Information Bottleneck-based Feature Weighting for Enhanced Medical Image Out-of-Distribution Detection

Brayden Schott[1](✉), Žan Klaneček[2], Alison Deatsch[1], Victor Santoro-Fernandes[1], Thomas Francken[1], Scott Perlman[3], and Robert Jeraj[1,2]

[1] Department of Medical Physics, School of Medicine and Public Health, University of Wisconsin, Madison WI 53706, USA
bschott@wisc.edu
[2] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
[3]Department of Radiology, Section of Nuclear Medicine, School of Medicine and Public Health, University of Wisconsin, Madison WI 53706, USA

Abstract. Deep learning models are subject to failure when inferring upon out-of-distribution (OOD) data, i.e., data that differs from the models' train data. Within medical image settings, OOD data can be subtle and non-obvious to the human observer. Thus, developing highly sensitive algorithms is critical to automatically detect medical image OOD data. Previous works have demonstrated the utility of using the distance between embedded train and test features as an OOD measure. These methods, however, do not consider variations in feature importance to the prediction task, treating all features equally. In this work, we propose a method to enhance distance-based OOD measures via feature importance weighting, which is determined through an information bottleneck optimization process. We demonstrate the utility of the weighted OOD measure within the metastatic liver tumor segmentation task and compare its performance to its non-weighted counterpart in two assessments. The weighted OOD measure enhanced the detection of artificially perturbed data, where greater benefit was observed for smaller perturbations (e.g., AUC = 0.8 vs. AUC = 0.72). In addition, the weighted OOD measure achieved better correlation to liver tumor segmentation Dice coefficient (e.g., ρ = -0.76 vs ρ = -0.21). In summary, this work demonstrates the benefit of feature importance weighting for distance-based OOD detection.

Keywords: Out-of-Distribution Detection, Uncertainty Quantification, Tumor Segmentation

1 Introduction

Deep learning models are known to fail when inferring on data that differs from the models' train data [1–3]. These differences may be obvious such as data from semantically different classes. Alternatively, differences may be subtle where the semantic meaning of the data is appropriate for a model, yet some fundamental data features differ from those present in the train data. In both cases, the input data would be considered out-of-distribution (OOD). The latter case, referred to as covariate shifted data, is especially relevant to the medical image setting where there are numerous sources of covariate shifts including differences in scanners, differences in image acquisition protocols, the use of small train datasets, and more. These sources likely contribute to the alarming lack of generalizability of medical image deep learning models [4]. To account for these shifts and their induced model failures, it is critical to implement OOD detection algorithms.

OOD detection methods generally follow one of four approaches [5]. Classification-based methods manipulate a model's output to detect OOD data, for example, by means of temperature scaling [6], gradient space calculations [7], or Bayesian modeling [8, 9]. These methods, however, are not typically invoked for covariate shift detection. Density estimator-based methods seek to approximate the probability density of a model's train data and flag test data with low probability likelihood as OOD. Unexpectedly, these methods have been shown to assign high likelihoods to OOD data [10]. Reconstruction-based approaches train a generative model (e.g., an autoencoder) to encode and reconstruct train data, where the reconstruction performance is used as an OOD metric. Reconstruction-based methods, however, tend to lack sensitivity for detecting subtle OOD shifts [11, 12]. On the other hand, distance-based methods generally have greater detection sensitivity than reconstruction-based methods [12] and operate by invoking some distance measure between a set of embedded train and test features (i.e., model activations), where large distances are associated with OOD data.

Several works have implemented a variety of distance-based OOD detection methods on natural [13–15] and medical images [16, 17]. Perhaps most notably, the Mahalanobis distance is often employed because it accounts for possible correlations between features. One limitation of these approaches is that the applied distance measures require dimensionality reduction, which potentially removes valuable information. Distance-based measures are also limited in that they consider embedded features equally, when it is known that some features are more important for the prediction task than others [18].

Distance-based OOD detection methods may be enhanced by weighting the distance measure by each feature's level of importance to the prediction task. One strategy for acquiring "feature importance weights" is to implement an information bottleneck process. This is an optimization process by which the complexity of an information system is reduced by the removal of unnecessary information [19]. If we consider a deep learning model, then the necessity of information within the model, determined by the information bottleneck process, can be used to characterize feature importance. Information bottlenecks have previously been used for deep attribution mapping [20, 21] but have not yet been integrated into an OOD detection algorithm.

In this work, we explored the utility of information bottlenecks to enhance distance-based OOD detection via feature importance weighting. We evaluated the benefit of this weighting mechanism for the metastatic liver tumor segmentation task, where we assessed the performance of detecting covariate shifted data in addition to assessing the correlation between the OOD measure and model segmentation performance.

2 Methods

2.1 Information Bottleneck Implementation

In this work, we adapted the information bottleneck implementation described in K. Schulz et al. [21]. The goal was to use a post hoc optimization routine to minimize the amount of feature information of a deep learning model according to the loss function:

$$\mathcal{L} = \mathcal{L}_{model} + \beta\mathcal{L}_{info} \tag{1}$$

where $\mathcal{L}_{model}$ is the model's standard loss (e.g., cross entropy), $\mathcal{L}_{info}$ describes the amount of model feature information, and β is set according to the desired trade-off between these two terms. Consequently, this loss function minimizes information while encouraging good model performance.

Feature information of selected model layers was minimized through the injection of noise, an information removal process [22]. Let R represent the embedded features of some model layer. Noise was injected into these features according to

$$Z = \lambda(\alpha)R + \big(1 - \lambda(\alpha)\big)\epsilon \tag{2}$$

where $\lambda(\alpha) = \text{sigmoid}(\alpha)$, α is a learnable parameter inserted at the model layer, and ϵ is replacement noise defined as $\epsilon \sim \mathcal{N}(\mu_R, \sigma_R^2)$, where μ_R and σ_R^2 are the estimated mean and variance of the layer features, sampled from the train data. The optimizable α parameter controls how much features are replaced with noise and can be used to characterize feature importance. The shared information between Z and R ($I[R, Z]$) describes the amount of information removed from R and can be approximated as

$$\mathcal{L}_{info} = I[R, Z] \cong \mathbb{E}_R\big[D_{KL}[P(Z \mid R) \mid\mid Q(Z)]\big] \tag{3}$$

where $P(Z \mid R)$ is the probability distribution of Z given R, $Q(Z) = \mathcal{N}(\mu_R, \sigma_R^2)$ is a variational approximation, and $D_{KL}[\cdot]$ represents the KL-divergence. A detailed derivation of equation 3 can be found in K. Schulz et al. [21].

We implemented information bottlenecks on a U-Net segmentation model. Due to the U-Net skip connections, the model does not have an architectural bottleneck where it may be fitting to insert an information bottleneck due to restricted information flow. As a modification, we inserted multiple information bottlenecks in the model, one after each encoder layer (**Fig. 1**). The information loss from each inserted bottleneck was then weighted and summed to define the total information loss as

$$\mathcal{L}_{info,total} = \textstyle\sum_{l=1}^{L} \overline{w}_l \cdot \mathcal{L}_{info,l} \tag{4}$$

where $\mathcal{L}_{info,l}$ is the information loss from the encoder layer, l, $\overline{w}_l$ are normalized layer-wise weights pre-defined using $\overline{w}_l = w_l / \sum_{l=1}^{L} w_l$ and $w = \left\{\frac{1}{2^{L-l}} \mid l = 1, 2, \ldots, L\right\}$, and L is the number of encoder layers. Under this formalism, information losses were weighted more heavily for deeper encoder layers, where distance measures are expected to be more stable.

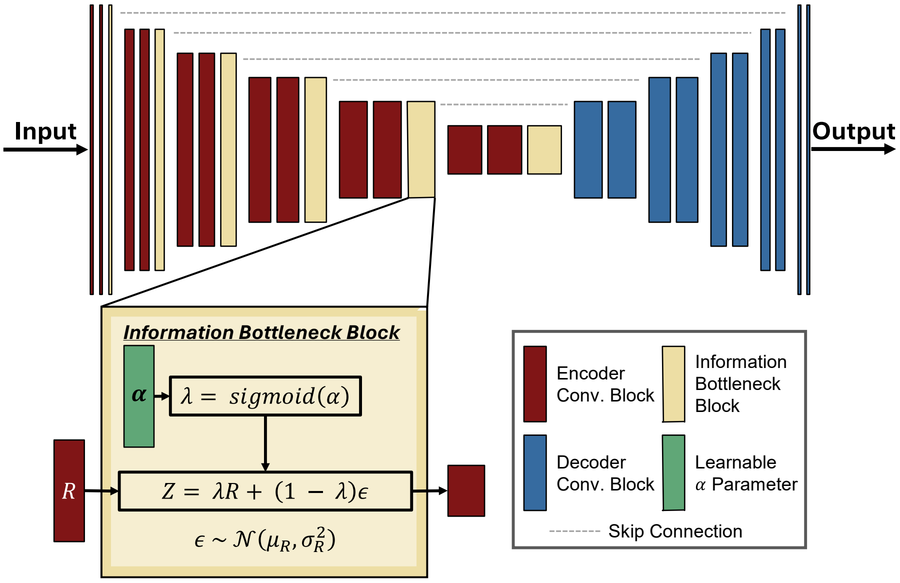

Fig. 1. Schematic describing the implementation of information bottlenecks on a U-Net architecture. An information bottleneck block is placed after each encoder layer. The information bottleneck block diagram was adapted from K. Schulz et al. [21].

This modified information bottleneck method was implemented in a post hoc manner on individual test data samples, where the only learnable parameters in the model were the inserted α parameters. We trained this process for 20 iterations using the Adam optimizer and a learning rate of 0.5, initialized all α parameters to 5.0, and set $\beta = 0.1$. The feature-wise Gaussian distributions in equations 2 and 3 were sampled from 1,000 train data samples. To enable use in a deployed setting, we set the ground-truth in the model loss of equation 1 as the model prediction before initiating the information bottleneck optimization. Thus, the optimization process was encouraged to maintain model performance while minimizing feature information.

2.2 OOD Distance Measures

Weighted OOD Measure. We assumed that individual embedded features in the model's encoder followed a Gaussian distribution [14, 23]. Using the estimated train feature-wise distributions, we calculated the number of standard deviations from the mean (i.e., z-scores) for each test feature and aggregated this into a layer-wise measure via a weighted average:

$$\overline{OOD}_{layer,l} = \frac{\sum_{n=1}^{N} \bar{\alpha}_{l,n} \cdot \frac{|R_n - \mu_n|}{\sigma_n}}{\sum_{n=1}^{N} \bar{\alpha}_{l,n}} \tag{5}$$

where N is the number of features in layer l, and $\bar{\alpha}_{l,n}$ is the learnt feature-wise importance parameter from index n and layer l, normalized between zero and one. This layer-wise measure was aggregated into an encoder-wise OOD measure through a weighted sum:

$$OOD_{weighted} = \sum_{l=1}^{L} \overline{w}_l \cdot \overline{OOD}_{layer,l} \tag{6}$$

where $\overline{w}_l$ are the same weights used in the total information loss (equation 4).

This distance measure was selected for its computational simplicity, and because it does not remove information via feature downsampling. Moreover, correlations among features are expected to be intrinsically removed via the information bottleneck process.

Non-weighted OOD Measure. A non-weighted OOD measure was defined using equations 5 and 6 and setting $\alpha_{l,n} = \{1, 1, \ldots, 1\}$ of length N, yielding the average of the feature deviations from the train distribution within each layer, aggregated into the encoder-wise measure. We refer to this non-weighted OOD measure as $OOD_{non\text{-}weighted}$.

Baseline OOD Measure. We implemented the established Mahalanobis distance as a baseline measure, which has been shown to outperform non-feature-based OOD detection approaches such as temperature scaling [16]. For this, we followed the method described in González et al. [16], where the Mahalanobis distance between train and test features from a segmentation model's bottleneck was computed. Bottleneck features were downsampled by average pooling until the number of features was less than the number of train samples (i.e., 1,000). We refer to this OOD measure as OOD_{maha}.

2.3 Datasets

Datasets utilized in this work consisted of abdominal CT scans from non-uniform acquisition protocols of patients with metastatic liver tumors. The base segmentation model was trained using $N = 104$ scans acquired from the LiTS liver tumor segmentation challenge [24]. This data defined the model's train distribution. An in-house dataset of $N = 31$ CT scans of Neuroendocrine Tumor patients with liver metastases was retrospectively collected and used as test data for all OOD evaluations. All ethical guidelines were followed, and internal review board authorization was approved for this data collection (IRB: 2015-0273, UWCCC: UW19146). Liver tumor contours on the in-house dataset were acquired under clinician guidance.

2.4 Segmentation Model

A three-dimensional segmentation model was trained in-house to segment the liver organ and metastatic liver tumors using the *nnUNet* repository [25]. The nnUNet model has demonstrated highly competitive results on a variety of segmentation tasks, making it well-suited to augment with and test OOD detection algorithms. The model was trained for 1,000 epochs with a batch size of 2 and using the sum of the Dice coefficient and cross entropy as the loss. Instance normalization was used between each convolution layer, the input image patch size was $[128 \times 128 \times 128]$ voxels, and data augmentation was applied during training. Model training took place on an Nvidia RTX

Titan GPU workstation with 24 GB of memory. Prior to training, all images were normalized to zero-mean-unit-variance, resampled to a common 2.5 mm^3 voxel spacing, and cropped about the center of the liver to the model patch size. Two patches were necessary for inference on each test image due to image padding to accommodate data augmentation. Consequently, OOD measures were averaged across these two patches to acquire an image-wise OOD measure for each test image.

2.5 OOD Detection Assessments

Image Perturbation Detection. In this first assessment, we evaluated each OOD measure's ability to distinguish adversarial attacked OOD test data from the model's train data. The purpose of this approach was to expose the model to unique perturbations not observed during training (via augmentations), mimicking encountering new data in deployed settings. The Fast Gradient Sign Attack (FGSM) method [26] was applied to the test data according to:

$$\tilde{x} = x + \epsilon \cdot sign\big(\nabla_x J(\theta, x, y)\big) \tag{7}$$

where $\tilde{x}$ is the perturbed image, x is the original image, y is the segmentation ground-truth label, θ is the model's learnable parameters, ϵ is the perturbation magnitude, and $sign(\cdot)$ describes the sign of the model loss gradient with respect to the input image, x. FGSM was performed on the test data using five perturbation magnitudes, $\epsilon = [0.0, 0.25, 0.50, 0.75, 1.0]$, constructing five OOD test sets (e.g., **Fig. 2**). For each perturbation magnitude, we assessed each OOD measure's performance in detecting perturbed data using receiver operating curve statistics including areas under the curve (AUC) and false positive rates at the 95% true positive threshold (FPR95). Bootstrapping was performed to acquire 95% confidence intervals for each detection metric.

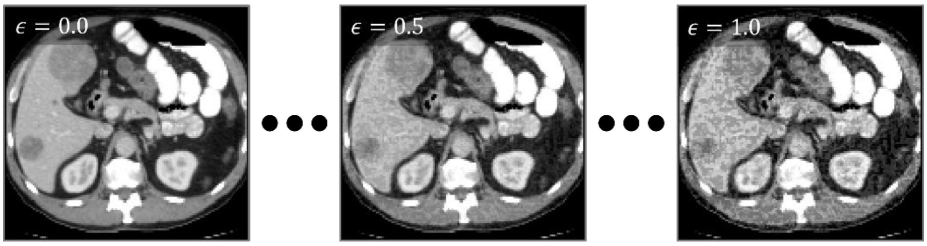

Fig. 2. A single example test scan with perturbations of different magnitudes.

Correlations with Segmentation Performance. In this second assessment, we evaluated each OOD measure's correlation with the trained model's segmentation performance. The Spearman correlation coefficients between individual OOD measures and the liver tumor Dice coefficient were obtained for each (unperturbed) test image. To observe the benefit of the information bottleneck optimization process on the weighted OOD measure, we reported this correlation when using the weights derived from each optimization iteration.

3 Results

3.1 Image Perturbation Detection

The OOD detection results for detecting perturbed data are shown in **Error! Reference s ource not found.**[1]. The OOD distance weighting enhanced detection performance for each perturbation magnitude except for the largest perturbation, where the performance between $OOD_{weighted}$ and $OOD_{non\text{-}weighted}$ was comparable. The detection of data with smaller perturbation magnitudes benefited more from weighting than for larger magnitudes. Meanwhile, the detection from the established OOD_{maha} measure showed little change across perturbation magnitudes and was consistently inferior to the $OOD_{weighted}$ and $OOD_{non\text{-}weighted}$ measures.

Table 1. Perturbed test image OOD detection results across distance measures and perturbation magnitudes (ϵ). Numbers in brackets indicate the 95% confidence intervals derived from bootstrapping. Bold text indicates the best performing metric.

	OOD_{maha}		$OOD_{non-weighted}$		$OOD_{weighted}$	
ϵ	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓
0.00	0.63 [0.52, 0.74]	0.99 [1.00, 0.60]	0.72 [0.62, 0.81]	0.80 [0.89, 0.46]	**0.80** **[0.71, 0.87]**	**0.62** **[0.74, 0.39]**
0.25	0.66 [0.54, 0.77]	0.99 [1.00, 0.56]	0.77 [0.67, 0.85]	0.76 [0.92, 0.39]	**0.83** **[0.75, 0.90]**	**0.59** **[0.68, 0.32]**
0.50	0.65 [0.53, 0.77]	0.99 [1.00, 0.58]	0.89 [0.83, 0.95]	0.34 [0.94, 0.15]	**0.95** **[0.92, 0.98]**	**0.21** **[0.32, 0.07]**
0.75	0.67 [0.55, 0.80]	0.99 [1.00, 0.54]	0.97 [0.95, 0.99]	0.15 [0.29, 0.02]	**0.98** **[0.97, 1.00]**	**0.07** **[0.17, 0.02]**
1.00	0.65 [0.53, 0.77]	0.99 [1.00, 0.76]	**0.99** **[0.99, 1.00]**	**0.03** **[0.10, 0.00]**	**0.99** **[0.98, 1.00]**	0.08 [0.16, 0.00]

3.2 Correlations with Segmentation Performance

The average number of tumors per test image predicted by the segmentation model was 9 (range: 1-30). The correlation analysis between image-wise liver tumor Dice coefficient and OOD distance as a function of information bottleneck optimization iteration is shown in **Fig. 3**.[1] Two test images with neither predicted nor ground-truth liver tumors were omitted from this analysis. Overall, the correlation coefficient magnitude of the weighted OOD distance enhanced from -0.21 (the non-weighted OOD distance) to -0.76 at the last iteration. Meanwhile, the established Mahalanobis distance yielded a correlation coefficient of 0.08.

[1] Additional figures supporting these results are included as supplementary material.

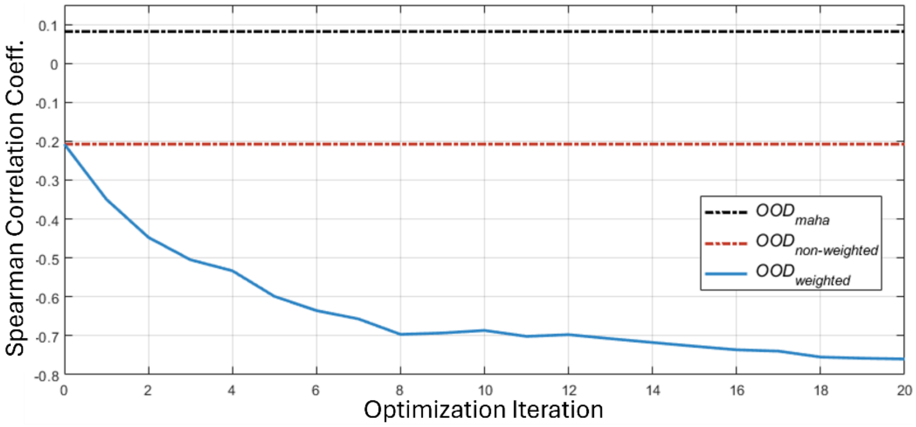

Fig. 3. The spearman correlation coefficient between $OOD_{weighted}$ and liver tumor segmentation Dice coefficient as a function of information bottleneck optimization iteration. OOD_{maha} and $OOD_{non\text{-}weighted}$ do not depend on information bottleneck optimization, and thus these are displayed as constants across optimization iterations.

4 Discussion and Conclusion

Our results indicate that weighting the OOD distance by feature importance enhances OOD detection performance. In the image perturbation detection analysis, we observed a greater benefit for the OOD distance weighting at smaller perturbation magnitudes. This implies that the weighting offers more benefit for detecting near-OOD than far-OOD data, where the former more closely resembles the types of shifts expected in deployed clinical settings. In the correlation analysis, we found that the weighting was essential to obtain a strong correlation ($|\rho| > 0.75$) between OOD distance and model segmentation performance. A strong correlation indicates that an OOD measure may serve as a proxy for model performance in the absence of ground truth data and may facilitate more trustworthy use of deployed clinical models. In both assessments, the established Mahalanobis distance measure was inferior to both $OOD_{weighted}$ and $OOD_{non\text{-}weighted}$.

In contrast to our results, González et al. [16] found that the Mahalanobis distance was superior to other OOD measures. However, their work compared the Mahalanobis distance measure to non-feature-based and general uncertainty quantification measures (e.g., Monte Carlo dropout), which have been shown to be insufficient in detecting OOD data [27–29]. As more OOD measures are established, a comprehensive comparison of feature-based OOD measures should be investigated. Additionally, our segmentation model may have been overcomplete, meaning, it did not need all the encoder layers for prediction. Consequently, the bottleneck features may have been dominated by noise, decreasing the utility of the Mahalanobis distance from the bottleneck layer.

A challenge of our work was the selection of the OOD distance measure. Distance measures are known to breakdown at high dimensions [30]. Our distance measure was implemented at each encoder layer, where the dimensions of shallower layers were high. However, the weighted sum of the distance measures from all encoder layers down-weighted distances with higher dimensions. In addition, the feature importance weighted averaging reduced the concern of using high dimension distances.

In summary, we demonstrated the importance of weighting an OOD distance measure by each feature's level of importance to the prediction task, where importance weights were acquired from a post hoc information bottleneck optimization process. Assessments regarding the relationship between the derived OOD measure and quantitative biomarkers (e.g., tumor volume) will be needed to help further understand the potential clinical impact of this work.

Acknowledgments. This research was supported by the University of Wisconsin Carbone Cancer Center.

Disclosure of Interests. Author Robert Jeraj, PhD is the Chief Scientific Officer and a co-founder of AIQ Solutions, a quantitative medical image analysis software company.

References

1. Nguyen, A., Yosinski, J., Clune, J.: Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In: CVPR. pp. 427–436 (2015)
2. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
3. Soin, A., Merkow, J., Long, J., Cohen, J.P., Saligrama, S., Kaiser, S., Borg, S., Tarapov, I., Lungren, M.P.: CheXstray: Real-time Multi-Modal Data Concordance for Drift Detection in Medical Imaging AI. arXiv. arXiv:2202.02833, (2022)
4. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 17, (2019). https://doi.org/10.1186/s12916-019-1426-2
5. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized Out-of-Distribution Detection: A Survey. arXiv. arXiv:2110.11334, (2021)
6. Liang, S., Li, Y., Srikant, R.: Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In: ICLR (2018)
7. Huang, R., Geng, A., Li, Y.: On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In: NeurIPS (2021)
8. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani. In: ICML (2016)
9. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: NeurIPS (2017)
10. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do Deep Generative Models Know What They Don't Know? In: ICLR (2019)
11. Meissen, F., Wiestler, B., Kaissis, G., Rueckert, D.: On the Pitfalls of Using the Residual Error as Anomaly Score. In: MIDL (2022)
12. Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., Vernekar, S.: Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance. arXiv. arXiv:1812.02765, (2018)

13. Huang, H., Li, Z., Wang, L., Chen, S., Dong, B., Zhou, X.: Feature Space Singularity for Out-of-Distribution Detection. arXiv. arXiv:2011.14654, (2020)

14. Lee, K., Lee, K., Lee, H., Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: NeurIPS (2018)

15. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-Distribution Detection with Deep Nearest Neighbors. In: ICML (2022)

16. González, C., Gotkowski, K., Fuchs, M., Bucher, A., Dadras, A., Fischbach, R., Kaltenborn, I.J., Mukhopadhyay, A.: Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation. Med Image Anal. 82, (2022). https://doi.org/10.1016/j.media.2022.102596

17. Karimi, D., Gholipour, A.: Improving Calibration and Out-of-Distribution Detection in Deep Models for Medical Image Segmentation. IEEE Transactions on Artificial Intelligence. 4, 383–397 (2023). https://doi.org/10.1109/TAI.2022.3159510

18. Samek, W., Wiegand, T., Müller, K.-R.: Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv. arXiv:1708.08296, (2017)

19. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv. arXiv:physics/0004057, (2000)

20. Zhmoginov, A., Fischer, I., Sandler, M.: Information-Bottleneck Approach to Salient Region Discovery. In: ECML PKDD (2020)

21. Schulz, K., Sixt, L., Tombari, F., Landgraf, T.: Restricting the Flow: Information Bottlenecks for Attribution. In: ICLR (2020)

22. Alemi, A.A., Fischer, I., Dillon, J. V., Murphy, K.: Deep Variational Information Bottleneck. In: ICLR (2017)

23. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-Normalizing Neural Networks. In: NeurIPS (2017)

24. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, et al.: The Liver Tumor Segmentation Benchmark (LiTS). Med Image Anal. 84, (2023). https://doi.org/10.1016/j.media.2022.102680

25. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 18, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

26. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. In: ICLR (2015)

27. Schwaiger, A., Sinhamahapatra, P., Gansloser, J., Roscher, K.: Is Uncertainty Quantification in Deep Learning Sufficient for Out-of-Distribution Detection? In: AISafety@IJCAI (2020)

28. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., Snoek, J.: Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In: NeurIPS (2019)

29. Liu, Y., Pagliardini, M., Chavdarova, T., Stich, S.U.: The Peril of Popular Deep Learning Uncertainty Estimation Methods. In: NeurIPS (2021)

30. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: ICDT (2001)