

# Confidence-based Out-of-Distribution Detection: A Comparative Study and Analysis

Christoph Berger<sup>1\*</sup>, Magdalini Paschali<sup>1</sup>,  
Ben Glocker<sup>2</sup>, and Konstantinos Kamnitsas<sup>2</sup>

<sup>1</sup> Computer Aided Medical Procedures, Technical University Munich, Germany

<sup>2</sup> Department of Computing, Imperial College London, UK

**Abstract.** Image classification models deployed in the real world may receive inputs outside the intended data distribution. For critical applications such as clinical decision making, it is important that a model can detect such out-of-distribution (OOD) inputs and express its uncertainty. In this work, we assess the capability of various state-of-the-art approaches for confidence-based OOD detection through a comparative study and in-depth analysis. First, we leverage a computer vision benchmark to reproduce and compare multiple OOD detection methods. We then evaluate their capabilities on the challenging task of disease classification using chest X-rays. Our study shows that high performance in a computer vision task does not directly translate to accuracy in a medical imaging task. We analyse factors that affect performance of the methods between the two tasks. Our results provide useful insights for developing the next generation of OOD detection methods.

## 1 Introduction

Supervised image classification has produced highly accurate models, which can be utilized for challenging fields such as medical imaging. For the deployment of such models in critical applications, their raw classification accuracy does not suffice for their thorough evaluation. Specifically, a major flaw of modern classification models is their overconfidence, even for inputs beyond their capacity. For instance, a model trained to diagnose pneumonia in chest X-rays may have only been trained and tested on X-rays of healthy controls and patients with pneumonia. However, in practice the model may be presented with virtually infinite variations of patient pathologies. In such cases, overly confident models may give a false sense of their competence. Ideally, a classifier should know its capabilities and signal to the user if an input lies out of distribution.

In this work, we first explore confidence- and distance-based approaches for out-of-distribution (OOD) detection on a standard computer vision (CV) task and afterwards evaluate the best OOD detection methods on a medical benchmark dataset. Moreover, we provide a set of useful insights for leveraging OOD approaches from computer vision to challenging medical datasets.

---

\* c.berger@tum.de

**Related work:** OOD detection methods can be divided in two categories. The first consists of methods that build a **dedicated model for OOD detection** [25]. Some works accomplish this via estimating density  $p(x)$  of ‘normal’ in-distribution (ID) data and then classify samples with low  $p(x)$  as OOD [10]. However, learning  $p(x)$  accurately can be challenging. An alternative is to learn a decision boundary between ID and OOD samples. Methods [27] attempt this in an unsupervised fashion using only ‘normal’ data. Nonetheless, supervised alternatives have also been introduced for CV and medical imaging [6,24,29], exposing the OOD classifier to OOD data during training. Such OOD data can originate from another database or be synthesized. However, collecting or synthesising samples that capture the heterogeneity of OOD data is challenging. Another approach for creating OOD detection neural networks (NNs) is *reconstruction-based* models [8,21]. A model, such as an auto-encoder, is trained with a reconstruction loss using ID data. Then, it is assumed that the reconstruction of unseen OOD samples will fail, thus enabling their detection. This approach is especially popular in medical imaging research [22,32,1,26,23], likely because it produces a per-pixel OOD score, allowing its use for unsupervised segmentation. It has shown promise for localisation of salient abnormalities but does not reach the performance of supervised models in more challenging tasks.

The second category of OOD detection methods, which this study focuses on, enhances a task-specific model to detect when an input is OOD. These approaches are commonly based on **confidence of model predictions**. They are compact, integrated straight into an existing model, and operate in the task-specific feature or output space. Their biggest theoretical advantage in comparison to training a dedicated OOD detector is that if the main model is unaffected by a change in the data, the OOD detector also remains unaffected. A subset of confidence-based methods has a probabilistic motivation, exploring the use of the predictive uncertainty of a model, such as Maximum Class Probability (MCP) [5], MCDropout [2] or ensembling [12]. Others derive confidence-scores based on distance in feature space [31], or learn spaces that better separate samples via confidence maximization [14] or contrastive losses [31,28]. In medical imaging, related work is mostly focused on improving uncertainty estimates by DNNs [30,17], or analysing quality of uncertainty estimates in *ID* settings [19,9]. In contrast, investigation of OOD detection based on model confidence is limited. A recent study compared MCDropout and ensembling [16] for medical imaging, finding the latter more beneficial. The potential of other OOD detection methods for medical imaging is yet to be assessed adequately, despite their importance for the field.

**Contributions:** This study assesses confidence-based methods for OOD detection. To this end, we re-implement and compare approaches, shown in Figure 1, in a common test-bed to accomplish a fair and cohesive comparison. We first evaluate those approaches on a CV benchmark to gain insights for their performance. Then, we benchmark all approaches on real-world chest X-rays [7]. We find that the performance of certain methods varies drastically between OOD detection tasks, which raises concerns about their reliability for real-world use,

and we identify a method that is consistently high performing across tasks. Finally, we conduct an empirical analysis to identify the factors that influence the performance of these methods, providing useful insights towards building the next generation of OOD detection methods.

## 2 Out-of-Distribution Detection Methods

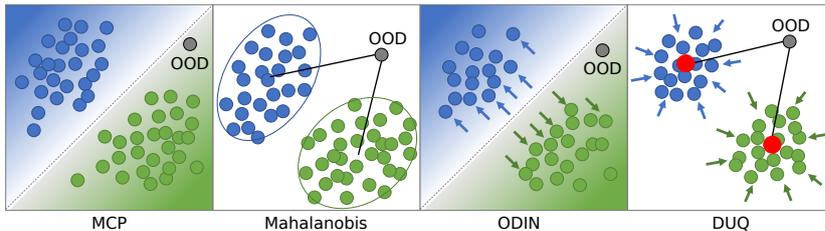


Fig. 1: Overview of the OOD detection methods studied: Maximum Class Probability (baseline), Mahalanobis Distance, ODIN and DUQ.

We study the following methods for OOD detection in image classification: **Maximum Class Probability (MCP)** [5]: Any softmax-based model produces an estimate of confidence in its predictions via its class posteriors. Specifically, the probability  $\max_y p(y|x)$  of the most likely class is interpreted as an ID score and, conversely, low probability indicates possible OOD input. Even though modern NNs have been shown to often produce over-confident softmax outputs [3], this method is a useful baseline for OOD detection.

**Mahalanobis Distance** [13]: Lee et al. propose the Mahalanobis distance as OOD metric in combination with NNs. The method can be integrated to any pre-trained classifier. It assumes that the class-conditional distributions of activations  $z(x; \theta) \in \mathbb{R}^Z$  in the last hidden layer of the pre-trained model follow multivariate Gaussian distributions. After training model parameters  $\theta$ , the model is applied to all training data to compute for each class  $c$ , the mean  $\hat{\mu}_c \in \mathbb{R}^Z$  of activations  $z$  over all training samples  $x$  of class  $c$ , and the covariance matrix  $\hat{\Sigma}$  of the class-conditional distributions of  $z$ . To perform OOD detection, the method computes the Mahalanobis distance between a test sample  $x$  and the closest class-conditional distribution as follows:

$$M(x) = \max_c - (z(x; \theta) - \hat{\mu}_c)^T \hat{\Sigma}^{-1} (z(x; \theta) - \hat{\mu}_c) \quad (1)$$

The threshold to decide whether an input is OOD or ID is then set as a certain distance from the closest distribution.

**Out-of-Distribution Detector for Neural Networks (ODIN)** [14]: This method is also applicable to pre-trained classifiers which output class-posteriors using a softmax. Assume  $f(x; \theta) \in \mathbb{R}^C$  are the logits for  $C$  classes. We write

$S(x, \tau) = \text{softmax}(f(x; \theta), \tau) \in \mathbb{R}^C$  for the softmax output calculated for temperature  $\tau$  ( $\tau_{tr} = 1$  for training), and  $S(x, \tau)_c$  is the value for class  $c$ . The method is based on the assumption that we can find perturbations of the input that increase the model’s confidence more drastically for ID samples than for OOD samples. The perturbed version of input  $x$  is given by:

$$\tilde{x} = x - \varepsilon \text{sign}(-\nabla_x \log \max_c S(x; \theta, \tau_{tr})_c) \quad (2)$$

Here, a gradient is computed that maximizes the softmax probability of the most likely class. The model is then applied on the perturbed sample  $\tilde{x}$  and outputs softmax probabilities  $S(\tilde{x}; \tau') \in \mathbb{R}^C$ . From this, the MCP ID score is derived as  $\max_c S(x; \tau')_c$ . Since the perturbation forces over-confident predictions, it negatively affects calibration. To counteract this, ODIN proposes using a different softmax temperature  $\tau'$  when predicting the perturbed samples, to recalibrate its predictions.  $\tau'$  is a hyperparameter that requires tuning. We assess the effect of the perturbation and  $\tau$  in an ablation study.

**Deep Ensembles [12]:** This method trains multiple models from scratch, while initialisation and order of training data is varied. During inference, predicted posteriors of all models are averaged to compute the ensemble’s posteriors. This in turn is used to compute MCP of the ensemble as an ID score. While deep ensembles have been shown to perform well for OOD detection, they come with high computational cost as training and inference times scale linearly with number of ensemble members. In our experiments, we also investigate an ensemble that uses a consensus Mahalanobis distance as OOD score instead of MCP.

**Monte Carlo Dropout (MCDP) [2]:** MCDP trains a model with dropout. At test time, multiple predictions are made per input with varying dropout masks. The predictions are averaged and MCP is used as ID score. The method interprets these predictions as samples from the model’s posterior, where their average is a better predictive uncertainty estimate, improving OOD detection.

**Deterministic Uncertainty Quantification (DUQ) [31]:** This method trains a feature extractor without a softmax layer. Instead, it learns a centroid per class and attracts samples towards the centroids of their class, similar to contrastive losses [4]. It uses a Radial Basis Function (RBF) kernel to compute the distance between the input’s embedding and the class centroids. The distance to the closest centroid defines classification, and is also used as the OOD score. Because RBF networks are prone to feature collapse, DUQ introduces a gradient penalty to regularize learnt embedding and alleviate the issue. Nonetheless, we still faced difficulties with DUQ convergence despite considerable attempts.

### 3 Benchmarking on CIFAR10 vs SVHN

We first show results on a common computer-vision (CV) benchmark to gain insights about methods’ performance, and validate our implementations by replicating results of original works before applying them to a biomedical benchmark.

Table 1: Out-of-distribution detection performance of WideResNet 28x10 trained on CIFAR10 with SVHN as OOD set. We report averages over 3 seeds.

Method	AUROC	AUCPR	ID Acc.
MCP (baseline)	0.939	0.919	0.952
MCDP	0.945	0.919	<b>0.956</b>
Deep Ensemble	0.960	0.951	0.954
Mahalanobis	0.984	0.960	0.952
Mahalanobis Ens.	<b>0.987</b>	<b>0.967</b>	0.954
ODIN	0.964	0.939	0.952
ODIN (pert. only)	0.968	0.948	0.952
ODIN (temp. only)	0.951	0.920	0.952
DUQ	0.833	-	0.890

### 3.1 Experimental Setup

**Dataset:** We use the training and test splits of CIFAR10 [11] as ID and SVHN [20] as OOD test set ( $n_{\text{test ID}} = 10000$ ,  $n_{\text{test OOD}} = 26032$ ). A random subset of 10% CIFAR training data is used as validation set, to tune method hyperparameters, such as temperature  $\tau$  for ODIN.

**Model:** We use a WideResNet (WRN) [33] with depth 28 and widen factor 10 (WRN 28x10), trained with SGD using momentum 0.9, weight decay 0.0005, batch normalization and dropout of 0.3 for 200 epochs with early stopping.

**Evaluation Metrics:** We use the following metrics to assess the performance of a method in separating ID from OOD inputs: (1) area under the receiver operating characteristic (AUROC), (2) area under the precision-recall curve (AUCPR), (3) accuracy (Acc) on ID test set. We also use (4) Expected Calibration Error (ECE) as a summary statistic for model calibration [18].

### 3.2 Results

In Table 1, we compare OOD detection performance for all studied methods. MCDP marginally improves over the baseline, with higher gains by Deep Ensembles. Interestingly, ODIN achieves comparable AUROC with Deep Ensembles and ODIN’s input perturbation is the component responsible for the performance (see ODIN (pert. only)). The results of only applying temperature scaling and no input perturbation are listed under ODIN (temp. only). The highest AUROC over all methods is achieved by Mahalanobis distance both as a single model and an ensemble. Moreover, none of the OOD detection methods compromised the accuracy on the classification task. We reproduced the results of original implementation of DUQ with ResNet50. However, we faced unstable training of DUQ on our WRN and did not obtain satisfactory performance despite our efforts.

Table 2: Performance of different methods for separation of out-of-distribution (OOD) from in-distribution (ID) samples for CheXpert in two settings. **Setting 1:** Classifier trained to separate *Cardiomegaly* from *Pneumothorax* (ID) is given samples with *Fractures* (OOD). **Setting 2:** Classifier trained to separate *Lung Opacity* from *Pleural Effusion* (ID) is given samples with *Fracture* or *Pneumonia* (OOD). We report average over 3 seeds per experiment. Best in **bold**.

Method	Setting 1			Setting 2		
	OOD		ID	OOD		ID
	AUROC	AUCPR	Acc	AUROC	AUCPR	Acc
MCP (baseline)	0.678	0.695	0.888	0.458	0.586	0.758
MCDP	0.696	0.703	0.880	0.519	0.637	0.756
Deep Ensemble	0.704	0.705	<b>0.895</b>	0.445	0.582	<b>0.769</b>
Mahalanobis	0.580	0.580	0.888	0.526	0.601	0.758
Mahalanobis Ens.	0.596	0.586	<b>0.895</b>	0.537	0.613	0.758
ODIN	<b>0.841</b>	<b>0.819</b>	0.888	0.862	0.856	0.758
ODIN (pert. only)	<b>0.841</b>	<b>0.819</b>	0.888	<b>0.865</b>	<b>0.856</b>	0.757
ODIN (temp. only)	0.678	0.695	0.888	0.444	0.575	0.757

## 4 Benchmarking on the X-ray Lung Pathology Dataset

### 4.1 Experimental Setup

**Dataset:** To simulate a realistic OOD detection task in a clinical setting, we use subsets of the CheXpert X-ray lung pathology dataset [7] as ID and OOD data, in two different settings. Since CheXpert images are multi-labeled, we only used samples where ID and OOD classes were mutually exclusive. **Setting 1:** We train a classifier to distinguish *Cardiomegaly* from *Pneumothorax* (ID), and use images with *Fracture* as OOD ( $n_{\text{test ID}} = 4300$ ,  $n_{\text{test OOD}} = 7200$ ). **Setting 2:** We train a classifier to separate *Lung Opacity* and *Pleural Effusion* (ID), and use *Fracture* and *Pneumonia* as OOD classes ( $n_{\text{test ID}} = 6000$ ,  $n_{\text{test OOD}} = 8100$ ). **Model:** We use WRN with depth 100 and a widen factor 2 (WRN 100x2). All other parameters remain the same as for the CIFAR10 vs SVHN benchmark. **Evaluation:** We analyse performance based on the same metrics as in Sec. 3.

### 4.2 Results

Results for the two ID/OOD settings in CheXpert are shown in Table 2. The baseline performance indicates that the ID and OOD inputs are harder to separate for Setting 2, and much harder than the CIFAR vs SVHN task. MCDP improves OOD detection in both Settings. Interestingly, Deep Ensembles, often considered the most reliable method for OOD detection, do not improve Setting 2, although the Mahalanobis Ensemble does. Moreover, ODIN shows best performance in both settings with a considerable margin, even when only using

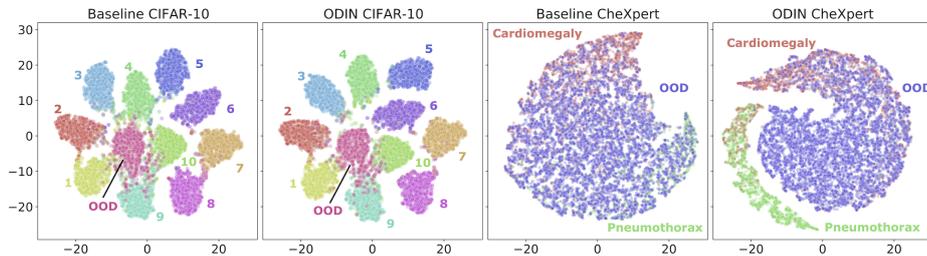


Fig. 2: T-SNE of embeddings for CIFAR10 vs SVHN and for CheXpert Setting 1. The OOD cluster is less separated for the latter, challenging benchmark. ODIN perturbations improve separation, which may explain its performance.

the adversarial-inspired component of the method without softmax tempering (see ODIN (pert. only) in Figure 2). Mahalanobis distance, which was the best method on the CIFAR10 vs SVHN task, performs worse than the Baseline on Setting 1 and only yields modest improvements in Setting 2. Reliability of OOD methods is crucial. Thus, the next section further analyses ODIN and Mahalanobis, to gain insights in the consistent performance of ODIN and the difference between the CV benchmark and CheXpert Setting 1 that may be causing the inconsistency of Mahalanobis distance.

### 4.3 Further Analysis

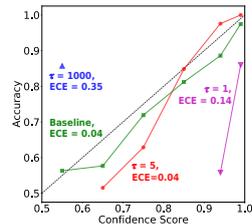
**Mahalanobis:** Our first hypothesis to explain the poor performance of Mahalanobis on the medical OOD detection task in comparison to the CV task was that the Mahalanobis distance may be ineffective in higher dimensional spaces. In the CIFAR10 vs SVHN task, the Mahalanobis distance is calculated in a hidden layer with [640,8,8] (40960 total) activations, whereas the WRN 100x2 for CheXpert has a corresponding layer with shape [128,56,56] (401408 total) activations. To test this hypothesis, we reduce the number of dimensions on which we compute the distributions by applying strided max pooling before computing the Mahalanobis distance and report the results in Table 3a. We find that this dimensionality reduction is not effective and conclude that this is not the major cause of Mahalanobis ineffectiveness in CheXpert.

To further investigate, we visualize with T-SNE [15] the last layer activations when trained models process perturbed samples for the CIFAR10 vs SVHN task and the CheXpert Setting 1. Figure 2 shows that activations for CIFAR10 classes are clearly separated and the OOD set is distinguishable from the ID clusters. For CheXpert, the baseline model achieves less clear separation of the two ID classes and the OOD class overlaps substantially with the ID classes. This suggests that fitting a Gaussian distribution to the ID embeddings is challenging, causing the Mahalanobis distance to not yield significant OOD detection benefits.

**ODIN:** We investigate how the perturbation that ODIN adds to inputs benefits OOD detection. For this, we also show T-SNE plots for both CIFAR10 and CheX-

Pooling	Layer shape	AUROC
None	[128,56,56]	<b>0.6018</b>
4x4, stride=2	[128,56,56]	0.5634
2x2, stride=4	[128,14,14]	0.5608
8x8, stride=1	[128,14,14]	0.5508
1x1, stride=4	[128,14,14]	0.545

(a) Results with dimensionality reduction



(b) Calibration curves

Table 3a: Results on CheXpert Setting 1 from experiments with dimensionality reduction in last hidden layer. Lower dimensionality did not improve OOD detection via Mahalanobis distance. Fig. 3b: Calibration of baseline and ODIN for varying temperature  $\tau$  and associated ECE, for CheXpert Setting 1. The baseline (green) is reasonably calibrated. Adding noise to the inputs with ODIN leads to highly overconfident model (purple, all samples very high confidence). For CheXpert,  $\tau = 1000$  as used for CIFAR10 leads to under-confident model, whereas  $\tau = 5$  restores good calibration. Interestingly, all ODIN settings achieve the same AUROC for OOD irrespective of  $\tau$  value and calibration.

pert Setting 1 in Figure 2. The added perturbation results in a better separation of ID classes in both datasets, with the effect more pronounced for CheXpert. While there is still overlap between the *Fracture* OOD class and the *Pneumothorax* ID class, the clusters are more pronounced which ultimately leads to better OOD detection. Finally, we investigate the effect of temperature variation in ODIN. Following [14], temperature 1000 was used for CIFAR10 and CheXpert. By comparing baseline, ODIN (temp. only) and (pert. only) on Tables 1 and 2, we find that OOD detection is primarily improved by perturbation, not temperature scaling, especially on CheXpert. We note, however, that the perturbations lead to a completely over-confident model using training temperature 1, with all predictions having very high confidence (Figure 3b). AUROC and AUCPR are calculated via ordering the OOD score (i.e. confidence) of predictions, so even slight differences between ID and OOD samples suffice to separate false and true detections. If only those metrics were taken into account, temperature scaling might have been considered redundant. However, to deploy an OOD system, a threshold on the confidence / OOD score needs to be chosen. Spreading the confidence estimates via temperature scaling ( $\tau = 5$  in Figure 3b) enables more reliable choice and deployment of a confidence threshold in practical settings.

## 5 Conclusion

This work presented an analysis of various state-of-the-art methods for confidence-based OOD detection on a computer vision and a medical imaging task. Our comprehensive evaluation showed that the performance of methods in a computer vision task does not directly translate to high performance on a medical

imaging task, emphasized by the analysis of the Mahalanobis method. Therefore, care must be given when a method is chosen. We also identified ODIN as a consistently beneficial OOD detection method for both tasks. Our analysis showed that its effect can be attributed to its input perturbation, which enhances separation of ID and OOD samples. This insight could lead to further advances that exploit this property. Future work should further evaluate OOD detection methods across other datasets and tasks to better understand which factors affect their performance and reliability towards real-world deployment.

## Acknowledgements

This work received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 757173, project MIRA, ERC-2017-STG), and the UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare.

## References

1. Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S.: Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis* p. 101952 (2021)
2. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) *Proceedings of The 33rd International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), <http://proceedings.mlr.press/v48/gal16.html>
3. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 70, pp. 1321–1330. PMLR, International Convention Centre, Sydney, Australia (06–11 Aug 2017), <http://proceedings.mlr.press/v70/guo17a.html>
4. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. vol. 2, pp. 1735–1742. IEEE (2006)
5. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations* (2017)
6. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606* (2018)
7. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 33, pp. 590–597 (2019)
8. Japkowicz, N., Myers, C., Gluck, M., et al.: A novelty detection approach to classification. In: *IJCAI*. vol. 1, pp. 518–523. Citeseer (1995)
9. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience* (2020)

10. Kobyzev, I., Prince, S., Brubaker, M.: Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020)
11. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-10 (canadian institute for advanced research) <http://www.cs.toronto.edu/~kriz/cifar.html>
12. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems*. vol. 30 (2017)
13. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: *Advances in Neural Information Processing Systems*. vol. 31 (2018)
14. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: *ICLR* (2018)
15. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* 9(86), 2579–2605 (2008)
16. Mehrtaash, A., Wells, W.M., Tempny, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging* 39(12), 3868–3878 (2020)
17. Monteiro, M., Folgoc, L.L., de Castro, D.C., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *arXiv preprint arXiv:2006.06015* (2020)
18. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. p. 2901–2907. AAAI’15, AAAI Press (2015)
19. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis* 59, 101557 (2020)
20. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
21. Pang, G., Shen, C., Cao, L., Hengel, A.v.d.: Deep learning for anomaly detection: A review. *arXiv preprint arXiv:2007.02500* (2020)
22. Pawlowski, N., Lee, M.C., Rajchl, M., McDonagh, S., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., et al.: Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders (2018)
23. Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain anomaly detection and segmentation with transformers. *arXiv preprint arXiv:2102.11650* (2021)
24. Roy, A.G., Ren, J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z., et al.: Does your dermatology classifier know what it doesn’t know? detecting the long-tail of unseen conditions. *arXiv preprint arXiv:2104.03829* (2021)
25. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. *arXiv preprint arXiv:2009.11732* (2020)
26. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: *International conference on information processing in medical imaging*. pp. 146–157. Springer (2017)
27. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. *Neural computation* 13(7), 1443–1471 (2001)

28. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. *NeurIPS* (2020)
29. Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B.: Detecting outliers with foreign patch interpolation. *arXiv preprint arXiv:2011.04197* (2020)
30. Tanno, R., Worrall, D.E., Ghosh, A., Kaden, E., Sotiropoulos, S.N., Criminisi, A., Alexander, D.C.: Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 611–619. Springer (2017)
31. Van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.: Uncertainty estimation using a single deep deterministic neural network. In: III, H.D., Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 119, pp. 9690–9700. PMLR (13–18 Jul 2020)
32. You, S., Tezcan, K.C., Chen, X., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. In: *International Conference on Medical Imaging with Deep Learning*. pp. 540–556. PMLR (2019)
33. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: *BMVC* (2016)