Out-of-Distribution Detection in Medical Imaging via Diffusion Trajectories

Lemar Abdi, Francisco Caetano, Amaan Valiuddin, Christiaan Viviers, Hamdi Joudeh, and Fons van der Sommen

Eindhoven University of Technology, Eindhoven 5612 AZ, The Netherlands
l.abdi@tue.nl

Abstract. In medical imaging, unsupervised out-of-distribution (OOD) detection offers an attractive approach for identifying pathological cases with extremely low incidence rates. In contrast to supervised methods, OOD-based approaches function without labels and are inherently robust to data imbalances. Current generative approaches often rely on likelihood estimation or reconstruction error, but these methods can be computationally expensive, unreliable, and require retraining if the inlier data changes. These limitations hinder their ability to distinguish nominal from anomalous inputs efficiently, consistently, and robustly. We propose a reconstruction-free OOD detection method that leverages the forward diffusion trajectories of a Stein score-based denoising diffusion model (SBDDM). By capturing trajectory curvature via the estimated Stein score, our approach enables accurate anomaly scoring with only five diffusion steps. A single SBDDM pre-trained on a large, semantically aligned medical dataset generalizes effectively across multiple Near-OOD and Far-OOD benchmarks, achieving state-of-the-art performance while drastically reducing computational cost during inference. Compared to existing methods, SBDDM achieves a relative improvement of up to 10.43% and 18.10% for Near-OOD and Far-OOD detection, making it a practical building block for real-time, reliable computer-aided diagnosis.

1 Introduction

Out-of-distribution (OOD) detection is a critical task in medical imaging, where real-world deployments often face data that diverges from the training distribution due to rare pathologies, acquisition variability, or previously unseen patient demographics. These challenges, which are aggravated by low prevalence, severe class imbalance, and the high cost of expert annotation, undermine the effectiveness of supervised approaches [15]. In response, unsupervised OOD detection has emerged as a promising alternative, which is solely trained on the readily available, typically more abundant non-pathological cases.

Unsupervised OOD approaches typically operate by modeling the inlier distribution of training data, explicitly through likelihood estimation [25,32], or implicitly through distance metrics at the pixel or feature level [5,14]. Reconstruction-

based methods attempt to flag anomalies based on reconstruction error, assuming that models trained on inlier data will fail to accurately reconstruct unseen or pathological structures. Among these, Denoising Diffusion Models (DDMs) have shown strong reconstruction performance in medical imaging [2, 19, 28], but suffer from key limitations: they require multiple of sampling steps in the forward and reverse process, are sensitive to noise level choices, and can also inadvertently reconstruct anomalous inputs faithfully—leading to false negatives. Moreover, conditioning mechanisms used during reconstruction may introduce dataset-specific biases. To overcome these challenges, recent works [9, 21] have proposed reconstruction-free OOD detection using the *forward* (noising) process of DDMs. These methods hypothesize that the diffusion trajectory of anomalous inputs deviates from that of nominal data, allowing effective OOD detection with significantly fewer model evaluations.

In this work, we extend forward diffusion-based OOD detection to the medical domain by leveraging the trajectories of DDMs, as realized through the estimated Stein scores. Our method avoids the limitations of reconstruction-based approaches and achieves strong performance, using as few as five diffusion steps. We further show that the learned score function generalizes across datasets, enabling OOD detection on multiple medical benchmarks with a single pre-trained model. The proposed approach outperforms state-of-the-art (SOTA) methods across twelve diverse medical imaging benchmarks in both Near-OOD and Far-OOD settings.

2 Related Work

2.1 Score-based Diffusion Models

Score-based diffusion models gradually perturb data $\mathbf{x}_0$ from data distribution p_0 into a simple prior distribution p_T, typically a standard Gaussian, through a stochastic differential equation (SDE). The forward diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ is governed by the SDE:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{f}(\mathbf{x}, t)$ denotes the drift term, $g(t)$ is a scalar diffusion coefficient, and $\mathbf{w}$ is a standard Wiener process. Sampling from the learned data distribution involves reversing this diffusion process. The reverse-time SDE is given by:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{2}$$

where $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the *Stein score* (or score function), and $\bar{\mathbf{w}}$ denotes the reverse Brownian motion over $[T, 0]$. Song *et al.* [24] showed that this stochastic process has a deterministic counterpart, known as the probability flow ODE, which yields the same marginal distributions as the original SDE. This formulation allows the diffusion process to be expressed as:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t. \tag{3}$$

Both the reverse-time SDE and the probability flow ODE require an accurate estimate of the score function. This is typically achieved by training a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x}, t)$ to approximate the Stein score function, using the score matching objective [26]:

$$\min_\theta \mathbb{E}_{t\sim\mathcal{U}[0,T]} \left\{ \mathbb{E}_{\mathbf{x}_0\sim p_0(\mathbf{x}_0)} \mathbb{E}_{\mathbf{x}_t\sim p_{0t}(\mathbf{x}_t|\mathbf{x}_0)} \left[\left\| \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right) - \boldsymbol{\epsilon} \right\|_2^2 \right] \right\}, \tag{4}$$

where the target denoising direction is defined as $\boldsymbol{\epsilon} = -\sigma_t \nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t \mid \mathbf{x}_0)$ and $p_{0t}(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, \sigma_t^2 \mathbf{I})$, following the DDPM formulation [10] for the fixed time-dependent variances $\sigma_t^2 = (1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)$.

2.2 Unsupervised OOD Detection

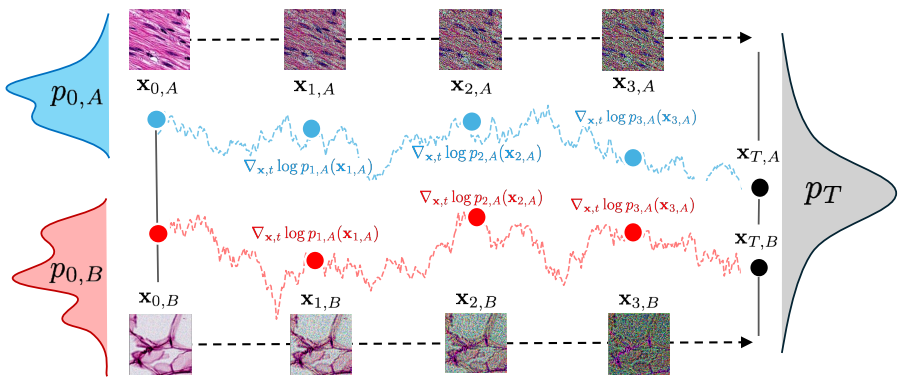
Given an in-distribution (ID) dataset sampled from p_{ID}, the goal of unsupervised OOD detection is to identify test-time samples that lie outside high-probability mass regions of p_{ID}, without requiring labels or prior knowledge of OOD samples. Existing approaches generally fall into three categories, described as follows. (1) Distance-based methods (also known as feature-based methods) compare test samples with training data in a learned embedding space, using metrics such as cosine similarity or Mahalanobis distance. These approaches rely on the assumption that OOD samples are farther from the ID feature manifold. Although effective, they often require storing training embeddings or class prototypes in memory [31]. (2) Density-based methods attempt to estimate $\log p_{\text{ID}}(\mathbf{x})$ directly. Normalizing Flows (NFs) enable exact likelihood computation, while VAEs and diffusion models yield tractable bounds on the likelihood. Despite their theoretical motivation, deep generative models often assign higher likelihoods to OOD inputs, undermining their effectiveness [3,13,17]. (3) Reconstruction-based methods assume that OOD samples will be poorly reconstructed by generative models trained on ID data, which can be quantified in pixel-space or feature-space. However, these methods often fail in complex settings—such as medical imaging—where anomalous samples can still be faithfully reconstructed [5,33].

OOD Detection with Diffusion Models: Diffusion models have gained traction in the context of OOD detection, primarily through reconstruction-based strategies. These methods typically employ class-conditioned DDMs, where the model learns to reconstruct inputs conditioned on non-pathological (often *healthy*) class labels [16,28]. Recent methods improve robustness by ensembling multiple reconstructions for distance computations or conditioning [2,19]. Nevertheless, these methods require many sampling steps to reconstruct an image—making real-time deployment less practical in clinical settings. Density-based approaches using DDMs are also feasible. The probability flow ODE (eq. (3)) allows for the exact likelihood computation under score-based diffusion models. However, this approach still suffers from the same limitations seen in other likelihood-based approaches [7]. More recently, reconstruction-free strategies have emerged that take advantage of the forward (noising) process of diffusion models [9,21]. These methods hypothesize that the forward trajectory of an anomalous sample

deviates significantly from that of inlier data. In particular, they achieve strong performance while requiring a fewer number of function evaluations (NFEs), highlighting their potential for clinical applications.

3 Method

3.1 Stein Score DDM OOD Detection



$$D_{\mathrm{KL}}(p_{0,A}\|p_{0,B}) = \frac{1}{2}\int_0^T \mathbb{E}_{\mathbf{x}\sim p_{t,A}} \frac{g(t)^2}{\sigma_t} \| -\sigma_t \nabla_{\mathbf{x},t} \log p_{t,A}(\mathbf{x}_{t,A}) - (-\sigma_t \nabla_{\mathbf{x},t} \log p_{t,B}(\mathbf{x}_{t,B}))\|_2$$

Fig. 1: Distinct data distributions induce different curvatures along the forward diffusion path. The estimated Stein score characterizes whether a given trajectory conforms to the curvature of the ID density, thereby enabling a reconstruction-free OOD detection approach with unconditional DDMs.

Given two distinct data distributions $p_{0,A}$ and $p_{0,B}$, the intermediate marginals $p_{t,A}$ and $p_{t,B}$ during the diffusion process are governed by the deterministic probability flow ODE of Equation (3). These marginals along the forward-time ODE path—i.e., from sample to noise—differ depending on the data distribution. Under regularity conditions, the divergence between these two distributions is equivalent to the squared difference between their Stein scores, as shown in [9]:

$$D_{\mathrm{KL}}(p_{0,A}\|p_{0,B}) = \frac{1}{2}\int_0^T \mathbb{E}_{\mathbf{x}\sim p_{t,A}} \frac{g(t)^2}{\sigma_t} \|\boldsymbol{\epsilon}_A - \boldsymbol{\epsilon}_B\|_2 + \underbrace{D_{\mathrm{KL}}(p_{T,A}\|p_{T,B})}_{=0}, \quad (5)$$

where $\boldsymbol{\epsilon}_i \approx -\sigma_t \nabla_{\mathbf{x},t} \log p_{t,i}(\mathbf{x}_{t,i})$ denotes the estimated Stein score at time t. The estimated Stein score reflects the local gradient of the marginals in the diffusion process, effectively capturing the direction and rate of change of the forward

trajectory. Its evolution over time—i.e., the variation of the score along the diffusion path—captures curvature, which further distinguishes *nominal* from *anomalous* samples, as illustrated in Figure 1. Using this, the anomaly score $s(\cdot)$ is defined for a test sample $\mathbf{x}_0^*$ as:

$$s(\mathbf{x}_0^*) = \left\| \sum_{t=1}^{S} -\sigma_t \nabla_{\mathbf{x},t} \log p_t(\mathbf{x}_t^*) \right\|^p + \left\| \sum_{t=1}^{S} \frac{\partial}{\partial t} \Big(-\sigma_t \nabla_{\mathbf{x},t} \log p_t(\mathbf{x}_t^*) \Big) \right\|^p, \tag{6}$$

where S denotes the number of forward-time diffusion steps, and $\|\cdot\|_p$ the p-norm. Similarly to Heng *et al.* [9], a Kernel Density Estimation (KDE) with a Gaussian kernel is fitted to the scores of the inlier validation samples $s(\mathbf{x}_0^{\text{val}})$ of each benchmark. At test time, a sample $\mathbf{x}_0^*$ is marked as OOD if the likelihood of its score $s(\mathbf{x}_0^*)$ under the KDE is less than $s(\mathbf{x}_0^{\text{val}})$.

We follow OpenAI's `improved diffusion`[1] codebase and architectural design. The models are trained using a cosine noise schedule, uniform time sampling, a learning rate of 1e-4, and a batch size of 64. For both training and inference, we use the DDIM sampler [23]. During testing, the Stein scores in Equation (6) are computed using the $\boldsymbol{\epsilon}$ output of the `ddim_reverse_sample` function. We fix $S = 5$, $p = 3$, and compute the time derivative using finite difference.

Unlike many existing methods, this approach does not require retraining on each inlier dataset, allowing application across multiple datasets with a single model. The proposed unconditional DDM relies on a well-trained score estimator and a KDE fitted to the validation data of the inlier dataset, which ensures that Equation (5) remains valid regardless of the dataset used to train $\boldsymbol{\epsilon}_\theta$. However, the choice of training data remains important: the model must learn features that generalize well within the medical domain. We empirically analyze this dependency and its impact on cross-dataset generalization in Section 4.2.

3.2 Medical OOD Benchmark

We adopt the OpenOOD framework [30] in the medical domain for unsupervised image-level OOD detection across the Near-OOD and Far-OOD regimes. We adopt five MedMNIST datasets [29] and the COVID-19 dataset [4, 20] to cover a spectrum of semantic and modality shifts. We define *Near-OOD* as a semantic shift within the same dataset: samples from classes not seen during training but drawn from the same support. Specifically, we split each dataset by aggregating training classes into majority/minority partitions and treat the excluded subset as OOD. For binary-class datasets, this corresponds to 'healthy' vs. 'pathological'; for multi-class datasets, it involves excluding minority or semantically disjoint classes. This partitioning aligns with the clinically relevant setting of detecting the pathologies absent in the training data. Our splitting protocol follows Narayanaswamy *et al.* [18] to ensure a consistent comparison. In contrast, *Far-OOD* corresponds to modality shifts across datasets. In this setting, one dataset is designated as ID, and all the remaining datasets serve as sources of OOD data. All images are resized to 128×128 pixels.

[1] https://github.com/openai/improved-diffusion

3.3 Baselines

We compare our proposed method with a range of state-of-the-art (SOTA) approaches, including both supervised and unsupervised methods:

Classification-based: We include two semi-supervised approaches with outlier-exposure—VOS++ and NDA++ [6, 8, 18, 22]—which are trained on additional synthetic samples compared to their vanilla models. G-ODIN [11] is an unsupervised confidence-based method that combines temperature scaling and input perturbation to improve softmax OOD discrimination.

Density-based: Mubarka *et al.* [18] propose an energy-based model trained on additional synthetic samples, achieving the SOTA OOD performance in MedMNIST prior to this work. We also evaluate GLOW [12], a normalizing flow (NF) model evaluated on the exact data log-likelihood, and Typ-Score [1, 27], an NF method that quantifies sample typicality using the Stein score of the data log-likelihood.

Reconstruction-based: We include a DDM conditioned on "healthy" samples [7], using a DDIM [23] sampler. For anomaly scoring, we evaluate both a combined feature-level LPIPS and pixel-level L2 error, as well as LPIPS alone, denoted as cDDM (L).

Forward diffusion-based: Sakai *et al.* [21] propose a reconstruction-free diffusion method, which similarly to SBDDM(-P), obtains anomaly scores using solely the forward-time diffusion path. We follow the same hyperparameter settings, with the model conditioned on healthy inlier images.

Table 1: AUROC scores (Near-OOD / Far-OOD). Color-coded results are obtained by SBDDM-P trained on PathMNIST or TissueMNIST. (*) denotes re-implementation.

Method	Blood-MNIST	Derma-MNIST	Path-MNIST	Tissue-MNIST	Pneu-monia	Covid-19	Avg.
G-ODIN	53.9 / 88.7	69.3 / 85.3	51.7 / 84.4	55.2 / 82.7	-	-	57.5 / 85.3
VOS++	38.2 / 84.2	72.9 / 85.3	71.7 / 71.0	28.0 / 60.2	-	-	52.7 / 75.2
NDA++	53.5 / 95.8	69.9 / 80.0	57.0 / 61.1	58.8 / 70.2	-	-	59.8 / 76.8
Mubarka	89.1 / 99.7	75.5 / 96.6	71.2 / **98.9**	**83.4** / 96.6	-	-	79.8 / 97.9
GLOW*	33.9 / 45.4	53.9 / 23.7	64.0 / 33.1	65.0 / 50.0	55.7 / 51.2	65.4 / 09.0	56.3 / 35.4
Typ-score*	55.7 / 91.4	70.8 / **100**	82.2 / 81.7	70.7 / 95.3	**90.4** / 91.7	93.7 / 57.3	77.3 / 86.2
cDDM (L)*	71.1 / 78.7	74.4 / 73.2	81.2 / 83.0	44.3 / 63.9	81.5 / 86.2	69.9 / 82.3	70.4 / 77.9
cDDM*	84.2 / 86.0	74.8 / 66.0	83.4 / 91.4	62.4 / 80.0	82.6 / 72.8	67.2 / 68.2	75.8 / 77.4
Sakai*	56.5 / 77.4	72.0 / 82.5	91.5 / 97.4	58.0 / 86.0	82.2 / 87.9	**95.6** / 87.7	76.0 / 86.4
SBDDM	88.6 / **100**	67.9 / 99.9	**92.1** / 96.6	64.6 / **99.9**	73.8 / 96	80 / **97.2**	77.8 / **98.3**
SBDDM-P	**89.6** / 95.6	**77.3** / 99.1	**92.1** / 96.6	64.6 / **99.9**	86.8 / **100**	89.3 / 70.7	**83.3** / 93.7

4 Results and Discussion

4.1 OOD Benchmark Performance

We present the OOD detection performance of the proposed Stein score-based DDM (SBDDM) method against several baselines in Table 1, using the Area Under the Receiver Operating Characteristic curve (AUROC) as the primary metric. Far-OOD AUROC results are averaged across all OOD datasets per benchmark. The classification-based baseline results are sourced from [18]. We report two variants of our method: SBDDM, which is retrained per inlier dataset (analogous to the baselines), and SBDDM-P, which uses a single model pre-trained on a large medical dataset (PathMNIST for RGB benchmarks, TissueMNIST for grayscale) across all benchmarks. Interestingly, SBDDM-P often outperforms retrained counterparts, underscoring the benefits of domain-consistent Stein score estimation; we further analyze this effect in Section 4.2. Among the baselines, Typ-score achieves the best Far-OOD result on DermaMNIST and the best Near-OOD result on Covid-19. Mubarka *et al.* [18] achieve the best Far-OOD performance on PathMNIST and the best Near-OOD performance on TissueMNIST, while Sakai *et al.* [21] achieve the best Near-OOD performance on Covid-19. In the remaining seven benchmarks, the proposed method outperforms all baselines across both Near-OOD and Far-OOD settings. Additionally, when running on an NVIDIA H100 GPU, SBDDM has a latency of 55 ms (18.2 FPS) using 5 NFEs, compared to 940 ms (1.1 FPS) with 100 NFEs for the reconstruction-based cDDM baseline.

Table 2: Effect of training distribution p_{train} of the SBDMM on Near-OOD performance.

p_{train}	**BloodMNIST**	**Pneumonia**
ImageNet	0.597	0.456
BloodMNIST	0.886	0.738
PathMNIST	0.896	0.868
MedMNIST RGB	0.775	-
MedMNIST Gray	-	0.597

4.2 Ablations

We evaluate the impact of the training distribution p_{train} on OOD detection performance in Table 2. Using a score-based DDM trained on ImageNet results in significantly lower performance, confirming that generic natural image statistics are insufficient to capture meaningful diffusion patterns in the medical domain. Training on the same dataset used for ID testing—such as BloodMNIST—yields strong OOD detection performance, as expected. Surprisingly, training on PathMNIST, a larger medical dataset with distinct semantics,

leads to even better results. However, aggregating all the RGB datasets from MedMNIST into one large training set (MedMNIST RGB) and combining the grayscale datasets into another (MedMNIST Grey) introduces inconsistencies and often degrades performance. This may stem from the distortion of transport paths and flattening of curvatures that are informative for detecting near-OOD samples. These findings suggest that p_{train} should be drawn from a sufficiently large and semantically consistent dataset, ensuring that the score estimation generalizes to other benchmarks within the same domain. Although score-based

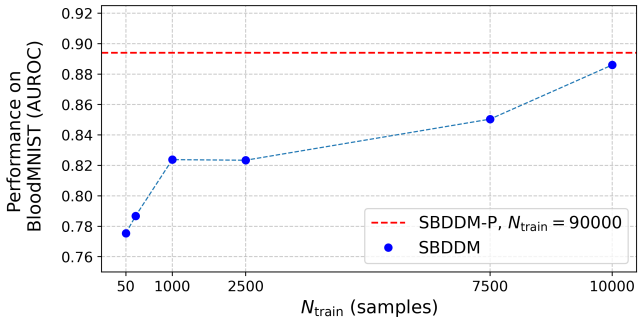


Fig. 2: When is retraining worth it? Performance on the BloodMNIST benchmark using SBDDMs trained on progressively larger subsets of 'healthy' images, compared against a SBDDM-P pre-trained on 90k PathMNIST samples.

DDMs trained on one dataset can generalize to others, we examine under what conditions retraining on the target inlier dataset becomes worthwhile. Specifically, we empirically show how much data is needed for a SBDMM to match the performance of a SBDDM-P model. Figure 2 presents performance on the BloodMNIST Near-OOD benchmark using SBDDMs trained on increasing subsets of 'healthy' BloodMNIST samples, compared to a SBDDM-P pre-trained on PathMNIST (approximately 90k samples). Results show that the SBDDM-P consistently outperforms SBDMMs trained on fewer than 10k samples. Only when the SBDDM model is trained on $\sim$10k samples does its performance begin to approach that of SBDDM-P. These findings highlight the proposed method's practicality under data-scarce conditions.

5 Conclusion

OOD detection is critical for medical imaging systems that must handle rare or unexpected anomalies without explicit supervision. We presented an efficient diffusion model for OOD detection in medical imaging, leveraging forward diffusion trajectories from score-based DDMs. Unlike prior DDMs, our method needs no class conditioning or costly reconstruction, using only five noising steps. Exploiting the Stein score captures both semantic and modality shifts between inlier

and anomalous samples. Our experiments show that a single SBDDM pre-trained on a large homogeneous medical dataset generalizes well across modalities and achieves strong performance across all benchmarks, requiring no fine-tuning beyond fitting a lightweight KDE. In contrast, SBDDM models trained on large heterogeneous or natural image datasets yield lower performance, highlighting the importance of domain-consistent training for robust score estimation. These findings highlight the potential of SBDDMs as practical, generalizable OOD detectors for clinical applications. Future work may explore extending this approach to volumetric imaging, semantic segmentation, and integrating uncertainty quantification to further enhance decision-making in real-world clinical settings. Additionally, while our method focuses on image-level OOD detection, it does not provide spatial localization of anomalies—a key advantage of DDM-based approaches. Incorporating localization into this framework is an important step toward interpretability in clinical workflows.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdi, L., Valiuddin, M.A., Viviers, C.G., de With, P.H., van der Sommen, F.: Typicality excels likelihood for unsupervised out-of-distribution detection in medical imaging. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 149–159. Springer (2024)
2. Behrendt, F., Bhattacharya, D., Mieling, R., Maack, L., Krüger, J., Opfer, R., Schlaefer, A.: Leveraging the mahalanobis distance to enhance unsupervised brain mri anomaly detection. In: International Conference on MICCAI. pp. 394–404 (2024)
3. Choi, H., Jang, E., Alemi, A.A.: WAIC, but Why? Generative Ensembles for Robust Anomaly Detection (May 2019), arXiv:1810.01392 [cs, stat]
4. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? Ieee Access **8**, 132665–132676 (2020)
5. Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., Vernekar, S.: Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. arXiv preprint arXiv:1812.02765 (2018)
6. Du, X., Wang, Z., Cai, M., Li, S.: Towards unknown-aware learning with virtual outlier synthesis. In: ICLR (2022), https://openreview.net/forum?id=TW7d65uYu5M
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 CVPR. pp. 2948–2957 (2023)
8. Hendrycks*, D., Mu*, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple method to improve robustness and uncertainty under data shift. In: ICLR (2020), https://openreview.net/forum?id=S1gmrxHFvB
9. Heng, A., Soh, H., et al.: Out-of-distribution detection with a single unconditional diffusion model. Advances in NeurIPS **37**, 43952–43974 (2024)

10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in NeurIPS **33**, 6840–6851 (2020)
11. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF CVPR. pp. 10951–10960 (2020)
12. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. NIPS **31** (2018)
13. Kirichenko, P., Izmailov, P., Wilson, A.G.: Why Normalizing Flows Fail to Detect Out-of-Distribution Data. In: NIPS. vol. 33, pp. 20578–20589. Curran Associates, Inc. (2020)
14. Li, J., Chen, P., He, Z., Yu, S., Liu, S., Jia, J.: Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In: Proceedings of the IEEE/CVF CVPR. pp. 11578–11589 (2023)
15. Li, J., Fong, S., Mohammed, S., Fiaidhi, J.: Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. The Journal of Supercomputing **72**(10), 3708–3728 (2016)
16. Mousakhan, A., Brox, T., Tayyub, J.: Anomaly detection with conditioned denoising diffusion models. In: DAGM German Conference on Pattern Recognition. pp. 181–195. Springer (2024)
17. Nalisnick, E.T., Matsukawa, A., Teh, Y.W., Görür, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: ICLR 2019, (2019)
18. Narayanaswamy, V., Mubarka, Y., Anirudh, R., Rajan, D., Thiagarajan, J.J.: Exploring inlier and outlier specification for improved medical ood detection. In: Proceedings of the IEEE/CVF ICCV. pp. 4589–4598 (2023)
19. Naval Marimont, S., Siomos, V., Baugh, M., Tzelepis, C., Kainz, B., Tarroni, G.: Ensembled cold-diffusion restorations for unsupervised anomaly detection. In: International Conference on MICCAI. pp. 243–253. Springer (2024)
20. Rahman, T.e.a.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in biology and medicine **132**, 104319 (2021)
21. Sakai, S., Hasegawa, T.: Reconstruction-free anomaly detection with diffusion models via direct latent likelihood evaluation. arXiv preprint arXiv:2504.05662 (2025)
22. Sinha, A., Ayush, K., Song, J., Uzkent, B., Jin, H., Ermon, S.: Negative data augmentation. In: ICLR (2021), https://openreview.net/forum?id=Ovp8dvB8IBH
23. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021), https://openreview.net/forum?id=St1giarCHLP
24. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021), https://openreview.net/forum?id=PxTIG12RRHS
25. Valiuddin, M.A., Viviers, C.G., van Sloun, R.J., de With, P.H., der Sommen, F.v.: Efficient out-of-distribution detection of melanoma with wavelet-based normalizing flows. In: CaPTion @ MICCAI2022 Workshop. pp. 99–107. Springer (2022)
26. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation **23**(7), 1661–1674 (2011)
27. Viviers, C., Valiuddin, A., Caetano, F., Abdi, L., Filatova, L., van der Sommen, F., et al.: Can your generative model detect out-of-distribution covariate shift? In: ECCV. pp. 184–201. Springer (2025)
28. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: MICCAI. pp. 35–45. Springer (2022)

29. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
30. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: Openood: Benchmarking generalized out-of-distribution detection. Advances in NeurIPS **35**, 32598–32611 (2022)
31. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024)
32. Zhao, Y., Ding, Q., Zhang, X.: Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection. In: The Eleventh ICLR (2022)
33. Zhou, Y.: Rethinking reconstruction autoencoder-based out-of-distribution detection. In: Proceedings of the IEEE/CVF CVPR. pp. 7379–7387 (2022)