International Skin Imaging Collaboration (ISIC) Challenge: using dermoscopic image context to diagnose melanoma : Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

International Skin Imaging Collaboration (ISIC) Challenge: using dermoscopic image context to diagnose melanoma

Challenge acronym

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

ISIC2020

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.

Skin cancer is one of the most frequent types of cancer and manifests mainly in areas of the skin most exposed to the sun. Despite being the less frequent of the common skin cancers, melanoma is responsible for 75% of deaths from skin tumors. It can appear at any age, and it is the first most diagnosed cancer among patients from 25-29 years old, the second among 20-24 years-old and the third solid tumor among 15-19 years-old.
Melanoma is one of the cancers with more years of productive life lost and is the most expensive cancer when expressed in terms of cost per death in Europe.

Since skin cancer occurs on the surface of the skin, its lesions can be evaluated by visual inspection. Dermoscopy is an imaging device composed of a magnifying glass coupled with polarized light, which permits visualizing more profound levels of the skin as its surface reflection is removed. Prior research has found that this technique permits improved visualization of the lesion structures, enhancing the accuracy of dermatologists. Typically, experts search for specific structural and color cues that help them determine if a lesion is of a particular type of skin cancer, rule sets such as the "ABCD rule" have been used to standardize the clinical procedure associated with the diagnosis of skin cancer.

Many medical institutions are not only using but also capturing images of the skin lesions using specialized dermoscopic adapters coupled with high-resolution cameras. The increase of imaging data of this modality has motivated the appearance of computer vision algorithms that aim to diagnose the lesions on the dermoscopic images automatically.

Earlier systems relied on the extraction of handcrafted features from the skin lesions, similar to the rule sets the clinicians were using to perform diagnosis. Researchers tried to develop highly specialized algorithms which would extract color, border features, symmetry, and a bunch of other types of diagnostic criteria that was later on appended on a machine learning classification algorithm. However, the increased availability of dermoscopic images has motivated the appearance of more sophisticated algorithms based on deep learning, mainly on convolutional neural networks. These algorithms are no longer based on rule sets since the feature extractor is already embedded in their architecture. A significant player in the adoption of these algorithms in the community has been the The International Skin Imaging Collaboration (ISIC), which has been organizing yearly challenges since 2016, where participants are asked to develop computer vision algorithms to help with the classification of dermoscopic images skin lesions.

Challenge keywords

List the primary keywords that characterize the challenge.

skin, neural networks, dermoscopy, image, context information

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.

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.

500+

Prior iterations of the ISIC challenge (hosted at ISBI in 2016 and MICCAI in 2018 and 2019) have received hundreds of submissions. In 2019, there were 169 submissions overall. We anticipate a significant increase in participation this year due to a new collaboration with the Society for Imaging Informatics in Medicine (SIIM). In 2019 the SIIM challenge (https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation) received 1,983 submissions so we expect a significantly increased participation list from last year in the proposed joint challenge.

Publication and future plans

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

Every year the organizing team publishes results on the challenge both in draft form to arXiv, and later final drafts at appropriate venues. For prior years, please see:

https://arxiv.org/abs/1605.01397
https://www.ncbi.nlm.nih.gov/pubmed/28969863
https://arxiv.org/abs/1710.05006 (Accepted to ISBI 2018)
https://www.ncbi.nlm.nih.gov/pubmed/31201137

For all challenges, we frame as both a competition and collaboration: fusion approaches among all participant submissions are evaluated to inspect the degree to which performance may improve in comparison to the best single approaches.

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.

Prior challenges have been hosted online using the Covalic Platform. For details, please see the challenge webpages from prior years:

http://challenge2019.isic-archive.com/
http://challenge2018.isic-archive.com/
http://challenge2017.isic-archive.com/
http://challenge2016.isic-archive.com/

During the challenge workshop, participants will be invited to give oral presentations. Projector display with podium PC and HDMI input will be required. Video camera for session recording if permitted. Room or area to support 150 individuals and 20+ posters.

TASK: Dermoscopic Image Classification without Contextual Information

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 goal of the combined tasks in this challenge is to help participants develop image analysis tools to enable the automated diagnosis melanoma using patient-level contextual information, a process more similar to clinical workflow. In practice, dermatologists frequently identify "ugly duckling" moles as most likely to be melanoma. Participants will be tested in two phases: in task 1 participants will be given dermoscopy images alone and asked to diagnose melanoma. Systems will be ranked according to AUC for melanoma and results will be compared to those of task 2 where contextual information will be provided.

Keywords

List the primary keywords that characterize the task.

skin, neural network, dermoscopic, image, contextual information, patient-centric analysis

ORGANIZATION

Organizers

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

- Noel C. F. Codella, Ph.D. (IBM Research, New York, USA)
- M. Emre Celebi, Ph.D. (University of Central Arkansas, Arkansas, USA)
- Kristin Dana, Ph.D. (Rutgers University, New Jersey, USA)
- David Gutman (Emory University, Georgia, USA)
- Brian Helba (Kitware, New York, USA)
- Harald Kittler (Medical University of Vienna, Vienna, Austria)
- Philipp Tschandl, M.D. Ph.D. (Medical University of Vienna, Vienna, Austria)
- Allan Halpern, M.D. (Memorial Sloan Kettering Cancer Center, New York, USA)
- Veronica Rotemberg, M.D. (Memorial Sloan Kettering Cancer Center, New York, USA)
- Josep Malvehy, M.D. (Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain)
- Marc Combalia, MsC. (Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain)

b) Provide information on the primary contact person.

nccodell@us.ibm.com

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.

Covalic

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

https://challenge.isic-archive.com/

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.

While both public and private external data is permitted, its use must be denoted upon submission.

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 participate but not eligible for awards and not listed in leaderboard.

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

1st Place	$6,000
2nd Place	$3,500
3rd Place	$2,000
4th - 10th Place	$500 each

In order to be eligible for prizes, the teams must participate in Task 1 and Task 2 although the tasks will be scored separately.

e) Define the policy for result announcement.

Examples:

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

A leaderboard will be made public on the challenge platform.

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 participants can optionally submit their work specific to their submission to journals or conferences.

Every year the organizing team publishes results on the challenge both in draft form to arXiv, and later final drafts at appropriate venues. There is no embargo time for participants. For prior years, please see:

https://arxiv.org/abs/1605.01397
https://www.ncbi.nlm.nih.gov/pubmed/28969863
https://arxiv.org/abs/1710.05006 (Accepted to ISBI 2018)
https://www.ncbi.nlm.nih.gov/pubmed/31201137

For all challenges, we frame as both a competition and collaboration: fusion approaches among all participant submissions are evaluated to inspect the degree to which performance may improve in comparison to the best single approaches.

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 be expected to submit a single CSV file, with one row for each test image. The first column of each row must contain the image ID. The remaining columns must contain predicted confidence scores (as a decimal value in the closed interval [0.0, 1.0], where 0.5 is used as the binary classification threshold and the maximum value used for classification).

Optionally, the participants may be invited to submit a docker file with their algorithms, which will be used after the challenge to better understand the results. The participants who submit a docker container may be invited to co-author a publication after the challenge celebration.

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 is allowed unlimited number of submissions per "approach" however only the most recent submission will be counted toward the leaderboard. Submissions will be capped to 1 per day.
Three approaches will be scored per team and each team will submit a manuscript highlighting the difference between approaches.

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
1st May 2019

TEST DATA AND VALIDATION DATA RELEASE
7th July 2019

SUBMISSION DEADLINE
14th August 2019

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).

Institutional review board approval was obtained for all the data providers, and specific information for each dataset will be included upon release of the training set.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC.

Additional comments: Some of the data is provided under CC-0. Each image or subset of the datasets will be annotated with each specific Creative Commons license.

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.

https://github.com/ImageMarkup/isic-challenge-scoring

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

The participants won't be required to provide their code.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

The challenge is devoted to an academic-industry partnership model and has in previous years obtained industry sponsorship, and avenues for industry sponsorship are currently being pursued. However, industry sponsors do not have access to the dataset especially gold standard labels for the test dataset with the exception of Noel Codella, the challenge organizer, who is also employed at IBM. While Noel will have access to the test data he will not participate in the challenge or be eligible for prizes.

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

Screening, Diagnosis, Decision support, Prevention.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

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.

Skin lesions from all anatomic sites in dermatology and general practice clinics.

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

Skin lesions from all anatomic sites at high risk melanoma clinics in Australia, Austria, Spain, and the United States.

Imaging modality(ies)

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

Dermoscopic images.

Context information

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

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

Individual dermoscopic images.

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

Gender, sex, anatomic location, lesion ID, obfuscated patient age.

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.

Target entities would be cutaneous lesions from all (surface) anatomical locations that are amenable to dermoscopic 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.

Identification of melanoma using dermoscopic images.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Hardware requirements, Usability, Robustness, Consistency, Reliability, Specificity, Sensitivity, Accuracy, Runtime.

Additional points: Identify highly accurate algorithms for the detection of melanoma. Ensemble level and clinical applicability information will be assessed for the participants who submit (optional) docker containers of their algorithms.

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).

1- Molemax - Derma Medical Systems
2- Vectra - Canfield Inc
3- DermEngine - MetaOptima

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

Cases will be formed from data collected in standard clinical practices and prospective clinical trials to evaluate lesion photography for diagnosis of melanoma.

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.

1. Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain
2. Memorial Sloan Kettering Cancer Center New York, NY
3. Department of Dermatology, Medical University of Vienna. Vienna, Austria
4. Melanoma Institute Australia. Sydney, Australia
5. The University of Queensland, Brisbane, Australia

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).

Board certified dermatologists have determined which lesions to photograph and will be responsible for annotation (using expert consensus verification). Dermatopathologists with experience in melanoma diagnosis will perform histopathologic confirmation.

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.

Training and test cases both represent a dermoscopic image of a lesion in the skin of a patient. Both training and test cases diagnoses have been annotated via expert consensus or histopathological verification.

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

1. There will be 15000 images used for training
2. There will be 100 images used for validation of submissions
3. There will be 5000 images used for test

This breakdown was determined due to the complexity of the challenge dataset in which patient-level information is included. Therefore, all images from a given patient (either one with or without melanoma) must be included together in either training, testing, or validation. The sets were designed to sufficiently incorporate melanomas and benign lesions with varying amounts of “geographical contextual” information to test the hypothesis that clinical information improves performance (especially specificity).

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

The cases were developed using a retrospective evaluation of the contributing institution datasets using pathology and image characteristics. Images were discarded only if they were of poor quality or if QA could not be performed.

We used all available and not previously released retrospective images from participating institutions from the past 2-5 years for which labels could accurately be identified.

2/3 Train
1/3 Test
100 images for held-out validation

This split was designed based on previous challenges, determining adequate numbers of melanomas in order to confidently determine algorithm accuracy, and patient-level information.

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.

Class distribution in classification tasks chosen according to real-world distribution in high risk melanoma clinics. Distributions will be balanced between train and test splits.

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.

Board certified dermatologists (for expert consensus verification) and dermatopathologists (for histopathological verification).

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.

Cases will be formed from data collected in standard clinical practices and prospective clinical trials to evaluate lesion photography for diagnosis of melanoma. Annotators will be board certified dermatologists who have experience evaluating dermoscopic lesions and images as part of their standard clinical practice so no additional training is needed. The annotators will be asked whether they believe a lesion is melanoma or not, and whether or not they would perform a biopsy.

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.

The images are submitted by the world experts in melanoma diagnosis across the world and image QA is performed prior to image submission. In a second step process board certified dermatologists will annotate benign lesions using expert consensus and malignant lesions will be labeled according to histopathology (reviewed by a board certified pathologist).

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.

If multiple annotations are performed on one case and consensus cannot be achieved a majority vote will be used to merge annotations.

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 will be performed on the data.

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.

Sources of error:

- Dermathology expert annotation error
- Dataset merge

Individual image QA is performed to reduce potential errors. We estimate less than 1 % of errors in the ground truth.

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

None

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)

AUC (to compute rankings), sensitivity, specificity.

Apart from the metric regarding algorithmic performance, we will also measure the running time and memory usage of the algorithms, but this information will not be considered for the leaderboard. In order to do so, we will ask for docker submissions and these will be run in the infrastructure from the organizing team.

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

Clinical application on skin lesion classification must be able to discriminate benign from malignant in order to determine which lesions to biopsy. The complexity of the 2020 dataset will allow for only benign melanocytic versus melanoma classification, as those are the two categories frequently confused for each other clinically. Algorithms will be ranked based on area under the curve for melanoma diagnosis.

AUC, which is the area of the receiver operating characteristic curve (over sensitivity and specificity), is a standard measure of classification performance that is inherently invariant to imbalanced classes (sensitivity is a measure over positives independently, and specificity is a measure over negatives independently). AUC also serves as a summary of multiple operating points along the ROC curve, as many thresholds along the curve are clinically relevant.

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.

In 2020 we have increased the diversity of the training and test set, including images of lesions that cannot be diagnosed from photographs beyond benignity. Therefore the classification scheme will be binary (melanoma vs not melanoma). Ranking will be performed by binary AUC for melanoma (each image submission will be incorporated on whether the algorithm correctly diagnosed the binary class categorization). Tied teams will be additionally ranked by sensitivity.

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

No missing data allowed.

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

In 2020 we have increased the diversity of the training and test set, including images of lesions that cannot be diagnosed from photographs beyond benignity. Therefore the classification scheme will be binary (melanoma vs not melanoma). Ranking will be performed by AUC where tied teams will be additionally ranked by sensitivity.

AUC, which is the curve over sensitivity and specificity, is inherently invariant to imbalanced classes. Sensitivity is a measure over positives, and specificity is a measure over negatives. No ranking indicator is a function of the ratio of positives to negatives (or the imbalance).

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.

All metrics will be computed automatically with full machine precision. Exact ties will be reflected as equal rankings (skipping the next rank) on the published leaderboard.

Statistical methods will be: T-test to compare submissions and performance between task 1 and task 2, sensitivity and specificity for melanoma, area under the curve.

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

In 2020 we have increased the diversity of the training and test set, including images of lesions that cannot be diagnosed from photographs beyond benignity. Therefore the classification scheme will be binary (melanoma vs not melanoma). Ranking will be performed by AUC where tied teams will be additionally ranked by sensitivity.

AUC, which is the curve over sensitivity and specificity, is inherently invariant to imbalanced classes. Sensitivity is a measure over positives, and specificity is a measure over negatives. No ranking indicator is a function of the ratio of positives to negatives (or the imbalance).

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 participants who provide a Docker Container will be eligible for a joint publication after the challenge. Their algorithms will be tested for a number of different situations.

The algorithms may be also aggregated to form ensembles to test the joint algorithmic accuracy.

TASK: Dermoscopic Image Classification with Contextual Information

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 goal of this challenge is to help participants develop image analysis tools to enable the automated diagnosis melanoma using patient-level contextual information, a process more similar to clinical workflow. In practice, dermatologists frequently identify “ugly duckling” moles as most likely to be melanoma.

In Task 2, the test set will be identified using patient-level information to provide patient context. Systems will be ranked according to AUC for melanoma. The novelty of the challenge this year is the additional information mimicking a clinical workflow in which many dermoscopic images from the same patient will be evaluated and linked using an (anonymized) patient identifier.

Keywords

List the primary keywords that characterize the task.

skin, neural network, dermoscopic, image, contextual information, patient-centric analysis

ORGANIZATION

Organizers

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

- Noel C. F. Codella, Ph.D. (IBM Research, New York, USA)
- M. Emre Celebi, Ph.D. (University of Central Arkansas, Arkansas, USA)
- Kristin Dana, Ph.D. (Rutgers University, New Jersey, USA)
- David Gutman (Emory University, Georgia, USA)
- Brian Helba (Kitware, New York, USA)
- Harald Kittler (Medical University of Vienna, Vienna, Austria)
- Philipp Tschandl, M.D. Ph.D. (Medical University of Vienna, Vienna, Austria)
- Allan Halpern, M.D. (Memorial Sloan Kettering Cancer Center, New York, USA)
- Veronica Rotemberg, M.D. (Memorial Sloan Kettering Cancer Center, New York, USA)
- Josep Malvehy, M.D. (Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain)
- Marc Combalia, MsC. (Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain)

b) Provide information on the primary contact person.

nccodell@us.ibm.com

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.

Covalic

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

https://challenge.isic-archive.com/

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.

While both public and private external data is permitted, its use must be denoted upon submission.

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 participate but not eligible for awards and not listed in leaderboard.

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

1st Place	$6,000
2nd Place	$3,500
3rd Place	$2,000
4th - 10th Place	$500 each

In order to be eligible for prizes, the teams must participate in Task 1 and Task 2 although the tasks will be scored separately.

e) Define the policy for result announcement.

Examples:

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

A leaderboard will be made public on the challenge platform.

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).

For all challenges, we frame as both a competition and collaboration: fusion approaches among all participant submissions are evaluated to inspect the degree to which performance may improve in comparison to the best single approaches.

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 be expected to submit a single CSV file, with one row for each test image. The first column of each row must contain the image ID. The remaining columns must contain predicted confidence scores (as a decimal value in the closed interval [0.0, 1.0], where 0.5 is used as the binary classification threshold and the maximum value used for classification).

Optionally, the participants may be invited to submit a docker file with their algorithms, which will be used after the challenge to better understand the results. The participants who submit a docker container may be invited to co-author a publication after the challenge celebration.

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 is allowed unlimited number of submissions per "approach" however only the most recent submission will be counted toward the leaderboard. Submissions will be capped to 1 per day.
Three approaches will be scored per team and each team will submit a manuscript highlighting the difference between approaches.

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
1st May 2019

TEST DATA AND VALIDATION DATA RELEASE
7th July 2019

SUBMISSION DEADLINE
14th August 2019

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).

Institutional review board approval was obtained for all the data providers, and specific information for each dataset will be included upon release of the training set.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC.

Additional comments: Some of the data is provided under CC-0. Each image or subset of the datasets will be annotated with each specific Creative Commons license.

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.

https://github.com/ImageMarkup/isic-challenge-scoring

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

The participants won't be required to provide their code.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

The challenge is devoted to an academic-industry partnership model and has in previous years obtained industry sponsorship, and avenues for industry sponsorship are currently being pursued. However, industry sponsors do not have access to the dataset especially gold standard labels for the test dataset with the exception of Noel Codella, the challenge organizer, who is also employed at IBM. While Noel will have access to the test data he will not participate in the challenge or be eligible for prizes.

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

Screening, Diagnosis, Decision support, Prevention.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

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.

Skin lesions from all anatomic sites in dermatology and general practice clinics.

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

Skin lesions from all anatomic sites with corresponding patient-level information at high risk melanoma clinics in Australia, Austria, Spain, and the United States.

Imaging modality(ies)

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

Dermoscopic images.

Context information

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

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

Individual dermoscopic images. Images will be annotated with (anonymized) patient ID information.

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

Gender, sex, anatomic location, lesion ID, obfuscated patient age, and patient ID.

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.

Target entities would be cutaneous lesions from all (surface) anatomical locations that are amenable to dermoscopic 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.

Identification of melanoma using all images from a given patient for contextual information.

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)).

Hardware requirements, Usability, Robustness, Consistency, Reliability, Specificity, Sensitivity, Accuracy., Runtime.

Additional points: Identify highly accurate algorithms for the detection of melanoma. Ensemble level and clinical applicability information will be assessed for the participants who submit (optional) docker containers of their algorithms.

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).

1- Molemax - Derma Medical Systems
2- Vectra - Canfield Inc
3- DermEngine - MetaOptima

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

Cases will be formed from data collected in standard clinical practices and prospective clinical trials to evaluate lesion photography for diagnosis of melanoma.

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.

1. Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain
2. Memorial Sloan Kettering Cancer Center New York, NY
3. Department of Dermatology, Medical University of Vienna. Vienna, Austria
4. Melanoma Institute Australia. Sydney, Australia
5. The University of Queensland, Brisbane, Australia

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).

Board certified dermatologists have determined which lesions to photograph and will be responsible for annotation (using expert consensus verification). Dermatopathologists with experience in melanoma diagnosis will perform histopathologic confirmation.

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.

Training and test cases both represent a dermoscopic image of a lesion in the skin of a patient. Both training and test cases diagnoses have been annotated via expert consensus or histopathological verification.

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

Task 2 will include the same training and test cases however the addition of patient information (relational "context") in which multiple images from the same patient will be identified will be provided.

1. There will be 15000 images used for training
2. There will be 100 images used for validation of submissions
3. There will be 5000 images used for test

This breakdown was determined due to the complexity of the challenge dataset in which patient-level information is included. Therefore, all images from a given patient (either one with or without melanoma) must be included together in either training, testing, or validation. The sets were designed to sufficiently incorporate melanomas and benign lesions with varying amounts of “geographical contextual” information to test the hypothesis that clinical information improves performance (especially specificity).

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

The cases were developed using a retrospective evaluation of the contributing institution datasets using pathology and image characteristics. Images were discarded only if they were of poor quality or if QA could not be performed.

2/3 Train
1/3 Test
100 images for held-out validation

This split was designed based on previous challenges, determining adequate numbers of melanomas in order to confidently determine algorithm accuracy, and patient-level information.

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.

Class distribution in classification tasks chosen according to real-world distribution in high risk melanoma clinics. Distributions will be balanced between train and test splits.

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.

Board certified dermatologists (for expert consensus verification) and dermatopathologists (for histopathological verification).

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.

Cases will be formed from data collected in standard clinical practices and prospective clinical trials to evaluate lesion photography for diagnosis of melanoma. Annotators will be previously trained board certified dermatologists who have experience evaluating dermoscopic lesions and images as part of their standard clinical practice so no additional training is needed. The annotators will be asked whether they believe a lesion is melanoma or not, and whether or not they would perform a biopsy.

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.

The images are submitted by the world experts in melanoma diagnosis across the world and image QA is performed prior to image submission. In a second step process board certified dermatologists will annotate benign lesions using expert consensus and malignant lesions will be labeled according to histopathology (reviewed by a board certified pathologist).

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.

If multiple annotations are performed on one case and consensus cannot be achieved a majority vote will be used to merge annotations.

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 will be performed on the data.

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.

Sources of error:
- Dermathology expert annotation error
- Dataset merge

Individual image QA is performed to reduce potential errors. We estimate less than 1 % of errors in the ground truth.

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

None

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)

AUC (to compute rankings), sensitivity, specificity.

Apart from the metric regarding algorithmic performance, we will also measure the running time and memory usage of the algorithms, but this information will not be considered for the leaderboard. In order to do so, we will ask for docker submissions and these will be run in the infrastructure from the organizing team.

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

Clinical application on skin lesion classification must be able to discriminate benign from malignant in order to determine which lesions to biopsy. The complexity of the 2020 dataset will allow for only benign melanocytic versus melanoma classification, as those are the two categories frequently confused for each other clinically. Algorithms will be ranked based on area under the curve for melanoma diagnosis.

AUC, which is the area of the receiver operating characteristic curve (over sensitivity and specificity), is a standard measure of classification performance that is inherently invariant to imbalanced classes (sensitivity is a measure over positives independently, and specificity is a measure over negatives independently). AUC also serves as a summary of multiple operating points along the ROC curve, as many thresholds along the curve are clinically relevant.

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.

In 2020 we have increased the diversity of the training and test set, including images of lesions that cannot be diagnosed from photographs beyond benignity. Therefore the classification scheme will be binary (melanoma vs not melanoma). Ranking will be performed by binary AUC for melanoma (each image submission will be incorporated on whether the algorithm correctly diagnosed the binary class categorization). Tied teams will be additionally ranked by sensitivity.

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

No missing data allowed.

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

In 2020 we have increased the diversity of the training and test set, including images of lesions that cannot be diagnosed from photographs beyond benignity. Therefore the classification scheme will be binary (melanoma vs not melanoma). Ranking will be performed by AUC where tied teams will be additionally ranked by sensitivity.

AUC, which is the curve over sensitivity and specificity, is inherently invariant to imbalanced classes. Sensitivity is a measure over positives, and specificity is a measure over negatives. No ranking indicator is a function of the ratio of positives to negatives (or the imbalance).

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.

All metrics will be computed automatically with full machine precision. Exact ties will be reflected as equal rankings (skipping the next rank) on the published leaderboard.

Statistical methods will be: T-test to compare submissions and performance between task 1 and task 2, sensitivity and specificity for melanoma, area under the curve.

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

In 2020 we have increased the diversity of the training and test set, including images of lesions that cannot be diagnosed from photographs beyond benignity. Therefore the classification scheme will be binary (melanoma vs not melanoma). Ranking will be performed by AUC where tied teams will be additionally ranked by sensitivity.

AUC, which is the curve over sensitivity and specificity, is inherently invariant to imbalanced classes. Sensitivity is a measure over positives, and specificity is a measure over negatives. No ranking indicator is a function of the ratio of positives to negatives (or the imbalance).

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 participants who provide a Docker Container will be eligible for a joint publication after the challenge. Their algorithms will be tested for a number of different situations.

The algorithms may be also aggregated to form ensembles to test the joint algorithmic accuracy.

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.

Further comments

Further comments from the organizers.

We are so grateful to MICCAI for hosting the challenge in 2018 and 2019. There were 30-40 attendees, all of whom attended instead of the dinner which was held at the same time. The challenge provided a wonderful opportunity for participants to provide feedback and engage with the technical aspects of the challenge. These have played a heavy role in informing future challenges including the 2020 proposal, and we would love to benefit from this engagement again this year.