MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images

David Zimmerer, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, Bjørn Sand Jensen, Alison Q. O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V. Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio Naval Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, Klaus Maier-Hein

***Abstract*— Detecting Out-of-Distribution (OoD) data is one of the greatest challenges in safe and robust deployment of machine learning algorithms in medicine. When the algorithms encounter cases that deviate from the distribution of the training data, they often produce incorrect and over-confident predictions. OoD detection algorithms aim to catch erroneous predictions in advance by analysing the data distribution and detecting potential instances of failure. Moreover, flagging OoD cases may support human readers in identifying incidental findings. Due to the increased interest in OoD algorithms, benchmarks for different domains have recently been established. In the medical imaging domain, for which reliable predictions are often essential, an open benchmark has been missing. We introduce the Medical-Out-Of-Distribution-Analysis-Challenge (MOOD) as an open, fair, and unbiased benchmark for OoD methods in the medical imaging domain. The analysis of the submitted algorithms shows that performance has a strong positive correlation with the perceived difficulty, and that all algorithms show a high variance for different anomalies, making it yet hard to recommend them for clinical practice. We also see a strong correlation between challenge ranking and performance on a simple toy test set, indicating that this might be a valuable addition as a proxy dataset during anomaly detection algorithm development.**

***Index Terms*— Anomaly Detection, Anomaly Localization, Biomedical Challenge, Out-of-Distribution Analysis.**

Submitted on 05.11.2021. The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health" and Helmholtz Imaging

D. Zimmerer is with the German Cancer Research Center, Heidelberg, Germany (e-mail: d.zimmerer@dkfz.de).

P. M. Full is with with German Cancer Research Center, Heidelberg, Germany and Heidelberg University, Heidelberg, Germany.

F. Isensee, P. Jaeger, T. Adler, J. Petersen, G. Koehler, T. Ross, A. Reinke, L. Maier-Hein, K. Maier-Hein are with German Cancer Research Center, Heidelberg, Germany.

A. Kascenas, B. Jensen are with University of Glasgow, Glasgow, UK.

A. O'Neil is with University of Edinburgh, Edinburgh, UK.

J. Tan, B. Hou, J. Batten, H. Qiu, B. Kainz are with Imperial College London, London, UK.

N. Shvetsova, I. Fedulova are with Philips Research, Moscow, Russia.

D. Dylov is with Skolkovo Institute of Science and Technology, Moscow, Russia.

B. Yu, J. Zhai, J. Hu, R. Si, S. Wang, X. Li, X. Chen, Y. Zhao are with College of Computer, National University of Defense Technology, Hunan, China.

S. Zhou is with College of Intelligence Science and Technology, National University of Defense Technology, Hunan, China.

S. Marimont is with CitAI Research Centre, University of London, London, UK.

G. Tarroni is with BioMedIA, Imperial College, London, UK.

V. Saase is with Heidelberg University, Heidelberg, Germany.

I. Introduction

The amount of medical images acquired in clinical routine doubled between 1997 and 2006 and continues to rise [1], [2]. At the same time, the review and annotation process for the acquired images is often prohibitively expensive due to its reliance on the valuable time of domain experts. Consequently, computer-assisted diagnosis systems are becoming more popular in the clinical workflow [3], [4]. However, many of the algorithms used in image analysis are vulnerable to Out-of-Distribution samples, resulting in wrong and overconfident decisions [5]–[8]. In addition, physicians overlook unexpected conditions in medical images, often termed 'inattentional blindness'. Indeed, [9] found that 50% of trained radiologists did not notice a gorilla image rendered into a lung CT scan when assessing lung nodules.

Out-of-Distribution (OoD) or anomaly detection, two terms which are used interchangeably in this context, can, trained on normal or representative data, recognize anomalies that have not been previously encountered. Therefore, OoD methods prove useful in situations where classic machine learning models may fail. By highlighting abnormal regions, anomaly detection can also guide the physician's attention to otherwise overlooked abnormalities in a scan and potentially reduce the time required to inspect medical images. Circumventing the need for labeled data, it can also sidestep the time-consuming labeling process and can therefore quickly be adapted to new modalities.

However, while there is much recent research on improving anomaly detection [10]–[18], some of which is focused on the medical imaging field [19]–[23], a publicly available dataset and benchmark to compare different approaches is missing.

Thus, currently, it is hard to draw a fair comparison of the various proposed approaches. While medical imaging still needs a common benchmark, benchmarks for tabular medical data [24], [25] as well as natural images, such as default detection [26] or abnormal traffic scene detection [27], have recently been proposed.

When designing an OoD detection benchmark in the medical imaging field, various additional aspects must be considered. First, as is the case in a real-life setting, the types of anomalies or distribution shifts appearing during application should not be known beforehand. This often proves an issue when choosing a dataset and testing it on only one single pathological condition, because this scenario is vulnerable to exploitation: if the type of anomalies occurring in the test set is known, one could perform fully supervised training on a separate dataset with the respective annotations (although this is prohibited by the challenge rules), and thus outperform other correctly trained anomaly detection approaches. This would lead to less robust algorithms scoring higher on the test set, a potentially dangerous outcome when deploying such algorithms in practice. Furthermore, making the exact types of anomalies known can cause a bias in the evaluation. Studies have shown that anomaly detection algorithms tend to overfit on a given task, if properties of the test set and types of anomalies are known beforehand [8], [23], [28], [29]. This further hinders the comparability of different algorithms. Secondly, combining test sets from different sources may also make it difficult to obtain a clean and meaningful evaluation, since different sources typically convey distribution shifts with respect to the training dataset due to large variations across medical image acquisition protocols.

In this work, we put forth the Medical-Out-of-Distribution-Analysis-Challenge (MOOD) as a standardized dataset and benchmark for anomaly detection. We propose two different tasks. In one task, we analyze sample-wise (i.e. patient-wise) anomalies, thus detecting OoD samples. Examples of anomalies in this task are previously unseen pathological conditions or any other condition not apparent in the training set. These phenomena can pose a problem for supervised algorithms. Robust identification of such cases could, for example, allow physicians to distrust results obtained from supervised algorithms or prioritize manual inspection of certain patients. As a second task, we propose a pixel-level analysis, i.e., predicting an anomaly score for each individual pixel, thereby highlighting regions with abnormal conditions in the image and providing further guidance to the physician.

To solve the previously described issues, we have provided two separate datasets containing over 500 scans each: one brain MRI-dataset and one abdominal CT-dataset. This enables a sound comparison of the generalization capabilities of submissions to be drawn across different anatomies and modalities. The training set was selected as a subset of scans in which no anomalies were identified. The remaining scans (some containing anomalies) were assigned to the test set. Thus, some scans in the test set did not contain anomalies, while others contained naturally occurring anomalies. In addition to the natural anomalies, we also added synthetic anomalies with different structures (e.g. a tumor or an image of a gorilla rendered into the brain scan [9]). We thus covered a wide variety of different anomalies which enabled the weaknesses and strengths of the methods to be analyzed using different factors (i.e. type, size, contrast, and others). Finally, we organized an international open challenge for a controlled and fair comparison of different algorithms (as recently similarly proposed by [27]). As a whole, this work effects a standardized comparison of anomaly detection approaches in a variety of both real-life and simulated cases. The following sections describe the data used in the challenge and the challenge setup. In Section IV the submitted approaches are described by the participants and the results are presented in Section V, which are discussed in Section VI.

II. Data

The challenge encompasses two datasets one brain MRI-dataset and one abdominal CT-dataset. The training set comprises hand selected scans of patients with no apparent anomalies or patients with common anatomical or pathological variations.

To prevent overfitting on the (types of) anomalies present in our test set, the test set was kept confidential at all times. As in reality, the types of anomalies should not be and were not known beforehand, to prevent a bias towards certain anomalies in the evaluation. Some scans in the test set did not contain any anomalies, while others contain naturally occurring or synthetic anomalies.

A. Datasets

Challenge participants were required to use the same algorithm/approach for both challenge datasets, but, individual hyperparameters and training on each dataset was allowed. Furthermore, we calculated the scores and ranking separately for each dataset, and combined the ranking using a consensus ranking.

Brain: Training and test cases both show MRI images of a human brain. The brain dataset is based on the HCP dataset [30], contains 3T MR imaging data from healthy young adult participants (ages 22-35). All participants were scanned on the same equipment and using the same protocol. The data was processed following the pipeline given in [30].

Abdominal: Training and test cases both show CT images of human abdominal tracts. For the study, male and female patients aged 50 years or older scheduled for a screening colonoscopy and which had not had a colonoscopy in the past 5 years were scanned at 15 study centers [31]. CT colonographic images were acquired using standard bowel preparation, stool and fluid tagging, mechanical insufflation, and multi-detector row CT scanners (with 16 or more rows). Consequently, these images may contain polyps, however, these were not considered abnormal (due to the training distribution) and only cases with severe or rare naturally occurring anomalies were considered to be abnormal.

B. Challenge Preprocessing

We applied the same additional challenge-specific preprocessing for both datasets. The transformations were cropping,

intensity shift and resampling. Since all patients within the same dataset were preprocessed using the same parametrization of our pipeline, there was no distribution shift here between the training and test cases. To prevent cheating (in this and future editions[1]), we will not disclose the exact details of our preprocessing and intentionally designed our preprocessing to produce a challenge dataset which is clearly distinct from both the original dataset and other existing datasets.

C. Anomalies

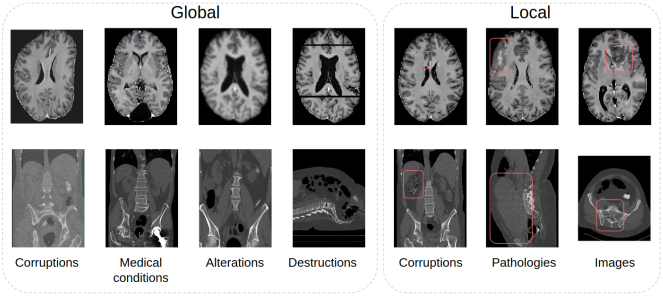

Fig. 1: Anomaly categories. The seven different categories of anomalies are presented here, divided in 4 global (affecting the whole scans) and 3 local (affecting only parts of the image) categories, visualized with brain and abdominal scan examples (some anomalies have been exaggerated for illustration purposes).

The training cases had no annotations and no conditions that we considered to be abnormal. The test cases either originated from the same training data distribution (normal data samples with no abnormal conditions) or from a different distribution (OoD data samples, i.e., exhibiting natural and synthetic abnormal conditions). The corresponding ground-truth labels for test cases were binary (0 = normal, 1 = abnormal/OoD).

The majority of the OoD data samples were generated by artificially modifying normal data samples, thus providing full information on the properties of the abnormality for those cases. In addition, a few selected naturally occurring conditions were excluded from the training set and added to the test set of OoD samples for the sample-level task. These conditions were checked multiple times by at least two human raters using a consensus annotation protocol. Since we plan to run new editions of the challenge and a continuous online benchmark we refrain from giving exact details on the anomalies.

We differentiated between local (specific location in the image, used in the test set for the sample-level and pixel-level task) and global (no specific location in the image, i.e. only used in the test set for the sample-level task) anomalies and sorted the anomalies into different (subjective) categories, see Fig. 1.

[1]MOOD currently is/was held in conjunction with MICCAI in 2020, 2021, and 2022

For the pixel-level case, annotations were generated by artificially introducing anomalies to the images locally. This enabled perfect ground truth to be obtained in the pixel-level scoring of the anomalies. For the local anomalies, we created the following categories:

- **Images**: Similar to [9], we rendered natural images into the scans.
- **(local) Pathologies**: We added different local pathologies such as tumors or lesions, to the healthy images.
- **Corruptions**: Local corruptions to the image, such as local contrast change or local pixel shuffling.

For the global anomalies, we created the following categories (sorted from strong to mild by level of corruption on the images):

- **Destructions**: Operations performed on the scan makes the complete scan corrupt or invalid, e.g. by omitting slices.
- **Alterations**: Global level alteration to the scan, which still results in a valid scan but should be directly noticeable, e.g., heavy blurring.
- **(global) Medical conditions**: Rare occurring medical conditions/variations were considered as global variations, as these abnormalities were often not to be restrained to a certain area.
- **Corruptions**: Small corruptions in the image which produce a valid image and are only recognizable using a vast amount of training data, such as deformations.

Despite our controlled setting, different sources of errors are related to our annotations. True anomalies may 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 overlooked since it was not the indication for the examination. The system would thus learn these cases and consider them to be normal since they are part of the training distribution. It could also be that an artificially introduced anomaly is, coincidentally, very similar to some of the true abnormalities which are missed during inspection of the training set. This is very unlikely, but if it does occur, we believe it will not influence the overall results too much given the size of the test set (and the fact that it is identical for all participants). We generated the anomalies artificially using software that undergoes stringent in-house testing with full control over their shape and appearance. Thus, we strongly believe that there are no errors in the annotation.

III. Challenge Setup

The MOOD Challenge was run as a MICCAI 2020 Challenge, and as such the challenge design was reviewed beforehand according to the MICCAI Challenge guidelines (two independent reviews and a meta review). The challenge design document [32] is available online. The MOOD Challenge consisted of two tasks, referred to as sample-level (or *global*) and pixel-level (or *local*) task respectively:

Sample-level Analyzing different scans/samples and reporting a score for each sample. The algorithm should process a single sample and give a "probability" indicating how likely it is that this sample is abnormal/OoD. The reported scores must

be within the range of [0-1], where 0 indicates no abnormality and 1 indicates the most abnormal input. In summary: one score per sample. Scores outside [0-1] were clamped to [0-1].

Pixel-level Analyzing different scans and reporting a score for each pixel of the sample (we use the term pixel here in analogy to anomaly detection on natural images even if voxel would be more appropriate). The algorithm should process a single sample and give a "probability" indicating, for each pixel, how likely it is that the pixel is abnormal/OoD. The reported scores must be within the range of [0-1], where 0 indicates no abnormality and 1 indicates the most abnormal input. In summary: $X \times Y \times Z$ scores per sample (where $X \times Y \times Z$ is the dimensionality of the data sample).

A. Dataset ratios

Since part of our test set was artificially created, we were able to generate a high number of different test cases. To prevent any fine-tuning of the scores on the normal/abnormal ratio, we chose not to disclose the exact number of cases. We roughly aimed for a 50%-50% split between training and test data. Considering the number of available samples and the time needed for evaluation, we opted for 800 training, 688 sample-level and 542 pixel-level test cases for the brain dataset, and for 550 training, 599 sample-level and 358 pixel-level test cases for the abdominal dataset (with each test set containing normal and abnormal samples and having an individual and fixed normal/abnormal ratio).

B. Evaluation process

The challenge submission was run via the synapse platform [33]. Test set submissions were made by submitting the inference code as a self-contained docker container which was then applied to the test set. Detailed submission explanations can be found on the challenge website [34] Thus participants could not get access to the test data at any time during the challenge. In case of a missing reported score or failure during the processing of a sample, the lowest possible anomaly score (= 0) was assigned to that sample. A runtime of 600 sec/case was allotted for the evaluation during the evaluation. Teams were allowed 10 submissions in total, however, only the most recent submission was considered, as previously announced.

A report of the submission was sent to the participants as soon as the submission was processed. This report contained the performance/scores on four toy-cases for each dataset and the computation time needed to process them. The toy-cases were not used in the challenge test set and consist of three scans with toy anomalies, i.e. a sphere with random intensity placed into a scan, and one normal scan. The toy-cases were made publicly available. In addition to a challenge submission, the participants could also make a submission on the toy dataset for development purposes (both algorithmic as well as containerized). Submitting to the toy-cases did not count towards the challenge and only returned the report of the toy-cases. This was done to eliminate "invalid" submissions, since the participants had access to the toy-cases scores and thus could validate the consistency of their submission on the evaluation platform.

C. Metrics & scoring

For each sample/pixel, the users should have reported an anomaly score, indicating the likelihood of detecting the anomaly for the given sample/pixels. We expected 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 were clamped to [0-1]). We used the predicted scores together with the ground truth labels to calculate the Average Precision (AP) for the whole dataset.

AP, which "summarizes a precision-recall curve as the weighted mean of the precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight" [35] and is calculated as follows:

$$AP = \sum_n (R_n - R_{n-1}) P_n, \tag{1}$$

where R_n is the recall and P_n is the precision at the n-th threshold. This is basically a finite approximation of the area under the precision-recall curve. For more information see [35].

A key advantage of the AP compared to other metrics is the fact that it does not require users to set a threshold for an output to be in or out of distribution. Instead, the metric integrates over all possible thresholds. Since it is more robust than AUROC in terms of class imbalance and has been suggested and used in many recent papers [11], [19], [23], [27], [36], we opted to implement AP as the primary metric.

For the sample-level task, the score was simply computed over all samples at once. Due to computational and time constraints in the pixel-level task, we computed the AP in batches of 20 samples each (randomly chosen but fixed and consistent across all submissions) instead of the whole dataset and then averaged the AP over the batched AP values. To validate the results and test the additional variance due to the division in batches (which is equivalent to sub-sampling points from the precision-recall curve and then calculating the mean AP, instead of calculating the AP over the whole dataset i.e. all points on the precision-recall curve), we validated the results with an additional randomized iteration over the dataset.

As a last step, we combined the rankings corresponding to the two datasets (brain and abdominal) by choosing a consolidation ranking schema, i.e. "determining the ranking that minimizes the sum of the distances of the separate rankings" [37].

Our validation code to reproduce the results on the toy cases can be found on our Github page [38].

IV. Participating teams

Overall, 65 participants registered with 11 actively participating, which resulted in 8 valid submissions for each task. All teams with valid submissions were invited to contribute to this section. In the following, a description of the submissions, as provided by the respective teams, is given. Teams which chose not to participate were anonymized for the later analysis (A1, A2).

A. Team: Canon Medical Research Europe

We propose an ensemble of two models. The first model is a denoising Autoencoder neural network, in which we treat the pixel-level reconstruction errors as the anomaly scores. The second model is a segmentation neural network trained to segment our diverse set of synthesised anomalies, for which we treat the segmentation class probabilities as the anomaly scores. The models are ensembled by averaging the respective scaled anomaly scores to obtain the final pixel-level results. We produce the sample-level results by averaging the pixel-level anomaly scores in each sample.

B. Team: FPI

In medical imaging, outliers can contain hypo/hyper-intensities, minor deformations, or completely altered anatomy. To detect these irregularities it is helpful to learn the features present in both normal and abnormal images. However, this is difficult because of the wide range of possible abnormalities and also the number of ways that normal anatomy can vary naturally. As such, we leverage the natural variations in normal anatomy to create a range of synthetic abnormalities. Specifically, the same patch region is extracted from two independent samples and replaced with an interpolation between both patches. The interpolation factor, patch size, and patch location are randomly sampled from uniform distributions. A wide residual encoder decoder is trained to give a pixel-wise prediction of the patch and its interpolation factor. This encourages the network to learn which features to expect normally and to identify where foreign patterns have been introduced. The estimate of the interpolation factor lends itself nicely to the derivation of an outlier score. Meanwhile, the pixel-wise output allows for pixel- and subject-level predictions to be made using the same model [39].

C. Team: Nina Tuluptceva

We based our solution on a Deep Perceptual Autoencoder [40] that had recently shown superior performance in the anomaly detection task on medical images. The Autoencoder was trained with the content-aware Perceptual Loss [41], with the reconstruction error being treated as the abnormality score. In this challenge, we applied the Deep Perceptual Autoencoder to 2D slices of the 3D volumes and therefore trained three separate models along each of the three axes. The prediction outcomes along each axis were then averaged to yield a single final abnormality score. The Perceptual Loss calculates the difference between two images as the distance between the deep features extracted by a pre-trained network. We used the VGG19 network [42] as a feature extractor trained using the unsupervised learning framework SimCLR [43] on the concatenated set of all slices. To calculate pixel-level abnormality scores, we averaged feature differences along the depth axis and then rescaled the map to the original image size.

D. Team: NUDT

To tackle these problems provided in this challenge, we opted for a reconstruction strategy to solve the anomaly detection task. By observing the discriminative reconstruction errors, we noted that the biomedical images with high reconstruction losses were most likely to be the abnormalities. Therefore, we adopt an U-Net architecture, which has an encoder-decoder structure with skip connections, to reconstruct the image. Moreover, we combine the image with the texture features extracted by a Canny operator and apply a masking-and-in-painting task. The score consists of the reconstruction errors, removing objects smaller than 100 voxels.

E. Team: Sergio Naval Marimont et al.

We propose an Out-of-Distribution detection method that combines density and restoration-based approaches using Vector-Quantized Variational Autoencoders (VQ-VAEs) [44]. The VQ-VAE model learns to encode images in a categorical latent space. The prior distribution of latent codes is then modelled using an Auto-Regressive (AR) model [45]. We found that the prior probability estimated by the AR model can be useful for unsupervised anomaly detection and enables the estimation of both sample and pixel-wise anomaly scores. The sample-wise score is defined as the negative log-likelihood of the latent variables above a threshold. Additionally, OoD images are restored as in-distribution images by replacing unlikely latent codes with samples from the prior model and decoding to pixel space. The average L1 distance between the generated restorations and the original image is used as the pixel-wise anomaly score [46].

F. Team: Victor Saase

We use a simple projection method which is equivalent to PCA and Linear Gaussian Process Regression. We first affinely register images to the MNI space and perform sample-wise z-normalization across all brain mask voxels. Then we executed a voxel-wise z-transformation with the mean and standard deviation estimated on the training samples. The resulting images are used to build a “healthy” vector space over the brain mask voxels and a testing sample is linearly projected on that space. The voxel-wise or sample-wise norm of the residual vector (test vector minus projection vector), after transforming it back from MNI space, is used as the score for pathology [47].

V. Results

This section provides an overview of the 8 valid submissions for the sample-level and the pixel-level tasks. We first present the final challenge results, and then question whether the toy examples alone already provide for a representative proxy ranking. Next, we investigate the performance of the algorithms across different anomaly sizes and color contrasts and across different anomaly types and judge the current state of the submitted OoD algorithms in a clinical application setting.

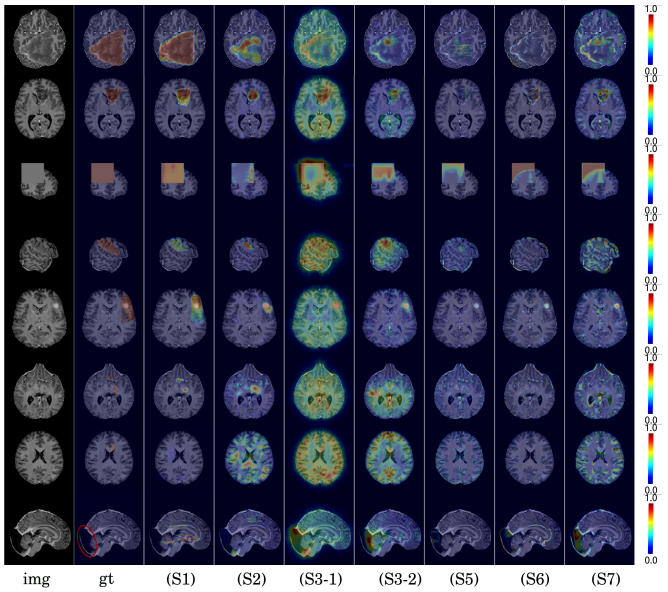

Fig. 2: Pixel-level result heatmap visualizations for the different valid submissions for exemplary and representative brain samples (some of these were solely created for this illustration). Each row corresponds to one example. The first column shows a raw image slice, the second column the ground-truth annotation and the next columns delineate predictions by different submissions (sorted by their pixel-level challenge ranking).

TABLE I: The ranking of sample-level task with the performance on each dataset given as AP.

Rank	Team	Brain	Abdom.
1.	FPI	0.962	0.874
2.	Sergio Naval Marimont, et al.	0.873	0.874
3.	Canon Medical Research Europe	0.845	0.871
4.	NUDT	0.792	0.876
5.	Nina Tuluptceva	0.840	0.861
6.	A1	0.831	0.780
7.	Victor Saase	0.800	0.770
7.	A2	0.634	0.816

A. Challenge ranking

1) Sample-level results: The sample-level results for each dataset and the corresponding consensus ranking obtained for the two target structures can be seen in Table I. While relatively large differences in performance can be observed for the brain, the best ranked teams perform comparably well for the abdomen.

2) Pixel-level results: The pixel-level results for each dataset and the following consensus ranking can be seen in Table II.

B. Toy samples as predictive validation set

We further investigated the predictability of toy examples in performance of the final task. We aimed to explore whether very simple toy examples alone already enable a fair and representative comparison of the approaches, without the need for a big, extensive test set with high anomaly variability. Therefore, we generated 100 abnormal examples using the same mechanism as the toy examples provided to the participants, i.e., adding either spheres or cubes with random intensity to the scans (e.g. see Fig. 2, 3rd row). We call these samples the toy-ish samples. These toy-ish samples vary greatly from most of our challenge test set anomalies.

Interestingly, in all cases (for each dataset & task), the challenge test set ranking and the toy-ish test set rankings had 2 of the top 3 approaches in common. Furthermore, the winning algorithm was also the same algorithm for both the total test set and the toy-ish test set. To quantify the results, we also calculated Kendall tau rank distance, as seen in Table III. Kendall's tau is a correlation coefficient that compares correlations between rankings. We used the tau-b version of Kendall's tau which can handle ties and results in a value of 1.0 for a completely positive correlation, 0.0 for no correlation, and -1.0 for a completely negative correlation.

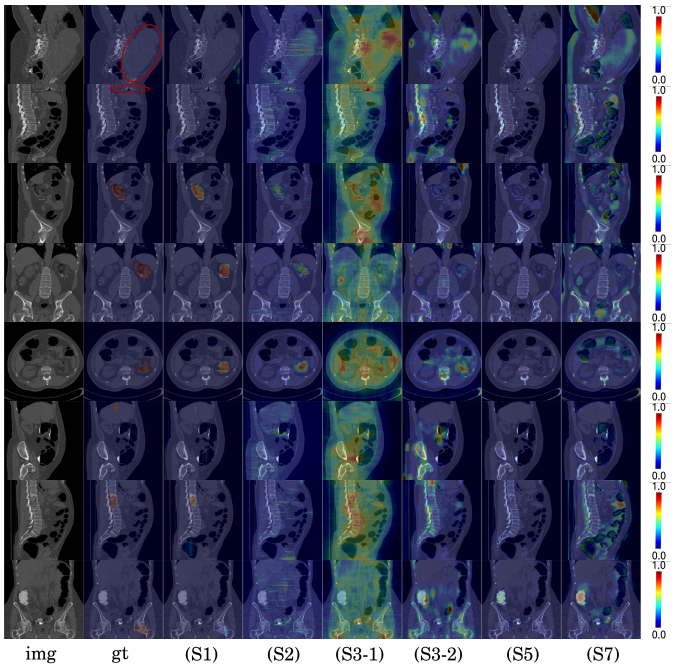

Fig. 3: Pixel-level result heatmap visualizations for the different valid submissions for exemplary and representative abdominal samples (some of these were solely created for this illustration). Each row corresponds to one example. The first column shows a raw image slice, the second column the ground-truth annotation and the next columns delineate predictions by different submissions (sorted by their pixel-level challenge ranking).

TABLE II: The ranking of pixel-level task with the performance on each dataset given as AP.

Rank	Team	Brain	Abdom.	Abbrev.
1.	FPI	0.449	0.394	*(S1)*
2.	Canon Medical Research Europe	0.416	0.288	*(S2)*
3.	Nina Tuluptceva	0.211	0.221	*(S3-1)*
3.	Sergio Naval Marimont, et al.	0.273	0.217	*(S3-2)*
5.	NUDT	0.201	0.239	*(S5)*
6.	Victor Saase	0.204	0.014	*(S6)*
7.	A1	0.160	0.072	*(S7)*
8.	A2	0.002	0.014	*(S8)*

Here, Kendall's tau indicates that there is some correlation between the challenge ranking and the toy-ish dataset ranking, given the limited data. There is a stronger correlation for the abdominal dataset in particular. We assume that this is because the toy cases proved more difficult to analyze overall in the abdominal dataset (see Fig. 8 for example). For both datasets, the toy-ish samples elicited a higher level of predictive accuracy for the pixel-level task. This raises the question whether a quite simple validation dataset can be used to develop generic anomaly detection algorithms. This is also strengthened by the performance of Team Sergio Naval Marimont et al. (2nd and 3rd place), which, in contrast to the other top ranking teams that developed their own sophisticated evaluation datasets, only used the three provided toy cases to evaluate their own submission performance.

TABLE III: Kendall tau rank distance between the rankings on the toy-ish test set and the challenge test set.

	Sample-level	**Pixel-level**
Brain	0.357	0.500
Abdominal	0.642	1.000

C. Analysis

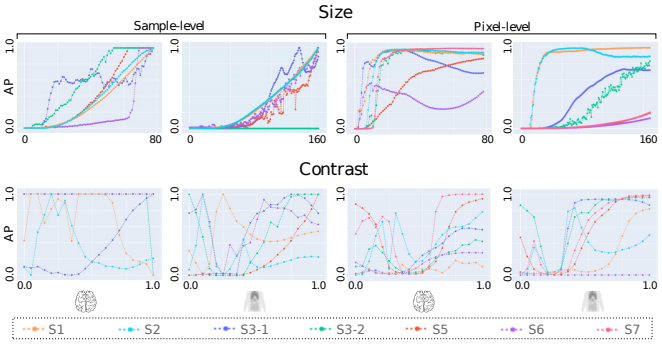

Fig. 4: AP for anomalies of different sizes and levels of contrast. Each line corresponds to a submitted algorithm (S1-S7). The top line of graphs shows the performance for a single toy-ish example which is always in the same position but varies in size from a radius of 0-80 pixels and 0-160 pixels for the brain and abdominal datasets respectively. In the bottom row, the performance for a toy-ish example which is always at the same position with a varying color value, and as such contrast from 0.0 to 1.0 in 0.05 steps, is shown.

1) Contrast & Size: Our hypothesis was that the size and contrast of anomalies would affect the anomaly detection performance. To test this hypothesis, we varied the color-contrast as well as the size for a toy-ish example and outline the results in Fig. 4. While performing this analysis on a more natural or sophisticated anomaly might have given slightly different results, this would require a very comprehensive time- and computing intensive analysis in order to prevent bias. Instead, we chose a simple but nonetheless informative analysis based on toy examples. As expected, the bigger the anomaly size, the better most submissions were able to detect the anomaly. Similarly, the more the contrast differs from the mean (0.5), the better the submissions performed. The top ranking submissions were particularly successful and show the expected bathtub curve. Interestingly, most algorithms tended to perform better on very bright (pixel value ≈ 1) anomalies compared to very dark (pixel value ≈ 0) anomalies, which is likely due to the background which was also assigned the value 0, but was also noted in [48].

2) Anomaly classes: Some anomaly categories proved more challenging than other anomaly classes. Exemplary (pixel-level) anomalies and submission outputs are shown in Fig. 2 and Fig. 3. To procure a quantitative comparison, we chose a dedicated test set with an exact 50%-50% normal-abnormal data sample split with a fixed and consistent number of samples for each subcategory in order to make the metrics as comparable as possible.

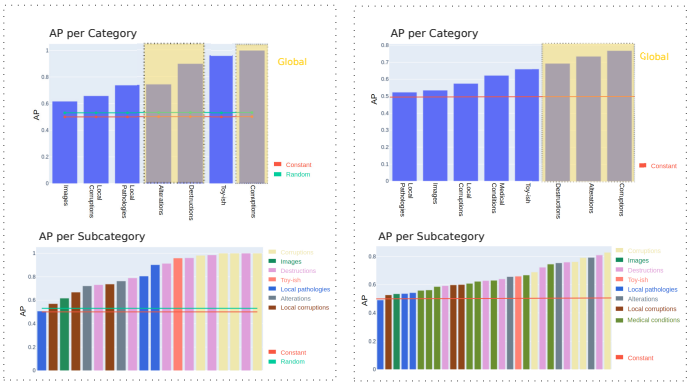

Fig. 5: Median sample-level AP for the different anomaly categories across all submissions. The top row shows the mean of the grouped categories, and the second row gives more detailed results for the subcategories, i.e. the top row categories split up in fine-grained subcategories. The median submission performance was used as a base for the subcategories.

a) Sample-level: The median sample-level performances for the reported categories in Sec. 2.3 are shown in Fig. 5. A clear difference in the performances for the local and global anomalies can be seen. The median performance on the global categories is better than that for the local categories in all cases. Furthermore, the median performances across all algorithms are better than the constant (always predicting the label '0', i.e. no anomaly) and a random (randomly predicting the label '0' or '1') algorithm. As an additional analysis, we also investigated whether the submitted algorithms perform in correlation with the difficulty of the anomalies, according to how they are judged subjectively. Therefore, we sorted the subclasses of some anomaly classes by human-perceived difficulty and show the median performance in Fig. 6. Additionally, in Fig. 5, we show the median performance on all subcategories, sorted by median performance. Again, the performance on global anomalies is almost always better than on local anomalies. The performance in almost all classes is better than a constant guess and is generally similar for the brain and the abdominal dataset. These results raise the question: can these approaches be translated and bring value to a clinical setting now?

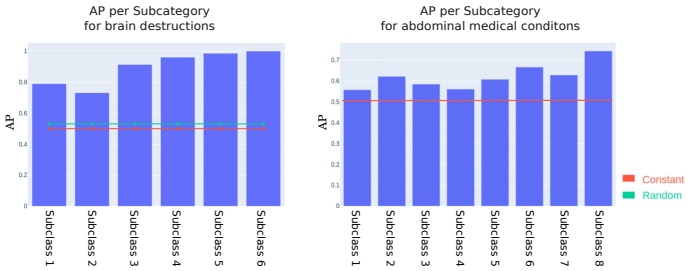

Fig. 6: Median submission performance on sub-classes of two different anomaly categories. The sub classes (classes 1-6, 1-8) are sorted by human perceived difficulty in descending order, i.e. class 1 is the class which was perceived as the hardest and classes 6 (and 8) are the classes perceived as being easiest.

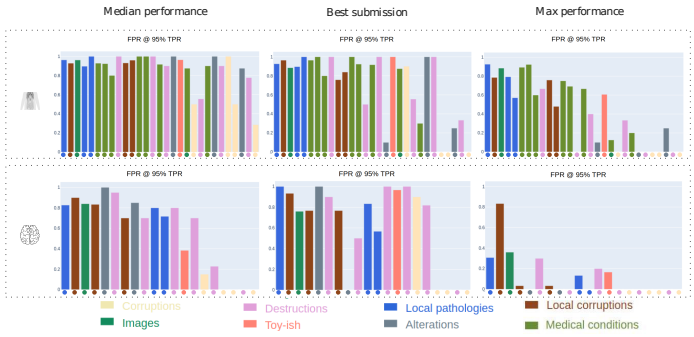

Fig. 7: False positive rate at 95% true positive rate for the different anomaly sub classes of the abdominal dataset (top) and the brain dataset (bottom). The median submission performance, the performance of the best sample-level submission and the maximal performance of algorithms (i.e. picking the best algorithm for each subclass) are shown.

b) FPR@0.95TPR: To investigate a further important aspect of the clinical applicability of the proposed approaches, we analyzed the 'FPR@0.95TPR' metric, which shows a false-positive rate at 95% true-positive rate. In our setting, a score of 0 would mean that an algorithm could detect 95% of the anomalies without diagnosing a single normal sample as abnormal, thus allowing physicians to accelerate their diagnostic processes greatly. A score of 0.5 would mean that prefiltering with an approach would still result in every second image being normal, thus giving a rough acceleration of just $\frac{1}{4}$. A score of 1.0 would mean that in order to detect 95% of the abnormal samples, the physician would have to inspect every sample, providing no acceleration. Whether a TPR of 95% is actually clinically relevant or a higher TPR would be required remains a topic for discussion, however this metric was often used in other OoD work [14], [27] and discussions with physicians have indicated this to be a metric of interest. We present the sample-level results with the FPR@0.95TPR metric in Fig. 7, using the same order as that in Fig. 5. The median performance is shown, along with the individual top subcategory performance, which is determined by choosing the best submission for each subcategory, and the overall best performing algorithm as realistic performance estimates of anomaly detection algorithms respectively. The results here mirror the results in the section above, i.e. for some classes with global destruction or corruptions, the performance is very good, in the best case, the model is able to find 95% of the anomalies without inspecting a single normal image. However for most cases, and especially the the local and medical cases, the amount of cases that would have to undergo inspection in order to find 95% of anomalies could not even be reduced by half.

We do not extend this FPR@0.95TPR analysis to a pixel- or object-level as this requires binarization and connect-component analysis, which might introduce some bias and has not yet been used or evaluated in this context in prior work. Thus, for the pixel-level task, we opt for conventional metrics only.

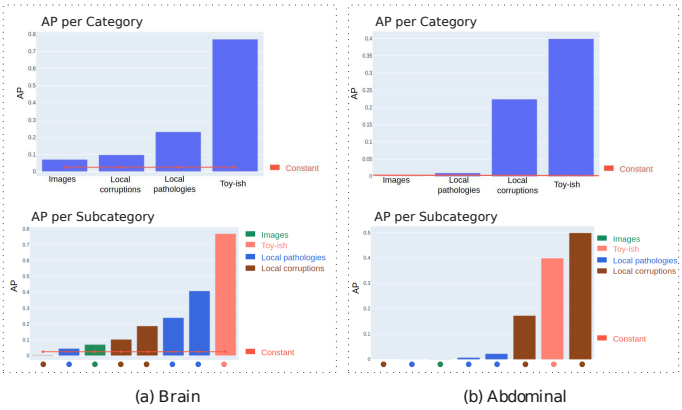

Fig. 8: Pixel-level AP for the different anomaly categories. The top row shows the mean of the grouped categories, and the second row gives more detailed results of subcategories, i.e. the top row categories split up in fine-grained subcategories. Median submission performance was used as the basis for the subcategories.

c) Pixel-level: The median submission performance for pixel-level anomaly categories can be seen in Fig. 8 (qualitative examples are shown in Fig. 2 and Fig. 3). In the pixel-level case especially, the submissions perform better than a constant algorithm in almost all cases. However, the performance differences between the abdominal and brain dataset are vast. While the top (median) performance on the toy-ish dataset is around 0.8 AP on the brain dataset, it is around 0.4 AP on the abdominal dataset. A performance analysis of the categories with subcategories is detailed in the second row of Fig. 8. Interestingly, while some benefits can be observed on most subcategories for the brain dataset, only a few selected types (mostly corruptions) seem to show great improvement compared to a constant guess on the abdominal dataset.

VI. Discussion

The objective of the challenge was to compare different approaches for OoD detection, to find how matured anomaly detection algorithms are and to measure their capabilities in a controlled yet realistic setting. We were also interested in assessing potential applicability and reliability in a clinical

setting. We found some OoD cases/categories which could readily be detected with very high reliability by the best submitted solutions. However, the clinical relevance of these easy-to-detect cases is debatable: these cases mostly contain very prominent global anomalies which mimic failures during the imaging process and corruption of the image files. These kinds of anomalies could be detected by a trained physician without much time and effort. Harder to detect were the local synthetic anomalies for which we can control properties of the anomaly such as intensity, contrast, and texture and can get a more detailed analysis when most approaches might fail. This performance analysis might not directly translate to a clinical setting but believe that this has clear benefits to a setting with only a certain kind of anomaly (e.g. brain tumors). Especially since in practice the type and properties of an anomaly (by definition) should not be known beforehand (which they were not before the challenge), this might give more indication of general performance than a dataset with few common types of pathologies/anomalies. There are submitted algorithms whose performance on harder semantic anomaly cases [11] and on cases with local anomalies (especially for the brain dataset) show very promising performance on some subcategories. Still, there is often great inter-subcategory variability in the anomaly categories and in the qualitative samples shown in Fig. 2 & Fig. 3. Evidently, inter-case and inter-participant variability is still quite high. We believe this high variance makes it hard to recommend a specific algorithm for general OoD detection in practice, and still leaves room for further improvements in OoD techniques in the future.

An interesting point in the results is the difference in performance between the abdominal and brain dataset. First, we included a quite homogeneous brain dataset as most papers published on medical anomaly detection focus on brain datasets. However, to test the robustness and generalizability we opted to include a more heterogeneous, not symmetrical, and closer to clinical practice dataset, the abdominal dataset. While the basic algorithms for creating the synthetic anomaly were kept consistent (with the size ranges adapted to the relevant data-samples' size accordingly), there are some differences between the datasets which might explain the performance gaps. The first point is that most participants developed their algorithm on the brain dataset and then extended it to the abdominal dataset. Similarly, brain datasets have established themselves as the main medical dataset for anomaly detection algorithms [19], [20], [22], [23], probably due to the number of available scans, high data quality, low inter-patient variance and homogeneity of the data samples. Here, the brain dataset contains young healthy patients whose scans were all recorded using the same scanner, whereas the abdominal dataset consists of scans from 18 different sites with elderly people who had a large number of varying anatomical (and pathological) conditions, both natural and unnatural. Furthermore, the brain samples were registered and contain less anatomical variance than abdominal scans by default. Additionally, a data sample of the abdominal dataset is 4 times bigger than a sample from the brain dataset and contains multiple organs and structures. This increased data sample size and complexity might necessitates a larger training sample size to achieve similar absolute performance, which we are unable to provide. We believe that these differences in the dataset characteristics and the fact that primary focus is placed on the brain dataset are the major factors behind the performance differences.

The deviating performance for global and local anomalies was also apparent in the sample-level results. For this case, [11] described a similar notion of semantic vs non-semantic OoD. Semantic OoD describes scenarios in which the OoD samples are contextually similar and roughly originate from the same domain, but contain semantic differences, while non-semantic samples stem from a different domain. The concept in [49], where they differ between near, i.e. from the same domain, and far, i.e from a different domain, OoD samples is similar to this. In our case, no abnormal sample stems from an entirely different domain. However, we would classify most global anomalies as near but nonetheless non-semantic outliers, while most local anomaly samples, e.g. which only have a small gorilla rendered into the image, still contain most of the contextual and statistical properties of the original scans and only exhibit semantic differences, and as such would be classified as near and semantic outliers. They would thus constitute the most interesting cases described in [11], [49]. This division is also broadly reflected in the performance of the algorithms. The subjectively harder the problem is and the more localized the anomalies are, the worse the performance of the algorithms will be. While the submitted algorithms can already almost entirely sort out abnormal inputs for certain categories of global anomalies, the benefit is more unclear for the more interesting and potentially medically relevant cases. In some settings these models can potentially enable the specialist supervision required to be reduced, however in other cases no relevant medical benefit would be expected as of today. Another interesting point raised by our analysis and in [48] is the correlation between intensity and localization performance. In [48], [50] the authors claim that often anomaly detection methods default to simple threshold-based intensity detectors. We do not believe this is in general the case here for all methods, especially with methods such as the winning FPI method, and also is often an inherent property of the datasets used during method development, but this further necessitates a "controlled" setting as in this challenge and [50].

One joint property of all submissions was that instead of processing a whole 3D sample at once, they processed 2D slices instead and then aggregated the anomaly scores to a sample-level score. While 3D processing has in many cases shown some benefits for segmentation [51], the additional compute and time constraints may have been the limiting factors in this case. This, however, can result in slice processing artifacts (see Fig. 3), and the additional information on the complete context and global position might show some potential for further medical OoD detection algorithm research.

Across all submissions, one trend, that is reflected in core machine learning and computer vision research, is the rise of self-supervised methods [43], [52]. Similarly, three of the top submissions here employed self-supervised techniques, either as pretraining to initialize the models or as a proxy

task during training of the algorithm. The extent to which these self-supervised tasks are beneficial is not entirely clear: perhaps, performance gains might also stem from the dedicated (synthetic) validation sets used by all teams or the (coincidental) similarity of the self-supervised tasks to our synthetic anomalies (but, these approaches still show top performance on naturally occurring anomalies). However, the follow-up papers on these approaches showed that the performance translates to other medical datasets as well [40], [53], [54]. Here, in contrast to the two purely self-supervised proxy task methods, the two other top performing methods use Autoencoder-based methods, which are another main direction in anomaly detection [19], [21], [23] and follow-up and consecutive work has also extended the methods to other datasets with great success [46], [55].

One finding which might be of interest for the further development of anomaly detection algorithms is that the simple toy dataset was a capable proxy for more generalized anomaly detection. We do not believe that an algorithm which is tuned on the toy test set and does well on this set will automatically generalize and perform well in other, more general anomaly settings. However, all submitted approaches still struggled to detect the toy examples perfectly (especially as the size decreased and the color contrast in relation to the context got worse) and as such they can be seen as an upper performance limit on the general anomaly detection performance. In addition, we were able to find some correlations between the final performance on the test set and the toy task. We believe that creating and using such a simple validation set might offer an easy way to benchmark anomaly detection algorithms during development, as most of the top teams did this during their development phase.

VII. Conclusion

We have presented the Medical-Out-of-Distribution-Analysis-Challenge 2020. The goal of the challenge was to create a standardized and comprehensive benchmark for OoD detection and anomaly detection algorithms in a controlled and fair medical setting. With eight valid and novel submitted algorithms, the challenge also provided the scope for an analysis of the strengths and weaknesses of current OoD approaches. While the results were quite promising for the global and 'easy' tasks, especially on a pixel level and on low-variance data, we still see room for improvement in clinical real-world scenarios.

References

[1] R. Smith-Bindman, D. L. Miglioretti, and E. B. Larson, "Rising use of diagnostic medical imaging in a large integrated health system," vol. 27, no. 6, pp. 1491–1502.

[2] R. Smith-Bindman, M. L. Kwan, E. C. Marlow, M. K. Theis, W. Bolch, D. L. Miglioretti, and et al, "Trends in use of medical imaging in US health care systems and in ontario, canada, 2000-2016," vol. 322, no. 9, pp. 843–856.

[3] B. van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," vol. 261, no. 3, pp. 719–732.

[4] Y. Gao, K. J. Geras, A. A. Lewin, and L. Moy, "New frontiers: An update on computer-aided diagnosis for breast imaging in the age of artificial intelligence," vol. 212, no. 2, pp. 300–307. Publisher: American Roentgen Ray Society.

[5] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," *arXiv:1610.02136 [cs]*, October 2016. arXiv: 1610.02136.

[6] A. Mehrtash, W. M. Wells III, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation," *arXiv:1911.13273 [cs, eess]*, November 2019. arXiv: 1911.13273 version: 1.

[7] R. Roady, T. L. Hayes, R. Kemker, A. Gonzales, and C. Kanan, "Are Out-of-Distribution Detection Methods Effective on Large-Scale Datasets?," *arXiv:1910.14034 [cs]*, October 2019.

[8] A. Shafaei, M. Schmidt, and J. J. Little, "A Less Biased Evaluation of Out-of-distribution Sample Detectors," *arXiv:1809.04729 [cs, stat]*, August 2019. arXiv: 1809.04729.

[9] T. Drew, M. L. H. Vo, and J. M. Wolfe, ""The invisible gorilla strikes again: Sustained inattentional blindness in expert observers"," *Psychological science*, vol. 24, pp. 1848–1853, September 2013.

[10] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent Space Autoregression for Novelty Detection," *arXiv:1807.01653 [cs]*, July 2018. arXiv: 1807.01653.

[11] F. Ahmed and A. Courville, "Detecting semantic anomalies," *arXiv:1908.04388 [cs]*, August 2019. arXiv: 1908.04388.

[12] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection," *arXiv:1901.08954 [cs]*, January 2019. arXiv: 1901.08954.

[13] L. Beggel, M. Pfeiffer, and B. Bischl, "Robust Anomaly Detection in Images using Adversarial Autoencoders," *arXiv:1901.06355 [cs, stat]*, January 2019. arXiv: 1901.06355.

[14] H. Choi, E. Jang, and A. A. Alemi, "WAIC, but Why? Generative Ensembles for Robust Anomaly Detection," *arXiv:1810.01392 [cs, stat]*, October 2018. arXiv: 1810.01392.

[15] S. Guggilam, S. M. A. Zaidi, V. Chandola, and A. Patra, "Bayesian Anomaly Detection Using Extreme Value Theory," *arXiv:1905.12150 [cs, stat]*, May 2019. arXiv: 1905.12150.

[16] L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther, "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling," *arXiv:1902.02102 [cs, stat]*, February 2019. arXiv: 1902.02102.

[17] C. Piciarelli, P. Mishra, and G. L. Foresti, "Image anomaly detection with capsule networks and imbalanced datasets," *arXiv:1909.02755 [cs]*, September 2019. arXiv: 1909.02755.

[18] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," *arXiv:1802.09088 [cs]*, February 2018. arXiv: 1802.09088.

[19] X. Chen, N. Pawlowski, M. Rajchl, B. Glocker, and E. Konukoglu, "Deep Generative Models in the Real-World: An Open Challenge from Medical Imaging," *CoRR*, vol. abs/1806.05452, 2018.

[20] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images," *arXiv:1804.04488 [cs]*, April 2018. arXiv: 1804.04488.

[21] C. Baur, B. Wiestler, M. Muehlau, C. Zimmer, N. Navab, and S. Albarqouni, "Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain mri," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e190169, 2021.

[22] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," in *IPMI* (M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, and D. Shen, eds.), Lecture Notes in Computer Science, pp. 146–157, Springer, 2017.

[23] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised Anomaly Localization using Variational Auto-Encoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 289–297, Springer, 2019.

[24] A. Avati, M. Seneviratne, E. Xue, Z. Xu, B. Lakshminarayanan, and A. M. Dai, "Beds-bench: Behavior of ehr-models under distributional shift–a benchmark," *arXiv preprint arXiv:2107.08189*, 2021.

[25] D. Ulmer, L. Meijerink, and G. Cinà, "Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data," in *Machine Learning for Health*, pp. 341–354, PMLR, 2020.

[26] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," pp. 9592–9600, 2019.

[27] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, "A Benchmark for Anomaly Segmentation," *arXiv:1911.11132 [cs]*, November 2019. arXiv: 1911.11132 version: 1.

[28] M. Goldstein, *Anomaly Detection in Large Datasets*. June 2014.

[29] V. Škvára, T. Pevný, and V. Šmídl, "Are generative deep models for novelty detection truly better?," *arXiv:1807.05027 [cs, stat]*, July 2018. arXiv: 1807.05027.

[30] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, WU-Minn HCP Consortiumand, and et al, "The Human Connectome Project: a data acquisition perspective," *NeuroImage*, vol. 62, pp. 2222–2231, October 2012.

[31] C. D. Johnson, M.-H. Chen, A. Y. Toledano, J. P. Heiken, A. Dachman, P. J. Limburg, and et al, "Accuracy of CT Colonography for Detection of Large Adenomas and Cancers," *New England Journal of Medicine*, vol. 359, no. 12, pp. 1207–1217, 2008. _eprint: https://doi.org/10.1056/NEJMoa0800996.

[32] D. Zimmerer, J. Petersen, G. Köhler, P. Jäger, P. Full, K. Maier-Hein, and et al, "Medical Out-of-Distribution Analysis Challenge," March 2020. Publisher: Zenodo.

[33] Synapse, "https://www.synapse.org/#!synapse:syn21343101/wiki/599515," 2020. Accessed: 2021-04-30 by D. Zimmerer.

[34] Website, "http://medicalood.dkfz.de/web/2020/," 2020. Accessed: 2021-04-30 by D. Zimmerer.

[35] Sklearn, "https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html - scikit-learn 0.24.1 documentation," 2020. Accessed: 2021-04-30 by D. Zimmerer.

[36] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings," *arXiv:1911.02357 [cs]*, November 2019.

[37] M. Wiesenfarth, A. Reinke, B. A. Landman, M. Eisenmann, A. Kopp-Schneider, and et al, "Methods and open-source toolkit for analyzing and visualizing challenge results," *Scientific Reports*, vol. 11, p. 2369, January 2021. Number: 1 Publisher: Nature Publishing Group.

[38] Github, "https://github.com/mic-dkfz/mood - MIC-DKFZ/mood," January 2021. Accessed: 2021-04-30 by D. Zimmerer.

[39] J. Tan, B. Hou, J. Batten, H. Qiu, and B. Kainz, "Detecting outliers with foreign patch interpolation," *arXiv:2011.04197 [cs]*, 2021.

[40] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders," *IEEE Access*, vol. 9, pp. 118571–118583, 2021.

[41] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," 2015.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv:2002.05709 [cs]*, 2020. arXiv: 2002.05709.

[44] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *arXiv:1711.00937 [cs]*, 2017. arXiv: 1711.00937.

[45] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "PixelSNAIL: An improved autoregressive generative model," 2017.

[46] S. N. Marimont and G. Tarroni, "Anomaly detection through latent space restoration using vector-quantized variational autoencoders," *arXiv:2012.06765 [cs]*, 2020. version: 1.

[47] V. Saase, H. Wenz, T. Ganslandt, C. Groden, and M. E. Maros, "Simple statistical methods for unsupervised brain anomaly detection on MRI are competitive to deep learning methods," *arXiv:2011.12735 [cs]*, 2020.

[48] F. Meissen, G. Kaissis, and D. Rueckert, "Challenging current semi-supervised anomaly segmentation methods for brain mri," 2021. arXiv: 2109.06023.

[49] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, O. Ronneberger, and et al, "Contrastive training for improved out-of-distribution detection," *arXiv:2007.05566 [cs]*, 2020. arXiv: 2007.05566.

[50] F. Meissen, B. Wiestler, G. Kaissis, and D. Rueckert, "On the pitfalls of using the residual as anomaly score," in *Medical Imaging with Deep Learning*, 2022.

[51] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, K. H. Maier-Hein, and et al, "nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation," *arXiv:1809.10486 [cs]*, September 2018.

[52] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9659–9669, 2021.

[53] J. Tan, B. Hou, T. Day, J. Simpson, D. Rueckert, and B. Kainz, "Detecting outliers with poisson image interpolation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds.), (Cham), pp. 581–591, Springer International Publishing, 2021.

[54] A. Kascenas, N. Pugeault, and A. Q. O'Neil, "Denoising autoencoders for unsupervised anomaly detection in brain MRI," in *Medical Imaging with Deep Learning*, 2022.

[55] W. H. L. Pinaya, P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso, "Unsupervised brain anomaly detection and segmentation with transformers," in *Medical Imaging with Deep Learning*, 2021.