Evaluating Reliability in Medical DNNs: A Critical Analysis of Feature and Confidence-Based OOD Detection

Harry Anthony[0009−0004−1252−7448],1,(✉),
Konstantinos Kamnitsas[0000−0003−3281−6509],1,2,3

[1]Department of Engineering Science, University of Oxford, Oxford, UK
[2]Department of Computing, Imperial College London, London, UK
[3]School of Computer Science, University of Birmingham, Birmingham, UK
harry.anthony@eng.ox.ac.uk

Abstract. Reliable use of deep neural networks (DNNs) for medical image analysis requires methods to identify inputs that differ significantly from the training data, called out-of-distribution (OOD), to prevent erroneous predictions. OOD detection methods can be categorised as either *confidence-based* (using the model's output layer for OOD detection) or *feature-based* (not using the output layer). We created two new OOD benchmarks by dividing the D7P (dermatology) and BreastMNIST (ultrasound) datasets into subsets which either contain or don't contain an artefact (rulers or annotations respectively). Models were trained with artefact-free images, and images with the artefacts were used as OOD test sets. For each OOD image, we created a counterfactual by manually removing the artefact via image processing, to assess the artefact's impact on the model's predictions. We show that OOD artefacts can boost a model's softmax confidence in its predictions, due to correlations in training data among other factors. This contradicts the common assumption that OOD artefacts should lead to more uncertain outputs, an assumption on which most confidence-based methods rely. We use this to explain why feature-based methods (e.g. Mahalanobis score) typically have greater OOD detection performance than confidence-based methods (e.g. MCP). However, we also show that feature-based methods typically perform worse at distinguishing between inputs that lead to correct and incorrect predictions (for both OOD and ID data). Following from these insights, we argue that a combination of feature-based and confidence-based methods should be used within DNN pipelines to mitigate their respective weaknesses. These project's code and OOD benchmarks are available at: `https://github.com/HarryAnthony/Evaluating_OOD_detection`.

Keywords: Out-of-distribution · Uncertainty · Distribution shift.

1 Introduction

Deep Neural Nets (DNNs) have emerged as powerful tools for analysing medical images, and have been found promising for various tasks such as classifying dis-

eases [32]. However, when they encounter data that differ significantly from the training data, out-of-distribution (OOD), their generalisation is unpredictable [9]. This motivated research in OOD detection, to create methods to identify when a model prediction is unreliable - mitigating risk of downstream errors.

We can separate OOD detection methods into two categories: *internal methods* (methods that use the parameters or outputs of a DNN which has been trained for a specific task e.g. classification) and *external methods* (external to the DNN). *External methods* encompass many approaches, such as one-class classifier methods [27,25], density-based methods [14,22] and reconstruction-based methods [30,18]. *Internal methods* can be further separated into those that do not require a DNN to be retrained (*post-hoc methods*) and those that require a specific type of training (*ad-hoc methods*). *Ad-hoc methods* cover a broad spectrum of techniques, from altering the network's architecture (e.g. Bayesian Neural Networks [31,2] and confidence enhancement methods [3,33]) to changing how a network is trained (e.g. outlier exposure [10,19]). This paper focuses on *post-hoc internal methods*, which have several benefits: they can be applied to pre-trained networks, they typically don't have restrictions on architecture design and are typically low computational cost. We can further separate *post-hoc internal methods* into *confidence-based methods*, which use a model's output layer for OOD detection, and *feature-based methods*, which use other features for OOD detection (e.g. hidden layer data) [23].

The primary goal of integrating OOD detection methods into a DNN pipeline for image classification is to identify more *trustworthy* predictions, which are likely to be accurate and less susceptible to unpredictable diagnoses caused by the model's interactions with OOD features. Most OOD detection studies evaluate methods on their ability to separate ID and OOD inputs, using a metric like AUROC [24]. However, there's a growing body of research that have argued the traditional OOD framework does not effectively reveal which method is best at detecting errors [6,36,4,12]. They propose evaluating methods based on their ability to specifically discard incorrect predictions, regardless of whether these predictions are from OOD inputs (known as failure detection). To this end, we analyse the strengths and weaknesses of both feature-based and confidence-based methods at OOD detection and failure detection. By identifying the respective weaknesses of these methods, we consider how they can be used to make DNN predictions more trustworthy. We make the following contributions:

- Develop two OOD benchmarks by categorising all images from the D7P and BreastMNIST datasets into those with and without artefacts (rulers and annotations respectively). We manually create modified versions of each of the 478 images with artefacts by removing the artefacts using a patch from the same image, allowing for an analysis of their impact on the model's predictions. As a contribution, we made this data publicly available.
- We challenge assumptions on which confidence and feature-based detection methods rely: OOD artefacts should lead to uncertain (high entropy) model outputs, and the distance of an input to the training data in the model's latent space is a reliable predictor of diagnosis accuracy.

- We use these false assumptions to explain and demonstrate that *feature-based methods* typically perform better at OOD detection than *confidence-based methods*, whereas *confidence-based methods* typically perform better at failure detection than *feature-based methods*.
- To our knowledge, be the first paper to motivate and demonstrate the benefit of combining both confidence and feature-based OOD detection methods in a DNN pipeline to mitigate their respective weaknesses.

2 Material and Methods

Primer on OOD detection. Consider a model's input, $\mathbf{x} \in \mathcal{X}$ and a label $y \in \mathcal{Y} = \{1, ..., K\}$ from a label-space with K classes. Data used to train a network, f, is from a sample $\mathcal{D}_{\text{train}} = \{(x_n, y_n)\}_{n=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$. This paper studies covariate-shifted OOD data $p_{\text{train}}(\mathbf{x}) \neq p_{\text{test}}(\mathbf{x})$, which has the same label-space as the training data, but a distribution shift in $\mathbf{x}$ (i.e. unseen artefact) - sometimes referred to as near OOD [35]. OOD detection can be viewed as a binary classification problem, where we get a confidence scoring function from our method $\mathcal{S}(\mathbf{x}, f)$ for an input $\mathbf{x}$. We label $\mathbf{x}$ as OOD when the scoring function $\mathcal{S}(\mathbf{x}, f)$ is below a threshold λ, and ID if it is above. The primary metric for evaluating the performance of OOD detection is $\text{AUROC}_{\text{OOD}}$, which assesses the scoring function's ability to distinguish ID from OOD inputs [9]. As outlined in Sec. 1, our focus extends to evaluating the scoring function's ability for failure detection, which we quantify using AUROC_f [12]. For $\text{AUROC}_{\text{OOD}}$ the true label is ID and the false label is OOD, and for AUROC_f the true label is a correct diagnosis and false label is an incorrect diagnosis [12].

Confidence-based OOD detection methods use the model's output layer for OOD detection. An example is Maximum Class Probability (MCP) [9], which uses the maximum class softmax probability as the scoring function

$$\mathcal{S}_{\text{MCP}}(\mathbf{x}, f) = \max_{y \in \mathcal{Y}} \text{softmax}[f(\mathbf{x}|\mathcal{D}_{\text{train}})]. \tag{1}$$

Other examples of confidence-based methods include Shannon Entropy (SE) [8], Max Logit Score (MLS) [8], Energy score [19], MCP from Monte Carlo Dropout (MCDP-MCP) [5], predicted entropy from Dropout (MCDP-PE) [12], Mutual Information from Dropout (MCDP-MI) [12], MCP from Deep Ensembles (DE-MCP) [15] and GradNorm [11]. Some methods increase the separation between ID and OOD data by either increasing the model's confidence in a diagnosis (ODIN [17]) or reducing it (ReAct [28] and DICE [29]) - these methods have hyperparameters that can be optimised on a validation OOD set.

In contrast, *feature-based* OOD detection methods don't use the model's output layer. An example is Mahalanobis score, which uses a feature extractor $\mathcal{F}$ (typically a section of the DNN) to extract feature maps from a hidden layer $h(\mathbf{x}) \in \mathbb{R}^{J \times J \times M}$, where the maps have size $J \times J$ with M channels. The feature map's means can be used to define a vector $\mathbf{z}(\mathbf{x}) \in \mathbb{R}^M = \frac{1}{J^2} \sum_J \sum_J \mathbf{h}(\mathbf{x})$. Firstly, the mean $\boldsymbol{\mu}_{\mathbf{y}}$ and covariance matrix Σ_y of each class in the training data $(\mathbf{x}, y) \sim \mathcal{D}_{\text{train}}$ is calculated.

The Mahalanobis distance between the vector $\mathbf{z}(\mathbf{x})$ of a test data point $\mathbf{x}$ and the training data of class y can be calculated as a sum over M dimensions. The Mahalanobis score $d_{\mathcal{M}}$ is defined as the minimum Mahalanobis distance between the test data point and the class centroids of the training data, and it can be used as an OOD scoring function [16].

$$d_{\mathcal{M}_y}(\mathbf{x}) = \sum_{i=1}^{M} (\mathbf{z}(\mathbf{x}) - \boldsymbol{\mu}_{\mathbf{y}}) \Sigma_y^{-1} (\mathbf{z}(\mathbf{x}) - \boldsymbol{\mu}_{\mathbf{y}}), \ \mathcal{S}_{\text{Mahal.}}(\mathbf{x}, f) = -\min_{y \in \mathcal{Y}} d_{\mathcal{M}_y}(\mathbf{x}) \quad (2)$$

Previous works suggest the OOD detection performance of Mahalanobis score could be improved by combining distances from different layers, called Multi-branch Mahalanobis (MBM) [1], or by measuring the distance relative to the distribution of all the training data, called Relative Mahalanobis Score (RMS) [21]. Another feature-based method uses the differences in GRAM matrices [26].
Explainable AI (XAI) methods generate saliency maps to weight the relevance of pixels of an image for a model's diagnosis, aiming to make the model's decision-making process more understandable - an example method is Layer Relevance Propagation (LRP) [20].

Datasets and Implementation. We manually annotated two datasets - BreastMNIST (ultrasound images) [34] and D7P (dermatology images) [13] - into subsets of images that contain an artefact (rulers and annotations respectively) and images that do not. Models were trained on 90% of the images without the artefact, with 10% used as held-out ID test cases (table 1). These tasks were selected for OOD analysis because the artefacts do not provide clinically useful information for diagnosing pathology, and are easy to localise and remove. The models used were ResNet18 and VGG16, where we trained 5 seeds.

Table 1: Summary of ID and OOD data used for OOD detection evaluation.

Dataset	Classes	# ID img	Train:Test	OOD Artefact	# OOD img
Breast-	Normal	126	90:10	Annotations	7
MNIST	Benign	269			168
[34]	Malignant	157			53
D7P	Nevus	832	90:10	Grid ruler	148
[13]	Not Nevus	571			102

For each of the 478 OOD images, we made pixel-wise segmentation masks for the artefact. We then manually replaced the pixels of the artefact with pixels in the same image (intra-image interpolation), using a Gaussian smoothing filter to ensure smooth boundaries. This was chosen over using a pre-trained generative model to remove the artefact because we can ensure we are not introducing a new unexpected OOD artefact into an image, and to prevent changing the true

label of the image. This was done to approximate the image without the artefact, allowing us to evaluate the impact of the artefact on a model's diagnosis. The dataset annotations, segmentation masks and synthetic image datasets have all been made publicly available.

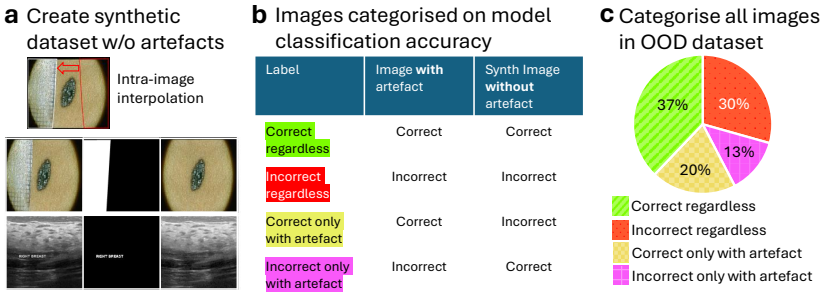

Fig. 1: Workflow for categorising data based on the model prediction impact, using intra-image interpolation to create synthetic images without artefacts.

Once the synthetic datasets were created, we compared the model's classification accuracy with the artefact (original image) and without the artefact. We categorised the OOD data depending if the diagnosis changed with the removal of the artefact (Fig. 1). After categorising the data, the next step involved integrating OOD detection methods into the DNN pipeline. The main focus was to evaluate which of these methods could effectively identify and dismiss potentially misleading predictions influenced by the OOD artefacts. To do this, we applied an OOD detection method and calculated the scoring function for each image in the ID and OOD test sets. We then set a threshold at the 75 percentile of the scoring function's for the held-out ID data $\lambda_{\text{ID-75}}$ [9], removing all model predictions below this threshold. We did this for a confidence-based method (MCP) and feature-based method (Mahalanobis score). The purpose of this study was to determine which OOD methods made the predictions of the DNN more trustworthy, defined here as being more accurate (diagnoses are more often correct) and less likely to have its predictions influenced by OOD features (it can reliably dismiss or handle artefacts that it wasn't trained with).

3 Results

We tested 16 OOD detection methods (described in Sec. 2) for the D7P and BreastMNIST OOD tasks, with results shown in Table 2 - evaluating their performance for OOD detection ($\text{AUROC}_{\text{OOD}}$) and failure detection (AUROC_f). For MC dropout we used $p = 0.3$ with 100 samples. Some methods like ODIN have hyperparameters, for which we show the optimised result as an upper bound. Previous works have shown that OOD artefacts, such as rulers, are optimally detectable in the early layers of a network, which are responsible for detecting

low-level features [1]. So we applied Mahalanobis score and RMS on an early layer of the network (module 5), and we apply MBM on the first branch [1].

Table 2: $\text{AUROC}_{\text{OOD}}$ and AUROC_f (mean of 5 seeds) for OOD detection methods for a) D7P (ruler OOD) and b) BreastMNIST (annotation OOD) tasks. **Bold** highlights best result. * methods with hyperparameters optimised on OOD data.

OOD-D method	D7P (ruler OOD)				BreastMNIST (anno. OOD)			
	ResNet18		VGG16		ResNet18		VGG16	
	AUC_{OOD}	AUC_f	AUC_{OOD}	AUC_f	AUC_{OOD}	AUC_f	AUC_{OOD}	AUC_f
Confidence-based Methods								
MCP [9]	49.3	64.0	51.9	**62.0**	55.8	66.3	52.4	62.4
SE [8]	49.5	64.0	52.8	**62.0**	55.8	66.7	51.4	61.9
MLS [8]	48.6	63.7	51.5	61.9	57.9	69.7	52.4	64.2
Energy Score [19]	48.5	63.6	51.5	61.9	57.6	69.8	51.9	64.1
MCDP-MCP [5]	49.3	64.0	52.0	61.8	55.8	66.3	51.9	62.2
MCDP-PE [12]	49.5	64.0	51.7	61.9	55.8	66.7	50.3	62.2
MCDP-MI[12]	49.5	64.0	51.7	61.8	55.8	66.7	50.3	62.1
DE-MCP [15]	49.9	64.2	52.7	61.9	56.0	66.4	53.3	62.3
GradNorm [11]	49.4	63.9	51.9	61.9	60.2	53.8	53.2	54.1
ODIN* [17]	64.6	58.6	52.0	**62.0**	58.7	67.4	53.6	62.2
ReAct* [28]	67.2	60.6	61.5	58.9	60.2	65.2	58.0	**64.4**
DICE* [29]	68.5	**67.8**	57.7	59.2	58.0	**70.9**	59.1	64.0
Feature-based Methods								
Mahal. Score [16]	76.9	62.1	72.5	57.8	77.1	52.7	72.5	52.2
MBM [1]	**80.7**	61.7	**73.8**	56.8	**77.4**	53.9	**76.8**	52.0
RMS [21]	70.2	57.1	60.5	57.1	62.7	50.5	52.7	51.9
GRAM [26]	53.6	54.8	72.3	55.8	63.6	51.4	71.3	52.0

From Table 2, it is observed that feature-based methods are typically more effective at detecting OOD inputs than confidence-based methods, quantified using $\text{AUROC}_{\text{OOD}}$. To explain why, we visualise the model predictions for two OOD images (both with and without the artefact) along with their eXplainable AI heatmaps using LRP (Fig 2). We use this to challenge the assumption that OOD artefacts will always cause a model to output a more uncertain (high entropy) output. The analysis shows that OOD artefacts can actually lead to high confidence predictions (high logit and hence softmax values), at comparable confidence to ID data. This phenomenon undermines the utility of confidence-based methods, that rely on the model's output layer, for detecting OOD inputs. There are several potential reasons for this phenomenon. One reason is the model can learn to identify correlations in the training data, such as specific intensity patterns in medical images. An OOD artefact that resembles these patterns can lead the model to make high-confidence predictions, even though the artefact is unrelated to the condition being diagnosed. Another cause is it has been theoretically demonstrated that ReLU networks inherently assign high confidence

to images that are far from the training data [7]. The results also show that feature-based methods perform comparatively worse at failure detection. This implies that it may be a false assumption to consider the Mahalanobis distance of an input to the training data as a reliable predictor of diagnostic accuracy [16] - we experimentally observe this phenomenon regardless of the network layer that the method is applied on (results not shown due to space constraints).

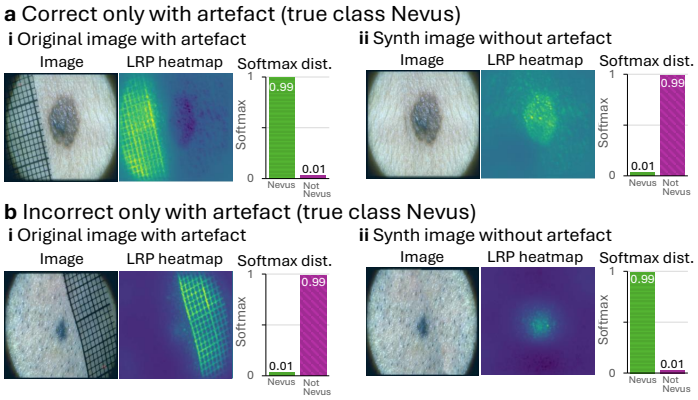

Fig. 2: Comparison of the model's (VGG16) output and XAI heatmap (LRP) for D7P images with and without artefacts, showing cases where predictions are correct (a) or incorrect (b) only with the artefact. Used to demonstrate that OOD artefacts can lead to high confidence predictions.

Following that, we integrated a confidence-based method (MCP) and a feature-based method (Mahalanobis score) into a DNN pipeline, placing threshold λ_{ID-75} and studying the predictions above the threshold (Fig. 3). We see that although the confidence-based method improves accuracy of predictions, it does not notably reduce the proportion of OOD images relative to ID images compared with the original dataset. For predictions on OOD images above λ_{ID-75}, the percentage which are correct only due to the presence of artefacts increases compared with the original dataset, which could cause an inflated AUROC_f metric. This could give a misleading impression that applying MCP would lead to much more trustworthy predictions, when in fact these predictions are heavily impacted by the OOD artefact. This could be an issue if this results in an overconfidence in the model's ability to handle OOD data, masking its vulnerability and potentially leading to performance breakdowns post-deployment if the correlation between OOD artefacts and correct diagnoses changes. This also raises concerns if the failure detection framework for evaluating OOD methods is too simplistic, as these metrics don't consider cases where the model is correct for the wrong reasons. We also see the feature-based method does better at reducing the number of OOD images, but the predictions above λ_{ID-75} can have worse diagnosis accuracy compared with the original dataset (for both ID and OOD sets).

Neither of these outcomes are ideal in terms of improving the trustworthiness of DNN predictions. If we consider feature-based methods as those capable of identifying images that look visually distinct from the training data (regardless of diagnosis accuracy) and confidence-based methods as those that are better at dismissing images with incorrect diagnoses (but struggle to identify visually distinct images), it motivates integrating both confidence and feature-based methods to compensate for their respective weaknesses. We test this by first removing predictions below λ_{ID-75} for Mahalanobis score, then removing predictions below λ_{ID-75} for MCP, to demonstrate that the remaining predictions are more accurate while reducing the number of predictions influenced by OOD features. Although it results in more predictions being dismissed, we argue that this configuration leads to more reliable DNN predictions. Hence, we suggest the community should consider systems which incorporate both a confidence-based and a feature-based method.

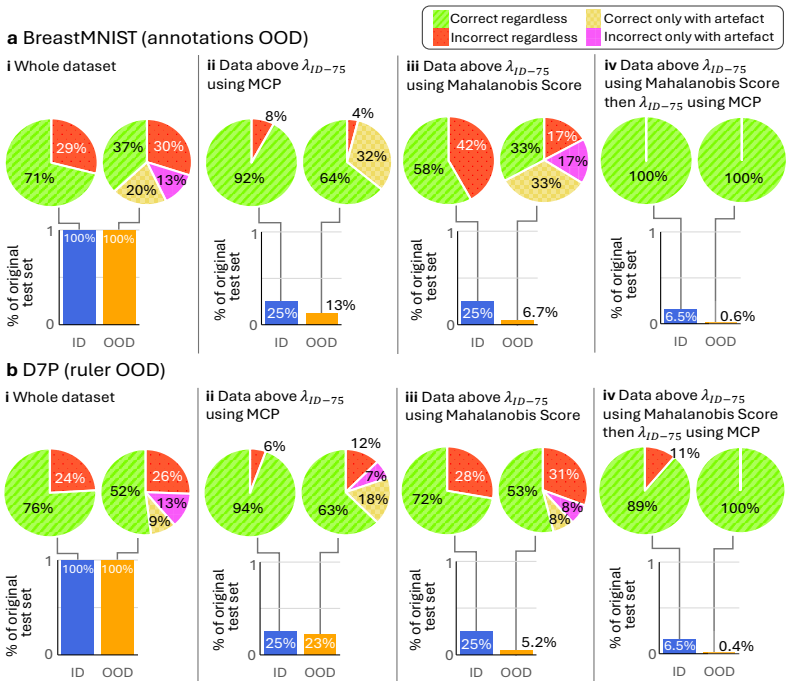

Fig. 3: a) BreastMNIST and b) D7P test sets were analysed using different OOD methods, removing predictions from a VGG16 model below λ_{75-ID}. The pie charts illustrate the distribution of predictions (see Fig. 1), while the bar charts display the percentage of ID and OOD data remaining after removing predictions below λ_{75-ID}, compared to the original dataset (i). The figure shows MCP's limitation in removing OOD data (ii) and Mahalanobis score's tendency to reduce prediction accuracy (iii). Combining these methods (iv) yields the most trustworthy predictions, but with a higher dismissal rate.

4 Conclusion

This paper sheds light into the weaknesses of current OOD detection methods. We show that confidence-based methods are less effective than feature-based methods in OOD detection, partly due to the false assumption that OOD artefacts consistently lead to higher model uncertainty. This paper also explains that feature-based methods, while superior at identifying OOD inputs, under-perform in failure detection compared to confidence-based methods, which can reduce accuracy of predictions when integrated into a DNN pipeline. The paper suggests a step forward could be to seek combinations of confidence- and feature-based methods that compensate for their respective shortcomings.

Acknowledgments. HA is supported by a scholarship via the EPSRC Doctoral Training Partnerships programme [EP/W524311/1, EP/T517811/1]. The authors also acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility for the work (http://dx.doi.org/10.5281/zenodo.22558).

References

1. Anthony, H. and Kamnitsas, K.: On the Use of Mahalanobis Distance for Out-of-distribution Detection with Neural Networks for Medical Imaging. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 136–146. Lecture Notes in Computer Science (2023)
2. Barber, D. and Bishop, C. M.: Ensemble learning in bayesian neural networks. Nato ASI Series F **168**, 215–238 (1998)
3. DeVries, T. and Taylor, G. W.: Learning Confidence for Out-of-Distribution Detection in Neural Networks (2018)
4. Ferreira, R. S. and Guerin, J.: SENA: Similarity-Based Error-Checking of Neural Activations. In: Frontiers in Artificial Intelligence and Applications. IOS Press (2023)
5. Gal, Y. and Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML. pp. 1050–1059 (2016)
6. Guérin, J., Delmas, K., Ferreira, R. and Guiochet, J.: Out-of-distribution detection is not all you need. In: AAAI. vol. 37, pp. 14829–14837 (2023)
7. Hein, M., Andriushchenko, M. and Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: IEEE. pp. 41–50 (2019)
8. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., et al.: Scaling Out-of-Distribution Detection for Real-World Settings (2022)
9. Hendrycks, D. and Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2022)
10. Hendrycks, D., Mazeika, M. and Dietterich, T.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
11. Huang, R., Geng, A. and Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. NeuIPS **34**, 677–689 (2021)
12. Jäger, P. F., Lüth, C., Klein, L. and Bungert, T.: A call to reflect on evaluation practices for failure detection in image classification. In: ICLR (2023)

13. Kawahara, J., Daneshvar, S., Argenziano, G. and Hamarneh, G.: 7-Point Checklist and Skin Lesion Classification using Multi-Task Multi-Modal Neural Nets. IEEE journal of biomedical and health informatics (2018)
14. Kobyzev, I., Prince, S. J. and Brubaker, M. A.: Normalizing Flows: An Introduction and Review of Current Methods. IEEE (11) (2021)
15. Lakshminarayanan, B., Pritzel, A. and Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: NeuIPS. vol. 30. Curran Associates, Inc. (2017)
16. Lee, K., Lee, K., Lee, H. and Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: NeuIPS (2018)
17. Liang, S., Li, Y. and Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2018)
18. Liang, Z., Anthony, H., Wagner, F. and Kamnitsas, K.: Modality cycles with masked conditional diffusion for unsupervised anomaly segmentation in mri. In: MICCAI 2023 Workshops (2023)
19. Liu, W., Wang, X., Owens, J. and Li, Y.: Energy-based out-of-distribution detection. NeuRIPS **33**, 21464–21475 (2021)
20. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., et al.: Layer-Wise Relevance Propagation: An Overview. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (2019)
21. Ren, J., Fort, S., Liu, J., Roy, A. G., et al.: A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection (2021)
22. Ren, J., Liu, P. J., Fertig, E., Snoek, J., et al.: Likelihood ratios for out-of-distribution detection. NeuRIPS **32** (2019)
23. Roschewitz, M. and Glocker, B.: Distance Matters For Improving Performance Estimation Under Covariate Shift. In: IEEE Workshops (ICCVW). Paris, France (2023)
24. Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., et al.: A Unifying Review of Deep and Shallow Anomaly Detection. IEEE **109**(5), 756–795 (2021)
25. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., et al.: Deep One-Class Classification. In: ICML (2018)
26. Sastry, C. S. and Oore, S.: Detecting out-of-distribution examples with gram matrices. In: ICML. pp. 8491–8501 (2020)
27. Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., et al.: Support Vector Method for Novelty Detection. In: NeuIPS. vol. 12 (1999)
28. Sun, Y., Guo, C. and Li, Y.: React: Out-of-distribution detection with rectified activations. NeuIPS **34**, 144–157 (2021)
29. Sun, Y. and Li, Y.: Dice: Leveraging sparsification for out-of-distribution detection. In: ECCV. pp. 691–708 (2022)
30. Tan, J., Hou, B., Day, T., Simpson, J., et al.: Detecting outliers with poisson image interpolation. In: MICCAI 2021. pp. 581–591. Springer (2021)
31. Tishby, , Levin, and Solla, : Consistent inference of probabilities in layered networks: predictions and generalizations. In: International 1989 Joint Conference on Neural Networks (1989)
32. Topol, E. J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine **25**(1), 44–56 (2019)
33. Van Amersfoort, J., Smith, L., Teh, Y. W. and Gal, Y.: Uncertainty estimation using a single deep deterministic neural network. In: ICML. pp. 9690–9700 (2020)

34. Yang, J., Shi, R., Wei, D. and Liu, : MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**(1) (2023)
35. Yang, J., Zhou, K. and Liu, Z.: Full-spectrum out-of-distribution detection. International Journal of Computer Vision **131**(10), 2607–2622 (2023)
36. Zhu, Y., Chen, Y., Li, X., Zhang, R., et al.: Rethinking out-of-distribution detection from a human-centric perspective. International Journal of Computer Vision pp. 1–18 (2024)