosting our challenge on grand-challenge.org

TASK: Analysis Towards Segmentation of Diabetic Foot Ulcers

SUMMARY

Keywords

List the primary keywords that characterize the task.

Diabetic foot ulcer, foot pathology, machine learning, deep learning, segmentation

ORGANIZATION

Organizers

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

Moi Hoon Yap, Manchester Metropolitan University
Neil Reeves, Manchester Metropolitan University
Andrew Boulton, University of Manchester and Manchester Royal Infirmary
Satyan Rajbhandari, Lancashire Teaching Hospital
David Armstrong, University of Southern California
Arun G. Maiya, Manipal College and Health Professions and Indian Podiatry Association
Bijan Najafi, Baylor College of Medicine in Texas
Eibe Frank, University of Waikato
Justina Wu, Waikato District Health Board

b) Provide information on the primary contact person.

Moi Hoon Yap, Manchester Metropolitan University
Email: M.Yap@mmu.ac.uk

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

Repeated annual event with different aspects of the challenge to address.

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.

grand-challenge.org

(will setup once the proposal is accepted)

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

https://dfu-challenge.github.io/

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.

Private data is allowed.

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

Any organizations/companies affiliated with members of the organizing committee are not excluded from participation in the challenge, but must ensure that their submissions are completely independent of the members of the organizing committee.

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

Certificates will be provided for the top 3 performing teams. We are actively seeking sponsorship and we anticipate being able to provide cash prizes and / or graphics cards. Our co-chairs are very well connected and we are confident in being able to attain sponsorship. Our first MICCAI challenge (2020) was sponsored by NVIDIA, who provided the prize for the winning team, so we do not anticipate any issues sourcing prizes for 2022.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

All results will be made available publicly, and the top 3 performing methods will be invited to the challenge event to present their work.

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 challenge organizers will publish at least one challenge journal paper and potentially more. The authors may publish their papers separately and decisions on publication strategy will be made according to achieving publication in the highest ranking journals.

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 purpose of result verification and to encourage reproducibility and transparency, all entries must submit the following:
- Mask images indicating image id, with pixel-wise label for background (0), ulcer region (1) and periwound region (2). If there is more than one periwound for a single ulcer, then only the largest periwound should be recorded. This is to ensure that all submissions are fairly and correctly evaluated for comparisons.
- A paper highlighting the contribution of the submission, but not limited to, the method, experimental results and analysis, prepared according to the format stipulated by MICCAI 2022. All challenge entries should be accompanied by a description of the method.
- GitHub repository URL containing all source code for their implemented method, and all other relevant files such as feature/parameter data. To help publicize our workshop and domain area, please mention (or add relevant links to) DFUC 2022 and MICCAI 2022. The participants may provide this URL in a simple text file while submitting.
For all files, participants should submit a single zip file and upload to the submission system as supplementary material. The submission link will be made available from 01/07/2022.

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.

The participating teams will be able to validate their results based on the validation set provided by the organizers. Submissions to DFUC 2022 are issued a validation score. This is to provide a sanity check of the submission (ensure the submission is in the correct format) and is not intended to be used for algorithm ranking or evaluation.

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

Training data release: 01/04/2022
Validation data release: 21/06/2022
Test images release: 01/07/2022
Submission deadline: 15/07/2022
Winner and invitation speakers: 15/08/2022
(subject to change depending on the MICCAI 2022 deadlines)

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

We have received approval from the UK National Health Service (NHS) Research Ethics Committee (REC) to use these images for the purpose of research. The NHS REC reference number is 15/NW/0539.

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)

Prospective participants will need to request the dataset by completing a license agreement and emailing a formal request to the data owners.

Additional comments: The dataset consists of 1500 labelled images available for training and 500 labelled images for validation. Additionally, there will be 1500 unlabelled images available for users to optionally use for training. The test dataset (planned release 01/07/2022) will contain an additional 3000 images. To download the dataset, please visit:
http://www2.docm.mmu.ac.uk/STAFF/M.Yap/dataset.php (to appear on 01/04/2022).
Download and complete the license agreement form, email to M.Yap@mmu.ac.uk with email subject: DFUC 2022.

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.

We will provide an online platform to evaluate the results. For transparency, we will release the source code used for calculating final scores after the closing date of the challenge.

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

GitHub repository URL containing source code for their implemented method, and all other relevant files such as feature/parameter data. To help publicize our workshop and domain area, participants have to mention (or add relevant links to) DFUC 2022 and MICCAI 2022. Participants may provide this URL in a simple text file while submitting.

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.

None

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

Research, 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.

People with or at risk of diabetic foot ulcers, their caregivers (e.g. family member) and care providers (e.g., podiatrists, physicians and wound nurses).

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

People with diabetes who developed foot ulcers.

Imaging modality(ies)

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

Photography.

Context information

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

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

Foot images.

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

People with diabetic foot ulcers.

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.

Diabetic foot shown in normal photography.

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.

Ulcer/wound on a foot.

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)).

Precision, Specificity, Sensitivity, Accuracy.

Additional points: Find highly accurate diabetic foot ulcer pathology segmentation algorithm for foot photographs.

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).

Three cameras were used for capturing the foot images, Kodak DX4530, Nikon D3300 and Nikon COOLPIX P100.

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

The images were acquired with close-ups of the full foot at a distance of around 30–40 cm with the parallel orientation to the plane of an ulcer. The use of flash as the primary light source was avoided, and instead, adequate room lights were used to get the consistent colors in images.

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.

We have received approval from the UK National Health Service (NHS) Research Ethics Committee (REC) to use these images for the purpose of research. The NHS REC reference number is 15/NW/0539. Foot images with DFU were collected from the Lancashire Teaching Hospitals over the past few years.

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).

Images were acquired by a podiatrist and a consultant physician with specialization in the diabetic foot, both with more than 5 years professional 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.

A case refers to one foot image. All training, validation and test cases are labelled with ulcer and periwound.

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

Training: 3000 DFU images (1500 labelled images and 1500 unlabelled images)
Validation: 500 labelled images (all DFU)
Testing: 3000 labelled images (DFU + non-DFU images)

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

For DFU segmentation, we follow a 50-50 split for the training and testing sets. For the segmentation task, it is an important aspect to create an algorithm that does not require laborious manual delineation. To encourage researchers in using other strategies (moving away from fully supervised learning), such as unsupervised, semi-supervised and self-supervised learning, we include 1500 unlabelled images for the training set.

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.

For DFU segmentation, a previous study [1] showed good accuracy with less than 1000 training images. Hence, 6000 images are sufficient for machine learning in DFU segmentation.

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.

The ground truth was produced by two healthcare professionals (a podiatrist and a consultant physician) who specialize in diabetic wounds and ulcers. The annotation process was completed by using the VGG Image Annotator (VIA) application developed by the University of Oxford Visual Geometry Group [https://www.robots.ox.ac.uk/~vgg/software/via/].

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.

Since these are the expert annotators in DFU, the instruction was to delineate all ulcers and ulcer periwounds on each image using the polygonal shape tool within the VGG Image Annotator application.

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.

A podiatrist and a consultant physician with specialization in the diabetic foot, both with more than 5 years professional experience.

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.

The most notable variation across annotations were the number of polygonal lines used to delineate ulcers and periwounds. Some cases exhibit more lines to represent the shape of the ulcer, whereas others exhibit fewer. We would consider this inter-rater variation to be normal as some podiatrists may have more time to delineate a wound than others. For cases where too few lines were used, we requested for the delineation to be completed again using more polygon lines that better reflected the actual shape of the wound. We will use the average of the delineated contours for the final ground truth.

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.

In this dataset, the size of images varies between 1600 × 1200 and 3648 × 2736. We will resize all the images to 640 × 480 (preserve the aspect ratio of the ulcers) to improve the performance and reduce computational costs.

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.

For inter-annotator variability measures, we compute the agreement of the contours produced by our two expert annotators by using Intersect over Union (IoU). We observed high inter-annotator reliability of >0.9. To produce the final contour, we average the delineated contours where they do not precisely match. The position of the resulting contour would be the average between the original contours.

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

We have discarded photographs with out of focus and blurry artefacts. To produce the final ground truth contours, in the case of disagreement, a third specialist podiatrist examined the photograph. The final decision was mutually settled with the consent of three experts.

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)

Users are challenged to build an algorithm to segment diabetic foot ulcers and its periwound from medical images. This competition will evaluate the similarity between the predicted contours produced by computer algorithms and the ground truth contours produced by the domain experts. The Dice similarity coefficient (DSC) is widely used to compare the pixel-wise agreement between a predicted segmentation (X) and its corresponding ground truth (Y). The DSC is defined to be 1 when both X and Y are 100% overlap (similar). The leader board score is the mean of the DSC for each image in the test set.

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

The Dice similarity coefficient is a statistical tool which measures the similarity between two sets of data. It is a commonly used metric to evaluate segmentation challenges and competitions on the grand challenge website and Kaggle, and is the most used metric in validating medical volume segmentation [5]. We have chosen this as we believe this is the best approach for the DFUC2022 challenge.

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.

We will evaluate the score using 4 decimal places on the final score to reduce the possibility of a tie. However, in the event of a tie, we will award the team who have submitted their solution first as the winner.

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

All the missing results on a DFU image, i.e. images with no segmented ulcer or no segmented periwound, will be treated as no prediction on the image. Therefore, such cases will be treated as false negatives.

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

These metrics provide a balanced judgement on whether an approach can segment the ulcers and periwound equally well, hence reducing the possibility that an approach could be well-fitted to only work for certain classes.

Additionally, the speed in providing the solution will be used for the ranking. The time stamp on the submission system will be used to reward the participants who provided the best solution in the shortest duration.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

For statistical analysis, we will use Kendall's tau analysis to quantify the agreement of the metrics and methods. We will be reporting the results during the challenge event with transparency.

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

The interpretation of Kendall's tau is very direct in observing the agreeable and non-agreeable pairs (in terms of probabilities). Additionally, bootstrapping will be used to analyze the variability of a ranking scheme. Bootstrapping allows estimation of the sampling distribution of almost any statistic using random sampling methods.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

N/A

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

[1] Goyal, M., Yap, M.H., Reeves, N.D., Rajbhandari, S. and Spragg, J., 2017, October. Fully convolutional networks for diabetic foot ulcer segmentation. In 2017 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 618-623). IEEE.

[2] Cassidy B. et al., 2020. DFUC2020: Analysis Towards Diabetic Foot Ulcer Detection. arXiv preprint arXiv:2004.11853. 2020 Apr 24.

[3] Yap, M.H. et al., 2020. Deep Learning in Diabetic Foot Ulcers Detection: A Comprehensive Evaluation. arXiv preprint arXiv:2010.03341.

[4] Yap, M. H., Chatwin, K. E., Ng, C. C., Abbott, C. A., Bowling, F. L., Rajbhandari, S., . . . Reeves, N. D. (2018). A New Mobile Application for Standardizing Diabetic Foot Images. Journal of Diabetes Science and Technology, 12(1),

169-173. doi:10.1177/1932296817713761

[5] Maier-Hein et al. (2020) BIAS: Transparent reporting of biomedical image analysis challenges. Medical Image Analysis, 101796. doi: https://doi.org/10.1016/j.media.2020.101796

Further comments

Further comments from the organizers.

The team has a strong track record in diabetes research (world leading researchers: Prof. Andrew Boulton, Prof. David Armstrong and Prof. Neil Reeves), computer vision and machine learning (Prof. Eibe Frank and Dr. Moi Hoon Yap), digital health and digital twin (Prof. Bijan Najafi).

Moi Hoon Yap has successfully conducted the following grand challenges in the past years:
DFUC2020: https://dfu-challenge.github.io/
The Facial Micro-Expressions Grand Challenges (2018-2020)
http://www2.docm.mmu.ac.uk/STAFF/m.yap/FG2018Workshop.htm
https://facial-micro-expressiongc.github.io/MEGC2019/
https://megc2020.github.io/

The team has experience in handling and sharing datasets to encourage reproducible research:
http://www2.docm.mmu.ac.uk/STAFF/M.Yap/dataset.php