Out-of-Clinical-Distribution Detection with a Softmax-Conditioned Variational Autoencoder Regulariser: Application to Fetal Ultrasound

Kangning Zhang[1]*, Jianbo Jiao[1,2], and J. Alison Noble[1]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
*kangning.zhang@eng.ox.ac.uk
[2] School of Computer Science, University of Birmingham, UK

Abstract. In medical image analysis, a reliable model is required to detect inputs containing important anatomical information and make accurate decisions based on it. Motivated by this, we introduce the concept of "out-of-clinical-distribution" (OCD) detection, where in-clinical-distribution data (ICD) is defined as images containing a "clinically interesting region" that is essential to clinical decision-making. We propose an OCD detection framework based on a classification-based model, enhanced by a novel softmax-conditioned variational autoencoder regulariser. In this framework, softmax scores are incorporated into the latent space with a mixture of learnable class-conditioned Gaussian distributions as prior. By embedding class information in feature reconstruction, this approach enforces feature compactness within ICD classes and enhances the separability between ICD and OCD features. The effectiveness of the proposed OCD detection method is demonstrated in the task of selecting anatomical views from real-time fetal ultrasound (US) videos, where it significantly outperforms both state-of-the-art classification-based and generative-based methods.

Keywords: OOD Detection · Ultrasound · Variational Autoencoder.

1 Introduction

Out-of-distribution (OOD) detection, which ensure reliable deployment of deep learning (DL) models in the open world, is crucial in medical image analysis [10]. A reliable DL model should refrain from making clinical decisions for cases outside its validated expertise, to ensure both patient safety and clinical accuracy [13]. Defining OOD data in the medical domain is challenging, as it must be clinically meaningful while accounting for factors such as data acquisition biases. Common examples of OOD data include inputs that are unrelated to the target evaluation, incorrectly prepared and previously unseen cases [3].

In this paper, we propose a new OOD definition tailored to the clinical setting. Specifically, we introduce the concept of a "*clinically interesting region (CIR)*", which refers to an area within a medical image containing critical and clear anatomical information that is essential for clinical decision-making and

holds particular clinical significance, such as abnormalities. Images containing a CIR are defined as *in clinical-distribution* (ICD) data. In real-world scenarios, such images demand extra attention from clinicians, as they are more likely to be linked to diseases or require further treatment. *Out-of-clinical distribution* (OCD) data refer to sample that lack of CIR. The goal of OCD detection is to identify the CIRs and discriminate ICD samples from OCD.

Here, we focus on OCD detection in fetal US application. During scanning, the sonographer adjusts the probe based on their expertise and standard plane guidelines to capture the optimal view of the fetal anatomy. Once the desired standard plane is achieved, a freeze-frame is saved and annotated for measurements and the medical report. Sonographers may zoom in, zoom out, and fine-tune the probe to achieve the optimal view, often holding the probe near standard planes for extended periods. Therefore, frames captured before and after the freeze-frame may contain relevant anatomical views, even if they don't fully meet standard criteria. These frames are referred to as "approximate standard planes" because they contain valuable, though not perfectly aligned, anatomical information. Fetal US examinations typically involve 13 anatomical views. These views include both standard planes and "approximate standard planes," which are considered ICD data as they contain relevant anatomical information. Background frames, on the other hand, are considered OCD data. A reliable deep learning model that mimics the sonographer's scanning and detection process should not only accurately detect and classify anatomical views but also remain vigilant to background frames.

Similar to OOD detection [14], an OCD detector defines a score function that maps input features to an uncertainty score, distinguishing inputs based on this score. Early approaches primarily focused on classification-based models [9][1][15], while generative models have gained popularity due to their promising results despite higher complexity [6][7]. For high-resolution medical imaging, this increased complexity demands significantly more computational resources, raising concerns. Meanwhile, anatomical views vary widely due to differences in probe positioning and fetal movements, often sharing contextual similarities with background frames captured in close succession during continuous scanning. This high heterogeneity within ICD classes and proximity to near-OCD frames makes detection more challenging [5]. To address these, we develope a classification-based model enhanced with a softmax-conditioned VAE as a feature regulariser. By incorporating softmax scores into the latent space, the VAE constructs a mixture of learnable class-conditioned Gaussians as priors, enhancing the constraint on ICD features and improving detection accuracy.

In summary, this paper makes the following three contributions: 1) We define a new concept of "out-of-clinical-distribution" detection in clinical setting, in which samples contain "clinical interesting region" are defined as ICD data. An OCD detector helps identify samples with clinically relevant information for decision-making; 2) We propose an OCD detection framework that combines a classification backbone with a softmax-conditioned VAE in the feature space. The VAE integrates softmax scores into both inference and generative models,

acting as a feature regulariser to enhance the compactness of ICD clusters; 3) Extensive experiments on fetal ultrasound video frames demonstrate the simplicity and effectiveness of the proposed OCD detection method.

2 OCD Detection Definition

Clinically Interested Region (CIR): We define a CIR as an area within a medical image that contains key anatomical structures with potential clinical significance. Such regions are of particular interest to clinicians due to their relevance to patient diagnosis or treatment and often require closer examination or analysis.

Out-of-clinical Distribution (OCD): We define ICD and OCD data as follows, based on the presence of a CIR:

- **ICD Data:** An image contains a CIR, where an abnormality or clinically significant finding is clearly present and requires further examination, diagnosis, or treatment.
- **OCD Data:** An image without a CIR, where no abnormality or clinically significant finding could be detected. It is a normal image that does not contain significant medical concerns or require further clinical investigation.

Here, anatomical views showing identifiable fetal structures are considered ICD data, while background frames lacking clear anatomical details are considered OCD data. The goal of OCD detection is to distinguish anatomical views from background frames in real-time ultrasound videos, which is crucial for two reasons. First, only standard planes are typically annotated and saved during scanning, yet anatomically informative frames appear throughout the scanning. Identifying these additional views offers deeper insights into fetal health and development. Second, detecting these frames provides valuable visual guidance for recognising critical anatomical information in real-time, potentially reducing the need for extensive training and experience for sonographers.

3 OCD Detection Methodology

Here D_{train} and D_{test} represent dataset used for classification training and OCD detection testing, respectively. D_{train} contains labeled ICD data only, where $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ and $y_i \in \{1, 2, ..., L\}$. D_{test} consists of both ICD and OCD data. The task is to distinguish whether a testing input is in or out of clinical distribution and consists of three steps as depicted in Fig. 1.

3.1 Training Stage

Classification Model. We adapte the approach from [15] to project ICD features obtained from the classification model onto a union of 1-dimensional subspaces, with each class occupying a distinct, mutually orthogonal subspace. Such projection provides a stronger constraint on the feature space compared to

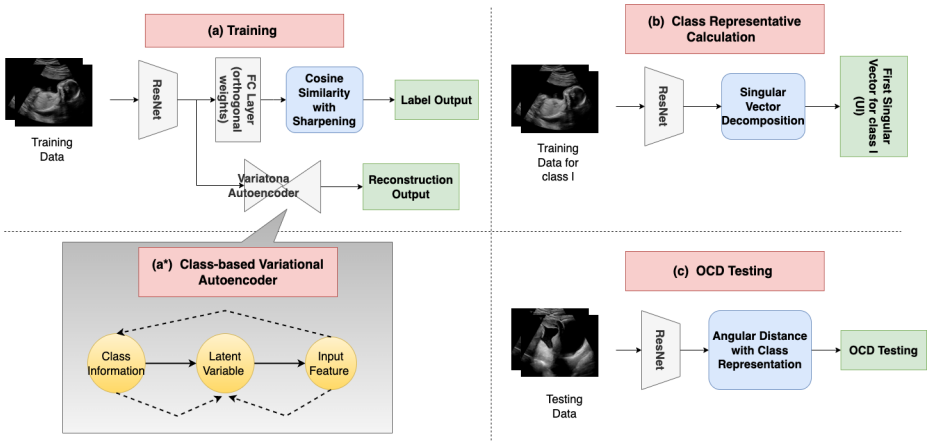


Fig. 1: OCD detection framework: (**a**): Training stage (**b**): Class representative calculation. (**c**): OCD detection stage. (**a**$*$): graphical model of the proposed SC-VAE ($\rightarrow$: generative model; $\dashrightarrow$: inference model).

traditional approach. To achieve this feature embedding, two key adjustments were made. First, the weights of the last fully connected layer are initialised with predefined orthonormal vectors and kept frozen during training. Second, cosine similarity is utilised in the calculation of the softmax function. For a given input x_i, the predicted probability of belonging to class l is calculated as $p_{i,l} = \frac{c_{i,l}}{\sum_j c_{i,j}}$, followed by a sharpening operation to enhance class discrimination. Here, $c_{i,l} = \frac{w_l^T f_i}{||f_i||}$ denotes the cosine similarity between the weight vector w_l of the last fully connected layer corresponding to class l and the feature vector f_i of input x_i. Therefore, the feature vectors from class l are constrained to form a 1-dimensional subspace aligned with the direction of w_l.

Softmax-Conditioned Variational Autoencoder (SC-VAE). In standard VAE, an isotropic Gaussian is commonly used as the prior over the latent space. However, this choice can hinder the model's ability to learn more complex representations [2]. For instance, when a shared prior is used for latent features of inputs from different classes, class-specific information may be lost during reconstruction. To address this, we propose incorporating softmax scores into the VAE to provide class membership information, thereby segregating the latent space into distinct classes to improve the quality of learned representations.

We use a mixture of class-conditioned Gaussian distributions as prior, with softmax scores providing a probabilistic measure of how likely a sample belongs to each Gaussian distribution. We assume that OCD features deviate from ICD features during training due to significant differences in reconstruction error. This deviation may be amplified by adding softmax scores, as ICD scores are typically biased toward the correct class, while OCD scores are more uniformly spread across incorrect classes, leading to larger reconstruction errors. This modified VAE, referred to as the softmax-conditioned VAE, is applied to the feature

space as a regulariser to enhance the model's ability to learn meaningful, class-specific features, therefore improving feature compactness within each class.

Here, we hypothesis that enforcing the compactness of each ICD class can improve OCD detection. Assuming that OCD samples and ICD samples from class l follow two multivariate Gaussian distributions, $\mathcal{N}(\mu_O, \Sigma_O)$ and $\mathcal{N}(\mu_l, \Sigma_l)$, respectively, the difference between these distributions can be quantified using the Bhattacharyya distance. This distance is calculated as the sum of the Mahalanobis distance between μ_O and μ_l and a measure of compactness for both distributions. Therefore, by making each $\mathcal{N}(\mu_l, \Sigma_l)$ more compact, the Bhattacharyya distance increases, enhancing the separation between OCD and ICD distributions and improving OCD detection performance.

Fig.1(a*) shows the graphical model of the proposed SC-VAE, where softmax scores are incorporated in both generative and inference processes. In the generative model, the latent variable z_i is sampled from class-conditioned Gaussian distributions, with the probability of sampling from a specific class distribution determined by its softmax score. The probability density function (pdf) of z_i conditioned on class l is formalised as:

$$p_\theta(z_i|l) \sim N(\mu_\theta(l), diag(\sigma^2_\theta(l))), \tag{1}$$

where $\mu_\theta(\cdot)$ and $\sigma^2_\theta(\cdot)$ are generated by a class-conditioned encoder parameterised by $\theta(\cdot) : \mathbb{R}^L \to \mathbb{R}^e$. $p_\theta(z_i|l)$ is used as the Gaussian prior for class l.

In the inference model, the posterior of z_i is dual-conditioned on both the input feature and class, with its pdf given by:

$$q_\phi(z_i|f_i, l) \sim N(\mu_\phi(f_i, l), diag(\sigma^2_\phi(f_i, l))), \tag{2}$$

where $\mu_\phi(\cdot)$ and $\sigma^2_\phi(\cdot)$ are generated by a feature-class dual-conditioned encoder parametrised by $\phi(\cdot) : \mathbb{R}^{d+L} \to \mathbb{R}^e$, d denotes the dimension of latent space.

The designed SC-VAE is trained using evidence lower bound optimisation (ELBO) [12], which involves minimising two losses: the reconstruction loss between the input features and generated targets (measured by MSE) and the distribution discrepancy between the prior and posterior in the latent space (measured by KL divergence). Given the softmax score $p_{i,l}$, which indicates the likelihood of input x_i belonging to Gaussian distribution conditioned on class l in the latent space. The loss function of the designed SC-VAE is defined as:

$$\mathcal{L}_{vae} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L} p_{i,l} \times \{MSE(f_i, \hat{f}_{i|l}) + KL[q_\phi(z_i|f_i, l)||p_\theta(z_i|l)]\}. \tag{3}$$

We jointly train the backbone classification model with the proposed SC-VAE regulariser, with the loss function defined as:

$$\mathcal{L}_{train} = \alpha_{cls} \times \mathcal{L}_{cls} + \alpha_{vae} \times \mathcal{L}_{vae}, \tag{4}$$

where α_{cls} and α_{vae} are the weights for classification loss and VAE loss. KL distances are estimated using a single latent sample [2]. A potential issue of "posterior collapse" is avoided as priors $p_\theta(z_i|l)$ are learnable and informative.

3.2 Class Representative Calculation

The first singular vector has been used as a robust estimator of mean and covariance in statistics [4]. According to [16], the first singular vector can also be treated as a class representative, as it preserves most of the information and is robust to noise and augmentation. Here, we use it as a representative for each ICD class. For ICD class l, its corresponding first singular vector u_l is obtained using singular value decomposition (SVD) on the matrix formed by aggregating the input features of the training data from that class.

3.3 OCD Detection

At test time, the minimum cosine similarity between the test feature f_t and each class representative $\{u_l\}_{l=1}^{L}$ is used as the uncertainty score for OCD detection, and it is calculated as: $\delta_t = \min_l(\arccos(\frac{f_t^T u_l}{||f_t^T||}))$. The corresponding test input x_t is classified as ICD or OCD based on the Eq. 5, where δ_T is the chosen threshold.

$$y_t = \begin{cases} 0(\text{ICD}) & \text{if } \delta_t < \delta_T \\ 1(\text{OCD}) & \text{otherwise} \end{cases} \tag{5}$$

The probability of a test input x_t belonging to ICD could also be calculated using probability function $P(\delta_t < \delta_T | x_t \in D_{test})$. By employing Monte Carlo sampling on δ_n, the probability function is estimated as $\frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(\delta_t^m < \delta_T)$, where M is the number of Monte Carlo samples and δ_t^m is the m-th sample.

4 Results

4.1 Ultrasound Data

We use a large-scale routine clinical fetal ultrasound dataset in our experiments. 342 second-trimester scans are used for training and testing. Around 6,650 freeze-frame images have been saved and manually annotated as one of 13 standard plane views - 4CH, 3VV, RVOT, LVOT, Brain (cb.), Brain (tv.), Lips, Abdominal, Kidney, Femur, Spine (Sag.), Spine (Cor.) and Profile. We used these frames to create the ICD dataset of 13 classes, splitting 80% for training and 20% for testing. Since a huge number of background frames and non-standard anatomical views remain unlabeled and largely indistinguishable, there is no ready-to-use OCD dataset. To address this, we manually created an OCD dataset by selecting 1,085 background frames based on three specific criteria: frames containing only dark amniotic fluid without visible fetal anatomy, frames with ultrasound artifacts (such as acoustic shadows or speckle noise) that obscure anatomical details, or frames with empty regions when the probe was briefly moved away from the fetus. These criteria ensured that the OCD dataset contained no clear anatomical information. The testing dataset for OCD detection includes both ICD testing data and OCD data.

Table 1: Evaluation Metrics for OCD detection

Metric	Description
Detection Error (%) ↓	Minimum misclassification rate over all possible thresholds.
AUROC ↑	Area under the FPR against TPR curve.
δ_T at 95% TPR ($\times 10^{-2}$) ↑	Angular distance based decision boundary at 95% TPR.
FPR at 95% TPR (%) ↓	Percentage of ICD data that wrongly classified as OCD at 95% TPR.

Table 2: Quantitative comparison of OCD performance on fetal US. The best results in each of the three method categories are marked in **bold**.

Method	FPR at 95% TPR (%) ↓	δ_T at 95% TPR ($\times 10^{-2}$) ↑	Detection Error (%) ↓	AUC ↑
Classification-based				
MSP [9]	25.8	38.92	11.32	94.32
ODIN [1]	18.64	43.58	8.92	96.13
Mahalanobis [11]	32.57	29.37	15.21	87.62
U1D [15]	12.7	50.25	7.86	97.19
SC-VAE Reg. (ours)	**9.17**	**57.45**	**6.86**	**97.53**
Generative-based				
AE	47.85	21.68	19.82	81.55
MemAE [6]	21.58	39.73	12.88	88.69
DDPM [7]	**12.06**	**54.18**	**7.41**	**95.51**
Dual Heads				
DDPM + ours	**9.12**	**57.92**	**6.69**	**97.64**

4.2 Implementation Details

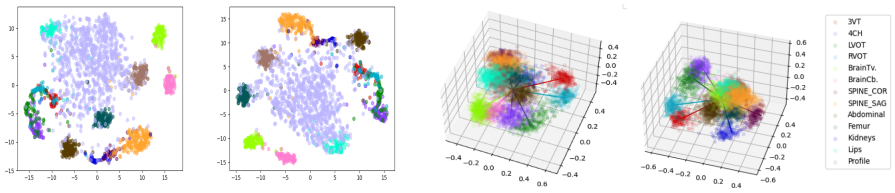
We use ResNet-18 [8] as the backbone architecture for the classification model. The proposed SC-VAE and classification model were jointly trained using the PyTorch framework and an SGD optimiser for 200 epochs with a batch size of 32, a moment of 0.9, a weight decay of 1×10^{-4} and an initial learning rate 0.1 that decayed by 0.1 after every 60 epochs. Data augmentation was performed including horizontal flipping, brightness variation, resizing, and center cropping.

4.3 Evaluation Metrics

During evaluation, OCD and ICD images are treated as positive and negative samples, respectively. Four metrics are used to evaluate OCD detection performance, as summarised in Table 1. Among these metrics, keeping a low FPR is particularly important clinically, as it helps avoid missing valuable information or questionable cases that could lead to severe consequences.

4.4 Quantitative Results

Table 2 presents a quantitative evaluation of OCD performance in comparison with different methods. The table indicates that our proposed mothod with the SC-VAE regulariser outperformed four selected state-of-the-art classification-based methods, i.e. MSP [9], ODIN [1], Mahalanobis [11], and Union of 1-D (U1D) [15] by 16.63%, 27.81%, 23.4%, and 3.53% respectively in FPR at 95% TPR. These improvements are statistically significant under the two-sided t-



(a) t-SNE (U1D) (b) t-SNE (ours) (c) PCA (U1D) (d) PCA (ours)

Fig. 2: PCA and T-SNE visualizations of baseline and PP-VAE boosted methods

test. The improvement in δ_T indicates a larger ICD conical region spanned from each class representative vector, while keeping 95% of OCD samples outside the region, which in line with the hypothesis that the proposed design enhances the separation between ICD and OCD data and is crucial for near-OCD cases. Among generative-based methods, AE performs the worst, likely due to its limitations in recognising and reconstructing complex medical images as a simple reconstruction approach. This is particularly challenging when near-OCD appears, as AE may fail to capture subtle differences. The improved performance of MemAE [6] highlights the importance of incorporating class information in OCD detection, which is also a key aspect of our design. Notably, our method also outperformed DDPM for OCD detection [7], a recent trend known for its effectiveness but complexity. Compared to DDPM, our approach offers a simpler yet effective alternative, benefiting from both computational complexity and performance. We also test a dual-head model where our method and DDPM are trained simultaneously. Although this setup shows a slight, but not statistically significant, improvement in performance, the substantially increase in computational cost might not worth the trade-off. Our method offers the best balance between computational complexity and performance.

4.5 Qualitative Results

Fig. 2 presents two types of visulisation of testing features, in which our method is compared with U1D, the second-best-performing classification-based model. The t-SNE plots provide a 2-D visualisation of both ICD and OCD features, showing greater separation of OCD features from ICD, especially for classes similar to OCD. This improvement in performance for classes near OCD is achieved without compromising performance on classes that are more distinct. The 3D PCA plots display both class representatives (solid lines) and ICD testing features (circles). The clustering of testing features centered around each solid line validates the choice of the first singular vector as class representatives. Additionally, the reduced spread of circles around the solid lines indicates increased compactness within each ICD class.

Table 3: Ablation study on the effectiveness of adding different feature regulariser. The best results are marked in **bold**.

Metric	No Reg.	+AE	+ vanilla CC-VAE	+CC-VAE	+SC-VAE (ours)
FPR at 95% TPR (%) ↓	12.70	12.84	11.34	12.17	**9.17**
δ_T at 95% TPR ($\times 10^{-2}$) ↑	50.25	54.87	56.93	55.13	**57.45**
Detection Error (%) ↓	7.86	8.23	7.72	7.72	**6.86**
AUROC ↑	97.19	96.84	97.03	97.07	**97.53**

4.6 Ablation Studies

An ablation study is conducted to evaluate the effectiveness of adding different regularisers to the feature space. The results are shown in Table 3. Here, CC-VAE refers to a class-conditioned VAE where classification labels are incorporated as class information. However, assigning incorrect classes to OCD samples could lead to biased results and hinder feature learning. Vanilla CC-VAE uses a standard Gaussian prior, while CC-VAE employs a learnable Gaussian prior. Our proposed SC-VAE regulariser achieves the best performance, aligning with our hypothesis and demonstrating the effectiveness and robustness of our design - incorporating a learnable Gaussian prior for each class and using softmax scores to provide more accurate class information.

5 Conclusion

In this paper, we introduce the concept of "out-of-clinical-distribution" detection for medical image analysis. We define ICD data as images containing clinically significant anatomical information, while OCD data refers to images lacking such information for clinical decision-making. We propose a novel OCD detection method by introducing a softmax-conditioned VAE as the features regulariser, which incorporates softmax scores into both the inference and generative models with a class-conditioned mixture Gaussian prior, to enhance ICD feature compactness and OCD detection performance. Experiments on fetal ultrasound video show the effectiveness of the proposed method over both classification-based and generative-based approaches. It achieves comparable results when using a dual-head model trained with DDPM, indicating that our method benefits from both computational complexity and performance. In the future, the proposed concept could be applied to different data modalities and tasks, and the proposed method could be adapted for use across various data types.

Acknowledgement

Kangning Zhang would like to thank the support from the ERC Project PULSE ERC-ADG-2015 694581 and the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1). Jianbo Jiao is supported by the Royal Society Short Industry Fellowship (SIF\R1\231009).

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In: Proceedings of the 2018 CHI conference on human factors in computing systems. pp. 1–18 (2018)
2. Bai, J., Kong, S., Gomes, C.P.: Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In: International Conference on Machine Learning. pp. 1383–1398. PMLR (2022)
3. Cao, T., Huang, C.W., Hui, D.Y.T., Cohen, J.P.: A benchmark of medical out of distribution detection. arXiv preprint arXiv:2007.04250 (2020)
4. Diakonikolas, I., Kamath, G., Kane, D.M., Li, J., Moitra, A., Stewart, A.: Being robust (in high dimensions) can be practical. In: International Conference on Machine Learning. pp. 999–1008. PMLR (2017)
5. Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. Advances in Neural Information Processing Systems **34**, 7068–7081 (2021)
6. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1705–1714 (2019)
7. Graham, M.S., Pinaya, W.H., Tudosiu, P.D., Nachev, P., Ourselin, S., Cardoso, J.: Denoising diffusion models for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2948–2957 (2023)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
10. Hong, Z., Yue, Y., Chen, Y., Cong, L., Lin, H., Luo, Y., Wang, M.H., Wang, W., Xu, J., Yang, X., et al.: Out-of-distribution detection in medical image analysis: A survey. arXiv preprint arXiv:2404.18279 (2024)
11. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
12. Mourelatos, Z.P., Zhou, J.: A design optimization method using evidence theory (2006)
13. Ulmer, D., Meijerink, L., Cinà, G.: Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In: Machine Learning for Health. pp. 341–354. PMLR (2020)
14. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4921–4930 (2022)
15. Zaeemzadeh, A., Bisagno, N., Sambugaro, Z., Conci, N., Rahnavard, N., Shah, M.: Out-of-distribution detection using union of 1-dimensional subspaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9452–9461 (2021)
16. Zaeemzadeh, A., Joneidi, M., Rahnavard, N., Shah, M.: Iterative projection and matching: Finding structure-preserving representatives and its application to computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5414–5423 (2019)