Medical Out-of-Distribution Analysis Challenge 2021: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Medical Out-of-Distribution Analysis Challenge 2021

Challenge acronym

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

MOOD

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.

Despite overwhelming successes in recent years, progress in the field of biomedical image computing still largely depends on the availability of annotated training examples. This annotation process is often prohibitively expensive because it requires the valuable time of domain experts. Additionally, this approach simply does not scale well: whenever a new imaging modality is created, acquisition parameters change. Even something as basic as the target demographic is prone to changes, and new annotated cases have to be created to allow methods to cope with the resulting images. Image labeling is thus bound to become the major bottleneck in the coming years. Furthermore, it has been shown that many algorithms used in image analysis are vulnerable to out-of-distribution samples, resulting in wrong and overconfident decisions [20, 21, 22, 23]. In addition, physicians can overlook unexpected conditions in medical images, often termed 'inattentional blindness'. In [1], Drew et al. noted that 50% of trained radiologists did not notice a gorilla image, rendered into a lung CT scan when assessing lung nodules.

One approach, which does not require labeled images and can generalize to unseen pathological conditions, is Out-of-Distribution or anomaly detection (which in this context is used interchangeably). Anomaly detection can recognize and outline conditions that have not been previously encountered during training and thus circumvents the time-consuming labeling process and can therefore quickly be adapted to new modalities. Additionally, by highlighting such abnormal regions, anomaly detection can guide the physicians' attention to otherwise overlooked abnormalities in a scan and potentially improve the time required to inspect medical images.

However, while there is a lot of recent research on improving anomaly detection [8, 9, 10, 11, 12, 13, 14, 15, 16, 17], especially with a focus on the medical field [4, 5, 6, 7], a common dataset/ benchmark to compare different approaches is missing. Thus, it is currently hard to have a fair comparison of different proposed approaches. While in the last few months common datasets for natural data were proposed, such as default detection [3] or abnormal traffic scene detection [2], we tried to tackle this issue for medical imaging with last year's challenge [25].

In a similar setting to last year we suggest the medical out-of-distribution challenge as a standardized dataset and benchmark for anomaly detection. We propose two different tasks. First a sample-wise (i.e. patients-wise) analysis, thus detecting out-of-distribution samples. For example, having a pathological condition or any other condition not seen in the training-set. This can pose a problem to classically supervised algorithms and detection of such could further allow physicians to prioritize different patients. Secondly, we propose a voxel-wise analysis i.e. giving a score for each voxel, highlighting abnormal conditions and potentially guiding the physician.

However, there are a few aspects to consider when choosing an anomaly detection dataset.
First, as in reality, the types of anomalies should not be known beforehand. This can be a particular problem when choosing a dataset and testing on only a single pathological condition, which is vulnerable to exploitation. Even with an educated guess (based on the dataset) and a fully supervised segmentation approach, trained on a not allowed separate dataset, one could outperform other rightfully trained anomaly detection approaches. Furthermore, making the exact types of anomalies known can cause a bias in the evaluation. Studies have shown that proposed anomaly detection algorithms tend to overfit on a given task, given that properties of the test set and the kind of anomalies are known beforehand. This further hinders the comparability of different algorithms [6, 18, 19, 23]. As a second point, combining test sets, from different sources with alternative conditions, may also cause problems. By definition, the different sources already propose a distribution shift to the training dataset, complicating a clean and meaningful evaluation.

To solve these issues we propose to provide two datasets with more than 600 scans each, one brain MRI-dataset and one abdominal CT-dataset, to allow for a comparison of the generalizability of the approaches. In order to prevent overfitting on the (types of) anomalies existing in our test set, the test set will be kept confidential at all times. The training set consists of hand-selected scans in which no anomalies were identified. The remaining scans will be assigned to the test set. Thus some scans in the test set do not contain anomalies, whilst others contain naturally occurring anomalies. In addition to the natural anomalies, we will add synthetic anomalies. We choose different structured types of synthetic anomalies (e.g. a tumor or an image of a gorilla rendered into the a brain scan [1]) to cover a broad variety of different anomalies and also allow for an analysis of weaknesses and strengths of the methods by different factors (type, size, contrast, ...). We believe that this allows for a controlled and fair comparison of different algorithms (as recently similarly proposed by [3]).

While in last year's edition [25] multiple different approaches were present, among which some showed clearly superior performance for certain kinds of anomalies, generally all algorithms still failed in (different) obvious cases and especially medically relevant classes of anomalies where still lacking in (clinically relevant) performance. In this year's edition we will introduce new synthetic anomalies and furthermore focus more on clinically relevant and not very well performing anomalies from last year's challenge. Consequently we will completely renew the synthetic part of the test set, keeping last years classes of anomalies (however generating new test samples for those) and introduce new anomaly classes, which we found as relevant judging from last year's challenge.

We hope that providing a such standardized dataset allows for a fair comparison of different approaches and can outline how well different approaches work in a realistic and clinical setting.

[1] Drew, Trafton, Melissa L. H. Vo, and Jeremy M. Wolfe. "'The Invisible Gorilla Strikes Again: Sustained Inattentional Blindness in Expert Observers.'" Psychological Science 24, no. 9 (September 2013): 1848–53. https://doi.org/10.1177/0956797613479386.

[2] Bergmann, Paul, Michael Fauser, David Sattlegger, and Carsten Steger. "MVTec AD -- A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," 9592–9600, 2019. http://openaccess.thecvf.com/content_CVPR_2019/html/Bergmann_MVTec_AD_--_A_Comprehensive_Real-World_Dataset_for_Unsupervised_Anomaly_CVPR_2019_paper.html.

[3] Hendrycks, Dan, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. "A Benchmark for Anomaly Segmentation." ArXiv:1911.11132 [Cs], November 25, 2019. http://arxiv.org/abs/1911.11132.

[4] Chen, Xiaoran, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. "Deep Generative Models in the Real-World: An Open Challenge from Medical Imaging." CoRR abs/1806.05452 (2018).

[5] Baur, Christoph, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images." ArXiv:1804.04488 [Cs], April 12, 2018. http://arxiv.org/abs/1804.04488.

[6] Zimmerer, David, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. "Unsupervised Anomaly Localization Using Variational Auto-Encoders." In International Conference on Medical Image Computing and Computer-Assisted Intervention, 289–297. Springer, 2019.

[7] Schlegl, Thomas, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," n.d. https://arxiv.org/pdf/1703.05921.pdf.

[8] Abati, Davide, Angelo Porrello, Simone Calderara, and Rita Cucchiara. "Latent Space Autoregression for Novelty Detection." ArXiv:1807.01653 [Cs], July 4, 2018. http://arxiv.org/abs/1807.01653.

[9] Ahmed, Faruk, and Aaron Courville. "Detecting Semantic Anomalies." ArXiv:1908.04388 [Cs], August 13, 2019. http://arxiv.org/abs/1908.04388.

[10] Akçay, Samet, Amir Atapour-Abarghouei, and Toby P. Breckon. "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection." ArXiv:1901.08954 [Cs], January 25, 2019. http://arxiv.org/abs/1901.08954.

[11] Beggel, Laura, Michael Pfeiffer, and Bernd Bischl. "Robust Anomaly Detection in Images Using Adversarial Autoencoders." ArXiv:1901.06355 [Cs, Stat], January 18, 2019. http://arxiv.org/abs/1901.06355.

[12] Bergmann, Paul, Michael Fauser, David Sattlegger, and Carsten Steger. "Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings." ArXiv:1911.02357 [Cs], November 6, 2019. http://arxiv.org/abs/1911.02357.

[13] Choi, Hyunsun, Eric Jang, and Alexander A. Alemi. "WAIC, but Why? Generative Ensembles for Robust Anomaly Detection." ArXiv:1810.01392 [Cs, Stat], October 2, 2018. http://arxiv.org/abs/1810.01392.

[14] Guggilam, Sreelekha, S. M. Arshad Zaidi, Varun Chandola, and Abani Patra. "Bayesian Anomaly Detection

Using Extreme Value Theory." ArXiv:1905.12150 [Cs, Stat], May 28, 2019. http://arxiv.org/abs/1905.12150.

[15] Maaløe, Lars, Marco Fraccaro, Valentin Liévin, and Ole Winther. "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling." ArXiv:1902.02102 [Cs, Stat], February 6, 2019. http://arxiv.org/abs/1902.02102.

[16] Piciarelli, Claudio, Pankaj Mishra, and Gian Luca Foresti. "Image Anomaly Detection with Capsule Networks and Imbalanced Datasets." ArXiv:1909.02755 [Cs], September 6, 2019. http://arxiv.org/abs/1909.02755.

[17] Sabokrou, Mohammad, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. "Adversarially Learned One-Class Classifier for Novelty Detection." ArXiv:1802.09088 [Cs], February 25, 2018. http://arxiv.org/abs/1802.09088.

[18] Goldstein, Markus, and Seiichi Uchida. "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data." PLOS ONE 11, no. 4 (April 19, 2016): e0152173. https://doi.org/10.1371/journal.pone.0152173.

[19] Škvára, Vít, Tomáš Pevný, and Václav Šmídl. "Are Generative Deep Models for Novelty Detection Truly Better?" ArXiv:1807.05027 [Cs, Stat], July 13, 2018. http://arxiv.org/abs/1807.05027.

[20] Hendrycks, Dan, and Kevin Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks." ArXiv:1610.02136 [Cs], October 7, 2016. http://arxiv.org/abs/1610.02136.

[21] Mehrtash, Alireza, William M. Wells III, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation." ArXiv:1911.13273 [Cs, Eess], November 29, 2019. http://arxiv.org/abs/1911.13273.

[22] Roady, Ryne, Tyler L. Hayes, Ronald Kemker, Ayesha Gonzales, and Christopher Kanan. "Are Out-of-Distribution Detection Methods Effective on Large-Scale Datasets?" ArXiv:1910.14034 [Cs], October 30, 2019. http://arxiv.org/abs/1910.14034.

[23] Shafaei, Alireza, Mark Schmidt, and James J. Little. "A Less Biased Evaluation of Out-of-Distribution Sample Detectors." ArXiv:1809.04729 [Cs, Stat], August 20, 2019. http://arxiv.org/abs/1809.04729.

[24] Maier-Hein, L., Eisenmann, M., Reinke, A. et al. "Why rankings of biomedical image analysis competitions should be interpreted with care." Nat Commun 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7

[25] http://medicalood.dkfz.de/

Challenge keywords

List the primary keywords that characterize the challenge.

Anomaly detection, Anomaly localization, Out-of-distribution detection, novelty detection

Year

The challenge will take place in ...

2021

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.

Around 70 registrations with 10 submissions (last year)

Publication and future plans

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

We plan a publication of the results after the challenge.

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 plan on using the Synapse platform for submission of docker files for testing. We will run a continuous evaluation pipeline after the challenge starts where participants can check if their submission is valid and their toy task scores.

TASK: SampleLevel

SUMMARY

Keywords

List the primary keywords that characterize the task.

Anomaly detection, Anomaly localization, Out-of-distribution detection, novelty detection

ORGANIZATION

Organizers

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

Jens Petersen, Gregor Köhler, Paul Jäger, Peter Full, David Zimmerer, Klaus Maier-Hein
Div. Medical Image Computing (MIC), German Cancer Research Center (DKFZ)

Tobias Roß, Tim Adler, Annika Reinke, Lena Maier-Hein
Div Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ)

b) Provide information on the primary contact person.

David Zimmerer (MIC, DKFZ) d.zimmerer@dkfz.de

Life cycle type

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

Examples:

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

Repeated event with fixed submission deadline.

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.

Synapse.org

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

https://www.synapse.org/#!Synapse:syn21343101/wiki/599515

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.

No prices yet determined. Currently, the winner will win the right to brag about the results. (last year's winner received a NVidia RTX 3080)

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.

By default (i.e. if they don't decline) the top 3 (winning) teams will be announced publicly. The remaining teams may decide if they choose to appear on the public Leaderboard.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All participating teams with a valid submission. Teams that reveal their identity, can nominate members of their team as co-author for the challenge paper. However, we reserve the right to exclude teams and team members if they do not adhere to the challenge rules and guidelines.

The method description submitted by the authors may be used in the publication of the challenge results. Personal data of the authors will include their name, affiliation and contact addresses.

Participating teams may publish their own 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.

Docker container on the Synapse platform. Link to submission instructions:
https://www.synapse.org/#!Synapse:syn21989704 or medicalood.dkfz.de/web/

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Teams are allowed three submissions per challenge task, however, only the latest submission will be considered. For each challenge submission, the teams will receive the results on a very simple "toy" test-case, which will not be included in the final testset. We will also make the "toy" case and testing script publicly available so that the participants can check beforehand if their submission won't cause any errors and to compare the reported results with their own to check for any additional occurring errors. Additionally a dedicated toy-case evaluation queue will be available where participants can submit their solutions to our cluster and only run the toy cases and receive toy-case results. This toy-case evaluation queue will not count as a challenge submission and will have no submission limit and only serve as a "will run successfully" check.

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 cases: 1 April 2021
Registration Until: 06 September 2021
Submission of Dockers: 07 September 2021

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).

The provided challenge data is anonymized.
According to [Euro2016], the anonymized data can be used for the challenge.

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)

For the Brain Data the following Licence applies:

"I will not attempt to establish the identity of or attempt to contact any of the included human subjects.
I understand that under no circumstances will the code that would link these data to Protected Health Information be given to me, nor will any additional information about individual human subjects be released to me under these Open Access Data Use Terms.
I will comply with all relevant rules and regulations imposed by my institution. This may mean that I need my research to be approved or declared exempt by a committee that oversees research on human subjects, e.g. my IRB or Ethics Committee. The released HCP data are not considered de-identified, insofar as certain combinations of HCP Restricted Data (available through a separate process) might allow identification of individuals. Different committees operate under different national, state and local laws and may interpret regulations differently, so it is important to ask about this. If needed and upon request, the HCP will provide a certificate stating that you have accepted the HCP Open Access Data Use Terms.
I may redistribute original WU-Minn HCP Open Access data and any derived data as long as the data are redistributed under these same Data Use Terms.
I will acknowledge the use of WU-Minn HCP data and data derived from WU-Minn HCP data when publicly presenting any results or algorithms that benefitted from their use.
Papers, book chapters, books, posters, oral presentations, and all other printed and digital presentations of results derived from HCP data should contain the following wording in the acknowledgments section: "Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University."
Authors of publications or presentations using WU-Minn HCP data should cite relevant publications describing the methods used by the HCP to acquire and process the data. The specific publications that are appropriate to cite in any given study will depend on what HCP data were used and for what purposes. An annotated and appropriately up-to-date list of publications that may warrant consideration is available at http://www.humanconnectome.org/about/acknowledgehcp.html
The WU-Minn HCP Consortium as a whole should not be included as an author of publications or presentations if this authorship would be based solely on the use of WU-Minn HCP data.
Failure to abide by these guidelines will result in termination of my privileges to access WU-Minn HCP data."

For the Abdominal data the following Licence applies:

"Citations & Data Usage Policy
This collection is freely available to browse, download, and use for commercial, scientific and educational purposes as outlined in the Creative Commons Attribution 3.0 Unported License. See TCIA's Data Usage Policies and Restrictions for additional details. Questions may be directed to help@cancerimagingarchive.net.

Please be sure to include the following citations in your work if you use this data set:

Data Citation
Smith K, Clark K, Bennett W, Nolan T, Kirby J, Wolfsberger M, Moulton J, Vendt B, Freymann J. (2015). Data From CT_COLONOGRAPHY. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.NWTESAY1 "

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 evaluation code for the challenge is publicly released (furthermore, exemplary "ready to go" algorithms are provided as well). See https://github.com/MIC-DKFZ/mood .

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

All participating teams appear on the leaderboard (if they choose). However (to prevent misconduct) only teams that open source their code will be eligible to win the challenge and receive any prizes.

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.

Currently, there is no explicit sponsoring of the challenge (the workforce is currently funded by DKFZ).
Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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

Decision support, Diagnosis, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Out-of Distribution Detection (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.

The goal of this approach is to detect (incidental) findings and improve inattentional blindness.
Thus, anyone having an MRI/CT scan, for which a model was trained, would be in a target cohort.

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

For our challenge cohort the subjects are:
Brain: healthy young adult participants (see more: https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release)

Abdominal: Male and female outpatients, aged 50 years or older, scheduled for screening colonoscopy, who have not had a colonoscopy in the past 5 years (see more: https://wiki.cancerimagingarchive.net/display/Public/CT+COLONOGRAPHY).

Imaging modality(ies)

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

Brain: MRT (T2)
Abdominal: CT

Context information

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

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

None

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

None

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.

The advantage and goal of the challenge is to evaluate models that are not specific to a distinct body part or scan. So if enough data or an already trained model is available all body regions appearing in medical imaging scans (CT/MRI) could be used as a target for the algorithms.
However, given the available data, we only provide two different data origins, head MRI and abdominal CT. For those datasets, the same algorithm should be trained and evaluated.

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 in our case is anything that deviates from the given training distribution, i.e Anomalies / Out-of-(training)distribution images. In particular, the idea is to differentiate between natural anatomic deviations and pathological conditions, thus detecting anomalies/ abnormal parts. However, to ensure a controlled setting and allow for a fair and highly controlled evaluation of the algorithms, we use samples for testing that are derived from the same distribution as was used for training purposes. Anomalies will be artificially introduced

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)).

Detect samples that differ from the training data distribution, therefore assigning each sample a likelihood / an abnormality score.

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).

Brain:

"All HCP subjects are scanned on a customized Siemens 3T "Connectome Skyra" housed at Washington University

in St. Louis, using a standard 32-channel Siemens receive head coil and a "body" transmission coil designed by Siemens specifically for the smaller space available using the special gradients of the WU-Minn and MGH-UCLA Connectome scanners. The scanner has a customized SC72 gradient insert and a customized body transmitter coil with 56 cm bore size (diffusion: Gmax = 100 mT/m, max slew rate = 91 mT/m/ms; readout/imaging: Gmax = 42 mT/m, max slew rate = 200 mT/m/ms). The HCP Skyra has the standard set of Siemen's shim coils (up to 2nd order). Relative to a standard commercial Skyra, the customized hardware includes a gradient coil and gradient power amplifiers that together increase the maximum gradient strength from 40 mT/m to 100 mT/m on the WU-Minn 3T. This specifically benefits diffusion imaging, and on theoretical grounds (gurbil et al. 2013) it should provide significant gains over the standard 40 mT/m though not as much as the 300 mT/m customized gradients used by the MGH/UCLA HCP consortium. Head motion and physiological monitoring. To address head motion, in most scan sessions we acquire dynamic head position information using an optical motion tracking camera system (Moire Phase Tracker, Kineticor). This system monitors head position precisely and in real-time using an infrared camera mounted in the scanner bore. Images of Moire interference fringes on a target affixed by clay to the bridge of the subject's nose are streamed in real-time to a computer that displays the current position of the sensor and stores the positional information in a data file linked to the associated MRI scan. The stored file of head position and head movement can be used for post-hoc analyses. We also use it as a feedback trigger in dMRI scans to interrupt the movie being viewed whenever suprathreshold displacement and/or rapid head movement occur. Positional information can also be routed to the MRI scanner computer and can in principle be used prospectively to update the MRI slice prescription in real-time (Zaitsev et al., 2006). However, prospective motion correction is not part of our 3T HCP acquisition protocol because the technology becaame available only late in the HCP method development phase and was not sufficiently tested and developed before the data collection protocol was finalized (Van Essen et al., 2013). "

see
https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf for more information

Abdominal:
"A total of 15 clinical sites participated in this HIPAA-compliant study, and approval was obtained from the Institutional Review Board at each site prior to activation. Subjects were recruited from among all asymptomatic patients 50 years or older, prescheduled for routine colonoscopy at the participating sites between February 2005 and December 2006. Exclusion criteria were melena and/or hematochezia on more than one occasion in the previous six months, lower abdominal pain, inflammatory bowel disease and/or familial polyposis syndrome, serious medical conditions associated with excessive colonoscopy risk, colonoscopy within the previous five years, anemia (hemoglobin less than 10 gm/dl), or positive fecal occult blood test (FOBT). Each enrolled study participant provided written informed consent.

[...]

Supine and prone data acquisitions were obtained. All examinations were performed using at least a 16 slice CT scanner. Images were acquired using 0.5–1.0 mm collimation, pitch of 0.98–1.5, matrix 512 × 512, field-of-view to fit, 50 effective mAs, 120 kVp, and standard reconstruction algorithm. Images for prone and supine acquisitions were reconstructed to slice thicknesses of 1–1.25 mm with a 0.8 mm reconstruction interval.(7)
"

For more information see: https://wiki.cancerimagingarchive.net/display/Public/CT+COLONOGRAPHY

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

See a)

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.

See a)

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).

See a)

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.

Brain:
Training and test cases both represent MRI images of human brains.

Abdominal:
Training and test cases both represent CT images of human abdominal tracts.

The training cases have no annotations. The test cases either come from the same training distribution or another distribution e.g. by altering the images, i.e. introducing anomalies. For the test images, information on the original distribution labels and anomaly segmentations are maintained.

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

Brain: Training: 800, Testing 688, (4 for Feedback)
Abdominal: Training 550, Testing: 599, (4 for Feedback)

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

Since a part of our test-cases is artificially created, we can generate a high number of different test-cases (however the minimal amount of test cases is determined by the normal part of the test-case, and to prevent any fine-tuning of the scores to this ratio, we choose not to disclose the exact number of cases. However, if it would be useful, we can share this information with the reviewers, if they agree not to take part in the challenge). To get a fair

comparison with high significance, we aim for a 50%-50% split between training and test data. Considering the time needed for evaluation, this results in 800 training and 688 test cases for the brain dataset and 550 training and 599 test cases for the abdominal dataset (each testset having a fixed normal/abnormal split).

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.

Since we use publicly available data and to prevent misconduct we do not fully disclose our testing data distribution (since the goal is to detect potentially unknown anomalies). However, we are willing to share this information with the reviewers, if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

For the sample-wise task, a few selected naturally occurring conditions were chosen and excluded from the test set. The results were checked by at least two humans (checked individually, then a fixed common labeling protocol was agreed on and then all cases checked again). Furthermore the results were checked semi automatically multiple times.

However, the majority of the anomalies are generated artificially thus giving perfect labels.

The labels are binary (0 = normal, 1 = abnormal).

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 this would reveal too much information and to prevent cheating we do not fully disclose this information and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

Since this would reveal too much information and to prevent cheating we do not fully disclose this information and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

Since this would reveal too much information and to prevent cheating we do not fully disclose this information and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

For all sets, we perform the same preprocessing. We resample, crop and intensity-shift the images. This makes the preprocessed training samples and the preprocessed (unmodified) testing samples still originate from the same distribution. However, since we use publicly available data and to prevent cheating we do not fully disclose our preprocessing and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

We generate the anomalies artificially with full control over their shape and appearance. Thus, there should be no errors in the annotation. We have already tested the anomaly generation code for some time and believe that it is free of errors. However, since we only use the labels for testing, potential remaining bugs do not affect training and scores could be simply recomputed after fixing cases and labels. This would lead to an update of the leaderboard.

It is possible that "true" anomalies appear in the training set. This could potentially include cases such as polyps, that were not detected by a radiologist, or a patient with an abnormal kidney, that was overseen since it was not the indication for the examination. Such cases would be expected to be learned as "normal" since they are part of the training distribution. There is the highly unlikely case, in which an artificially introduced anomaly by coincidence is very similar to some of these "true" abnormalities missed during inspecting of the training set. Even if this unlikely case will occur, it will not influence the overall results too much given the size of the testset (and it is the same for all participants).

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

see above

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)

Average Precision (AP), which "summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight" (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html) . For more information see [9] or https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html.

Our evaluation code can be found here: https://github.com/MIC-DKFZ/mood

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

We chose AP since it is more robust to AUROC with regard to class imbalance (despite us being able to control it and keep it at 1/2 to 2/3) and has been suggested by many recent papers [2,3,4,6,9].
However, since this is the among most used and the often recommended metric we chose AP, but after the challenge might investigate different metrics, such as object-level detection metrics (choosing an abnormality threshold-based on a hold-out validation set similar to [23]) and present it in the paper as well as on the leaderboard (e.g. AUROC, FPR@95).

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.

For each sample, the users should report an anomaly score, indicating the certainty of detecting the anomaly in the given image. We expect the scores to be in the interval [0-1], where 0 is the lowest score indicating no abnormality and 1 is the highest score indicating the most abnormal input (Scores above and below the interval will be clamped to [0-1]). We will use the reported scores together with the ground truth labels to calculate the AP over the whole dataset. The ranking for one dataset will then be given by the AP over the whole dataset. We combine the two datasets by choosing a consolidation ranking schema.

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

Since we use docker based testing we assume that almost all participants will process all cases. However, should one case be not considered, we will assign it the lowest given anomaly score (= 0).

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

We chose AP since it is more robust to AUROC with regard to class imbalance (despite us being able to control it and keep it at 1/2 to 2/3) and has been suggested by many recent papers [2,3,4,6,9].
However, since this is the among most used and the often recommended metric we chose AP, but after the challenge might investigate different metrics, such as object-level detection metrics (choosing an abnormality threshold-based on a hold-out validation set similar to [23]) and present it in the paper as well as on the leaderboard (e.g. AUROC, FPR@95).

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.

Stability will be investigated via bootstrapping and hypothesis testing (using challengeR, code available at github: https://github.com/MIC-DKFZ/mood, results from last year available at: http://medicalood.dkfz.de/docs/).

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

Bootstrapping was identified as an appropriate approach to investigate ranking variability in [24, Maier-Hein et al 2018].

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.

Since we can control and differentiate different kinds, scales, and intensities of anomalies, even for the same test cases, similar to last year we plan to we further plan to investigate which algorithm is able to detect which kind of anomaly reliable or with less certainty.

TASK: PixelLevel

SUMMARY

Keywords

List the primary keywords that characterize the task.

Anomaly detection, Anomaly localization, Out-of-distribution detection, novelty detection

ORGANIZATION

Organizers

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

Jens Petersen, Gregor Köhler, Paul Jäger, Peter Full, David Zimmerer, Klaus Maier-Hein
Div. Medical Image Computing (MIC), German Cancer Research Center (DKFZ)

Tobias Roß, Tim Adler, Annika Reinke, Lena Maier-Hein
Div Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ)

b) Provide information on the primary contact person.

David Zimmerer (MIC, DKFZ) d.zimmerer@dkfz.de

Life cycle type

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

Examples:

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

Repeated event with fixed submission deadline.

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.

Synapse.org

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

https://www.synapse.org/#!Synapse:syn21343101/wiki/599515

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.

No prices yet determined. Currently, the winner will win the right to brag about the results. (last year's winner received a NVidia RTX 3080)

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.

By default (i.e. if they don't decline) the top 3 (winning) teams will be announced publicly. The remaining teams may decide if they choose to appear on the public Leaderboard.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All participating teams with a valid submission. Teams that reveal their identity, can nominate members of their team as co-author for the challenge paper. However, we reserve the right to exclude teams and team members if they do not adhere to the challenge rules and guidelines.

The method description submitted by the authors may be used in the publication of the challenge results. Personal data of the authors will include their name, affiliation and contact addresses.

Participating teams may publish their own 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.

Docker container on the Synapse platform. Link to submission instructions:
https://www.synapse.org/#!Synapse:syn21989704 or medicalood.dkfz.de/web/

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Teams are allowed three submissions per challenge task, however, only the latest submission will be considered. For each challenge submission, the teams will receive the results on a very simple "toy" test-case, which will not be included in the final testset. We will also make the "toy" case and testing script publicly available so that the participants can check beforehand if their submission won't cause any errors and to compare the reported results with their own to check for any additional occurring errors. Additionally a dedicated toy-case evaluation queue will be available where participants can submit their solutions to our cluster and only run the toy cases and receive toy-case results. This toy-case evaluation queue will not count as a challenge submission and will have no submission limit and only serve as a "will run successfully" check.

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 cases: 1 April 2021
Registration Until: 06 September 2021
Submission of Dockers: 07 September 2021

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).

The provided challenge data is anonymized.
According to [Euro2016], the anonymized data can be used for the challenge.

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)

For the Brain Data the following Licence applies:

"I will not attempt to establish the identity of or attempt to contact any of the included human subjects.
I understand that under no circumstances will the code that would link these data to Protected Health Information be given to me, nor will any additional information about individual human subjects be released to me under these Open Access Data Use Terms.
I will comply with all relevant rules and regulations imposed by my institution. This may mean that I need my research to be approved or declared exempt by a committee that oversees research on human subjects, e.g. my IRB or Ethics Committee. The released HCP data are not considered de-identified, insofar as certain combinations of HCP Restricted Data (available through a separate process) might allow identification of individuals. Different committees operate under different national, state and local laws and may interpret regulations differently, so it is important to ask about this. If needed and upon request, the HCP will provide a certificate stating that you have accepted the HCP Open Access Data Use Terms.
I may redistribute original WU-Minn HCP Open Access data and any derived data as long as the data are redistributed under these same Data Use Terms.
I will acknowledge the use of WU-Minn HCP data and data derived from WU-Minn HCP data when publicly presenting any results or algorithms that benefitted from their use.
Papers, book chapters, books, posters, oral presentations, and all other printed and digital presentations of results derived from HCP data should contain the following wording in the acknowledgments section: "Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University."
Authors of publications or presentations using WU-Minn HCP data should cite relevant publications describing the methods used by the HCP to acquire and process the data. The specific publications that are appropriate to cite in any given study will depend on what HCP data were used and for what purposes. An annotated and appropriately up-to-date list of publications that may warrant consideration is available at http://www.humanconnectome.org/about/acknowledgehcp.html
The WU-Minn HCP Consortium as a whole should not be included as an author of publications or presentations if this authorship would be based solely on the use of WU-Minn HCP data.
Failure to abide by these guidelines will result in termination of my privileges to access WU-Minn HCP data."

For the Abdominal data the following Licence applies:

"Citations & Data Usage Policy
This collection is freely available to browse, download, and use for commercial, scientific and educational purposes as outlined in the Creative Commons Attribution 3.0 Unported License. See TCIA's Data Usage Policies and Restrictions for additional details. Questions may be directed to help@cancerimagingarchive.net.

Please be sure to include the following citations in your work if you use this data set:

Data Citation
Smith K, Clark K, Bennett W, Nolan T, Kirby J, Wolfsberger M, Moulton J, Vendt B, Freymann J. (2015). Data From CT_COLONOGRAPHY. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.NWTESAY1 "

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 evaluation code for the challenge is publicly released (furthermore, exemplary "ready to go" algorithms are provided as well). See https://github.com/MIC-DKFZ/mood .

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

All participating teams appear on the leaderboard (if they choose). However (to prevent misconduct) only teams that open source their code will be eligible to win the challenge and receive any prizes.

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.

Currently, there is no explicit sponsoring of the challenge (the workforce is currently funded by DKFZ).
Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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

Decision support, Diagnosis, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Out-of Distribution Detection (Localization)

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 goal of this approach is to detect (incidental) findings and improve inattentional blindness.
Thus, anyone having an MRI/CT scan, for which a model was trained, would be in a target cohort.

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

For our challenge cohort the subjects are:
Brain: healthy young adult participants (see more: https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release)

Abdominal: Male and female outpatients, aged 50 years or older, scheduled for screening colonoscopy, who have not had a colonoscopy in the past 5 years (see more: https://wiki.cancerimagingarchive.net/display/Public/CT+COLONOGRAPHY).

Imaging modality(ies)

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

Brain: MRT (T2)
Abdominal: CT

Context information

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

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

None

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

None

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.

The advantage and goal of the challenge is to evaluate models that are not specific to a distinct body part or scan. So if enough data or an already trained model is available all body regions appearing in medical imaging scans (CT/MRI) could be used as a target for the algorithms.
However, given the available data, we only provide two different data origins, head MRI and abdominal CT. For those datasets, the same algorithm should be trained and evaluated.

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 in our case is anything that deviates from the given training distribution, i.e Anomalies / Out-of-(training)distribution images. In particular, the idea is to differentiate between natural anatomic deviations and pathological conditions, thus outlining anomalies/ abnormal parts. However, to ensure a controlled setting and allow for a fair and highly controlled evaluation of the algorithms, we use samples for testing that are derived from the same distribution as was used for training purposes. Anomalies will be artificially introduced

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)).

Find and localize anomalies in test cases, by assigning each voxel a likelihood / an abnormality score.

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).

Brain:

"All HCP subjects are scanned on a customized Siemens 3T "Connectome Skyra" housed at Washington University in St. Louis, using a standard 32-channel Siemens receive head coil and a "body" transmission coil designed by

Siemens specifically for the smaller space available using the special gradients of the WU-Minn and MGH-UCLA Connectome scanners. The scanner has a customized SC72 gradient insert and a customized body transmitter coil with 56 cm bore size (diffusion: Gmax = 100 mT/m, max slew rate = 91 mT/m/ms; readout/imaging: Gmax = 42 mT/m, max slew rate = 200 mT/m/ms). The HCP Skyra has the standard set of Siemen's shim coils (up to 2nd order). Relative to a standard commercial Skyra, the customized hardware includes a gradient coil and gradient power amplifiers that together increase the maximum gradient strength from 40 mT/m to 100 mT/m on the WU-Minn 3T. This specifically benefits diffusion imaging, and on theoretical grounds (gurbil et al. 2013) it should provide significant gains over the standard 40 mT/m though not as much as the 300 mT/m customized gradients used by the MGH/UCLA HCP consortium. Head motion and physiological monitoring. To address head motion, in most scan sessions we acquire dynamic head position information using an optical motion tracking camera system (Moire Phase Tracker, Kineticor). This system monitors head position precisely and in real-time using an infrared camera mounted in the scanner bore. Images of Moire interference fringes on a target affixed by clay to the bridge of the subject's nose are streamed in real-time to a computer that displays the current position of the sensor and stores the positional information in a data file linked to the associated MRI scan. The stored file of head position and head movement can be used for post-hoc analyses. We also use it as a feedback trigger in dMRI scans to interrupt the movie being viewed whenever suprathreshold displacement and/or rapid head movement occur. Positional information can also be routed to the MRI scanner computer and can in principle be used prospectively to update the MRI slice prescription in real-time (Zaitsev et al., 2006). However, prospective motion correction is not part of our 3T HCP acquisition protocol because the technology becaame available only late in the HCP method development phase and was not sufficiently tested and developed before the data collection protocol was finalized (Van Essen et al., 2013). "

see
https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf for more information

Abdominal:
"A total of 15 clinical sites participated in this HIPAA-compliant study, and approval was obtained from the Institutional Review Board at each site prior to activation. Subjects were recruited from among all asymptomatic patients 50 years or older, prescheduled for routine colonoscopy at the participating sites between February 2005 and December 2006. Exclusion criteria were melena and/or hematochezia on more than one occasion in the previous six months, lower abdominal pain, inflammatory bowel disease and/or familial polyposis syndrome, serious medical conditions associated with excessive colonoscopy risk, colonoscopy within the previous five years, anemia (hemoglobin less than 10 gm/dl), or positive fecal occult blood test (FOBT). Each enrolled study participant provided written informed consent.

[...]

Supine and prone data acquisitions were obtained. All examinations were performed using at least a 16 slice CT scanner. Images were acquired using 0.5–1.0 mm collimation, pitch of 0.98–1.5, matrix 512 × 512, field-of-view to fit, 50 effective mAs, 120 kVp, and standard reconstruction algorithm. Images for prone and supine acquisitions were reconstructed to slice thicknesses of 1–1.25 mm with a 0.8 mm reconstruction interval.(7)
"

For more information see: https://wiki.cancerimagingarchive.net/display/Public/CT+COLONOGRAPHY

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

See a)

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.

See a)

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).

See a)

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.

Brain:
Training and test cases both represent MRI images of human brains.

Abdominal:
Training and test cases both represent CT images of human abdominal tracts.

The training cases have no annotations. The test cases either come from the same training distribution or another distribution e.g. by altering the images, i.e. introducing anomalies. For the test images, information on the original distribution labels and anomaly segmentations are maintained.

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

Brain: Training: 800, Testing 542, (4 for Feedback)
Abdominal: Training 550, Testing: 358, (4 for Feedback)

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

Since a part of our test-cases is artificially created, we can generate a high number of different test-cases (however the minimal amount of test cases is determined by the normal part of the test-case, and to prevent any fine-tuning of the scores to this ratio, we choose not to disclose the exact number of cases. However, if it would be useful, we can share this information with the reviewers, if they agree not to take part in the challenge). To get a fair comparison with high significance, we aim for a 50%-50% split between training and test data. Considering the time needed for evaluation, this results in 800 training and 542 test cases for the brain dataset and 550 training

and 358 test cases for the abdominal dataset (each testset having a fixed normal/abnormal split).

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.

Since we use publicly available data and to prevent misconduct we do not fully disclose our testing data distribution (since the goal is to detect potentially unknown anomalies). However, we are willing to share this information with the reviewers, if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

For the sample-wise task, a few selected naturally occurring conditions were chosen and excluded from the test set. The results were checked by at least two humans (checked individually, then a fixed common labeling protocol was agreed on and then all cases checked again). Furthermore the results were checked semi automatically multiple times.

For the voxel-wise case, annotations are generated by artificially introducing anomalies to the images. This allows for perfect ground truth in the voxel-wise scoring of anomalies.

The labels are binary (0 = normal, 1 = abnormal).

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 this would reveal too much information and to prevent cheating we do not fully disclose this information and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

Since this would reveal too much information and to prevent cheating we do not fully disclose this information and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

Since this would reveal too much information and to prevent cheating we do not fully disclose this information and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

For all sets, we perform the same preprocessing. We resample, crop and intensity-shift the images. This makes the preprocessed training samples and the preprocessed (unmodified) testing samples still originate from the same distribution. However, since we use publicly available data and to prevent cheating we do not fully disclose our preprocessing and are willing you share it with the reviewers if it would be helpful (and if the reviewers agree not to take part in the challenge).

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.

We generate the anomalies artificially with full control over their shape and appearance. Thus, there should be no errors in the annotation. We have already tested the anomaly generation code for some time and believe that it is free of errors. However, since we only use the labels for testing, potential remaining bugs do not affect training and scores could be simply recomputed after fixing cases and labels. This would lead to an update of the leaderboard.

It is possible that "true" anomalies appear in the training set. This could potentially include cases such as polyps, that were not detected by a radiologist, or a patient with an abnormal kidney, that was overseen since it was not the indication for the examination. Such cases would be expected to be learned as "normal" since they are part of the training distribution. There is the highly unlikely case, in which an artificially introduced anomaly by coincidence is very similar to some of these "true" abnormalities missed during inspecting of the training set. Even if this unlikely case will occur, it will not influence the overall results too much given the size of the testset (and it is the same for all participants).

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

see above

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)

Average Precision (AP), which "summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight" (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html) . For more information see [9] or https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html.

Our evaluation code can be found here: https://github.com/MIC-DKFZ/mood

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

We chose AP since it is more robust to AUROC with regard to class imbalance (despite us being able to control it and keep it at 1/2 to 2/3) and has been suggested by many recent papers [2,3,4,6,9].
However, since this is the among most used and the often recommended metric we chose AP, but after the challenge might investigate different metrics, such as object-level detection metrics (choosing an abnormality threshold-based on a hold-out validation set similar to [23]) and present it in the paper as well as on the leaderboard (e.g. AUROC, FPR@95).

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.

For each voxel, the users should report an anomaly score, indicating the certainty of detecting the anomaly for the given voxels. We expect the scores to be in the interval [0-1], where 0 is the lowest score indicating no abnormality and 1 is the highest score indicating the most abnormal input (Scores above and below the interval will be clamped to [0-1]). We will use the reported scores together with the ground truth labels to calculate the AP over the whole dataset. Due to computational and time constraints in the pixel-level task, we go over the dataset in batches of 20 samples (randomly chosen but fixed for all participants) and average the performance. To offset some additional variance we go over the dataset twice. The ranking for one dataset will then be given by the AP over the whole dataset. We combine the two datasets by choosing a consolidation ranking schema.

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

Since we use docker based testing we assume that almost all participants will process all cases. However, should one case be not considered, we will assign it the lowest given anomaly score (= 0).

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

We chose AP since it is more robust to AUROC with regard to class imbalance (despite us being able to control it and keep it at 1/2 to 2/3) and has been suggested by many recent papers [2,3,4,6,9].
However, since this is the among most used and the often recommended metric we chose AP, but after the challenge might investigate different metrics, such as object-level detection metrics (choosing an abnormality threshold-based on a hold-out validation set similar to [23]) and present it in the paper as well as on the leaderboard (e.g. AUROC, FPR@95).

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.

Stability will be investigated via bootstrapping and hypothesis testing (using challengeR, code available at github: https://github.com/MIC-DKFZ/mood, results from last year available at: http://medicalood.dkfz.de/docs/).

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

Bootstrapping was identified as an appropriate approach to investigate ranking variability in [24, Maier-Hein et al 2018].

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.

Since we can control and differentiate different kinds, scales, and intensities of anomalies, even for the same test cases, similar to last year we plan to we further plan to investigate which algorithm is able to detect which kind of anomaly reliable or with less certainty.

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.