Typicality Excels Likelihood for Unsupervised Out-of-Distribution Detection in Medical Imaging

Lemar Abdi, M.M. Amaan Valiuddin*, Christiaan G.A. Viviers*, Peter H.N. de With, and Fons van der Sommen

Eindhoven University of Technology, Eindhoven 5612 AZ, The Netherlands
m.m.a.valiuddin@tue.nl

Abstract. Detecting pathological abnormalities in medical images in an unsupervised manner holds potential for advancing modern medical diagnostics. However, supervised methods encounter challenges with exceedingly unbalanced training distributions due to limited clinical incidence rates. Likelihood-based unsupervised Out-of-Distribution (OOD) detection with generative models, especially Normalizing Flows, in which pathological abnormalities are considered OOD, could offer a promising solution. However, research in this direction has shown limited success as prior work has revealed that the likelihood does not accurately reflect the degree of anomaly for OOD samples, where in many instances higher likelihoods are assigned to anomalous samples compared to training samples. In this study, we present the first exploration of *typicality* (i.e. determining if samples belong to the *typical set*) for OOD detection in medical imaging, where test samples are juxtaposed against the probability mass rather than the density. The obtained findings demonstrate the superiority of evaluating typicality against likelihood for finding pathological abnormalities. We achieve state-of-the-art performance on the ISIC, COVID-19, and RSNA Pneumonia datasets, while being robust against significant data imbalances.

Keywords: OOD Detection · Typicality · Normalizing Flows

1 Introduction

The integration of data-driven deep learning algorithms in medical imaging has revolutionized healthcare at an unprecedented pace. However, a major challenge in the medical domain remains the strong imbalance between normal data and data with pathological abnormalities encountered in clinical practice. Although supervised methods have shown potency [25,34], both the extremely low incident rate of diseases and the challenges involved in collecting malignant samples lead to an inherent data imbalance, under which such models inadvertently fail [20]. To address these challenges, unsupervised out-of-distribution (OOD) detection

*Equal contribution

emerges as a popular approach, which solely models the abundantly available data of healthy cases. Thus, pathological abnormal examples can be identified by their low likelihood on the (implicit) model support. This alleviates the strenuous requirement of collecting scarcely available data and, therefore, it offers much greater potential in the upcoming second generation of AI for medical diagnosis.

Many unsupervised approaches implicitly evaluate OOD data through embedding-level distance metrics or image-level distances [1, 19], as well as explicit likelihood estimation with density modeling [15,37]. Unsupervised approaches assume that the training objective will exhibit higher cost for OOD data compared to in-distribution (ID) samples. Nonetheless, when dealing with intricate semantics, this disparity is hardly accurately reflected for distance-based approaches as shown in [9, 38]. In fact, cases exist where anomalous samples are faithfully reconstructed, thereby challenging the generalization of such approaches [27].

Density-based models aim to explicitly learn the approximate density of the available data, where generative models such as Normalizing Flows (NFs) are employed to directly evaluate the data log-likelihood on the model-inferred distribution. While this methodology is much more intuitive and theoretically justified, related literature presents strikingly contrasting findings [4,17,23]. For instance, studies have demonstrated that deep generative models trained solely on ID data with Maximum Likelihood Estimation (MLE), can assign higher likelihoods to OOD data [4, 23]. Furthermore, it is argued that this phenomenon occurs naturally, due to images not being high-likelihood samples, but rather elements of the *typical set* of the data distribution, where the empirical entropy of typical samples closely matches the entropy of the source distribution. Consequently, methods evaluating the typicality of data instances have been investigated as a surrogate for likelihood estimation [3, 12, 22].

Although OOD detection plays a crucial role in the adoption of Machine Learning in the medical domain, this phenomenon has unfortunately not been extensively explored with clinical data. In the context of medical images without pathological abnormalities, we argue that those samples are typical rather than most likely, and it is imperative that they are evaluated as such. The obtained findings clearly favour typicality as the superior learning objective over likelihood-based estimation on various medical image datasets, achieving state-of-the-art (SOTA) in the Area Under Curve (AUC) of the Receiver Operating Characteristics (ROC). The main contributions of this research are listed below:

- *Superior over likelihood evaluation*: Demonstration of the superiority of typicality over likelihood for OOD detection in medical imaging with NFs.
- *Bias removal*: Addressing the unfavourable intrinsic bias of NFs, which assigns likelihood based on textural complexity rather than semantic content.
- *Outperforms SOTA models*: SOTA results across four different medical image datasets of different modalities.
- *Assessment of data imbalance effects*: Exploration of data imbalance on the comparative performance of OOD detection w.r.t. supervised methods, exemplifying the performance of the proposed approach subject to exceedingly limited data.

2 Theoretical Background

2.1 Normalizing Flows

Consider image data $\mathbf{x} \sim P_{\mathbf{X}}$ taking values in high-dimensional space $\mathbb{R}^D$. To accurately model the data distribution from dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$, Normalizing Flows (NFs) are a family of models with fully tractable marginals. Training such a model entails optimizing a function that maps data $\mathbf{x}$ to the target density $p_{\mathbf{Z}}$ (usually a normal density). With K invertible functions $f_k : \mathbb{R}^D \to \mathbb{R}^D$ (for $k = 1, 2, \ldots, K$), intermediate variables $\mathbf{z}_k$ and the sequential mapping $\mathbf{z}_0 = f_K \circ f_{K-1} \circ ... \circ f_1(\mathbf{x})$, the exact likelihood can be determined by

$$\log p_{\mathbf{X}}(\mathbf{x}) = \log p_{\mathbf{Z}}(\mathbf{z}_0) - \sum_{i}^{K} \log \left| \det \frac{\partial f_k(z_{k-1})}{\partial \mathbf{z}_{k-1}} \right|. \tag{1}$$

Choosing the right bijection, f that balances expressivity and relatively cheap evaluation of the Jacobian determinant in Equation (1) is a crucial design choice. While many have been introduced, affine coupling layers have shown to work especially well for image data [10, 14–16].

Although other generative models have been explored for OOD detection in medical imaging, GANs [21] lack explicit density estimation, DDPMs require labeled guidance for competitive performance [32] or rely on less favorable reconstruction-based methods [11], and VAEs do not guarantee accurate inference on the lower bound of the likelihood [7, 36]. Given the advantage of obtaining exact likelihoods, our research explores NFs for OOD detection within the clinical setting. In this domain, NFs have been used for likelihood-based OOD detection of malignant Melanoma [30, 37]. According to the authors, NFs are mainly limited by the significant influence of textural information on the likelihood. In a similar way, various studies [2, 4, 23] indicate other severe limitations of NFs for OOD detection, by often assigning higher likelihood to OOD data rather than training data. Several explanations such as intrinsic biases [17] or entropic mismatch [2] have been discussed in the literature. Moreover, the information-theoretic perspective, which redirects attention from likelihood to typicality, demonstrates the most promising results [3, 12, 22].

2.2 Typicality

Given a random variable $\mathbf{X} \in \mathbb{R}$, we can define $\mathcal{X}^{(N)}$ as the set containing sequences of N i.i.d. datapoints $\{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N\}$. The typical set $A_\epsilon^{(N)} \in \mathcal{X}^{(N)}$ is said to contain sequences that satisfy

$$H(\mathbf{X}) - \epsilon \leq -\frac{1}{N} \sum_{n=1}^{N} \log_2 p(\mathbf{x}_n) \leq H(\mathbf{X}) + \epsilon, \tag{2}$$

for any small value ϵ and where $H(\mathbf{X})$ is the Shannon entropy. In other words, the empirical entropy is close to the entropy of the source distribution. As a

consequence of the Asymptotic Equipartition Property, it can be stated that

$$\frac{1}{N}\sum_{n=1}^{N}\log_2 p(\mathbf{x}_n) \to H(\mathbf{X}) \quad \text{s.t. } N \to \infty, \tag{3}$$

and thus $P(\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\} \in A_\epsilon^{(N)}) \approx 1$. Regardless of the fact that the typical set is only a small fraction of all possible sequences, any sequence of i.i.d. samples drawn of sufficient length is almost certainty in $A_\epsilon^{(N)}$, and thus typical. A common misconception attributes the most likely sequence as a member of $A_\epsilon^{(N)}$. A simple counterexample can be provided through a biased coin toss $\Omega = \{H, T\}$, with probabilities $P_H = 0.8$ and $P_T = 0.2$. Indeed, the sequence with all outcomes being H is most probable, with entropy $-\frac{1}{N}\log_2 0.8^N \approx 0.32$. However, this value remains far from the information content of the source $H(\Omega) \approx 0.72$, regardless of sequence length N.

An analogous argument can be made on the typicality of a sequence of images in dataset $\mathcal{D}$. Hence, Nalisnick *et al.* [22] argue that OOD samples should explicitly be evaluated on it. Consider a d-dimensional Gaussian distribution $\mathcal{N}(0, \mathbf{I}_d)$, where its $(\epsilon, 1)$-typical set $A_\epsilon^{(1)}$ primarily resides at a radius of $\sqrt{d}$ [22,31]. The *atypicality* of a sample is consequently measured by its deviation from this radius, quantified in latent space as $\left|\, ||\mathbf{z}||_2 - \sqrt{d} \,\right|$. This score is known as the typicality test in latent space (TTL) [**?**, 24]. However, the specific method of calculating distances as a measure of atypicality has been shown to be unreliable due to its susceptibility to image complexity [4, 22, 35].

Alternatively, an image's typicality can be quantified by using the model likelihood over training data space. Grathwhol *et al.* [12] argue that typical images are localized around regions of increased mass in the probability distribution. The high probability density area at the mode of the Gaussian annulus has a considerably smaller volume than the typical set region. This density-volume relationship results in the typical set residing at neighborhoods of high probability mass. Conversely, atypical data points reside in sparsely populated areas with consequently less uniform distribution of mass. The fully tractable nature of NFs present an excellent opportunity to estimate the mass distribution by leveraging the gradients w.r.t. the input. This is also known as the *gradient score*, denoted as $\nabla_{\mathbf{x}} \log p(\mathbf{x})$.

3 Methods

Despite criticisms of typicality-based approaches [35], NFs demonstrated success on benchmark datasets when evaluated with typicality rather than likelihoods, particularly when using the gradient score [3, 12]. Anomalous samples are expected to have higher gradients w.r.t. the input image. Thus, we mark samples as OOD if the gradient score exceeds the gradient score of the ID data. Similar to Chali *et al.* [3], we appropriately adjust the minimization objective to penalize

high gradients for training data by specifying:

$$\mathcal{L} = -\log p(\mathbf{x}) + \alpha \cdot ||\nabla_{\mathbf{x}} \log p(\mathbf{x})||_2 \ , \tag{4}$$

with hyperparameter α to further encourage accurate convergence.

The four datasets employed in this study to test the proposed approach encompass a diverse range of medical imaging modalities and conditions. The (1) ISIC Melanoma Dataset [6] which is utilized for skin lesion analysis, comprises 2,000 benign and 1,500 malignant RGB images. We use two datasets for detecting pathologies in Chest X-ray (CXR) images: the (2) COVID-19 Dataset [5,26] with 10,000 healthy and 4,000 COVID-affected images, and the (3) Pneumonia Dataset [28] with 10,000 healthy and 6,000 pneumonia-affected images (the aforementioned numbers are rounded). Additionally, for neurological examinations the (4) HeadCT dataset [18] is incorporated, consisting of 100 normal and 100 hemorrhage-affected head-computed tomography (CT) images. Each dataset is uniformly downscaled to contain 128×128-pixel images and are standardized with statistics of the training set.

The proposed method* uses `PyTorch` to train GLOW [16], utilizing a multi-scale setup with depth of flow $K = 32$ and the number of levels $L = 3$. To determine the gradient score, we apply attribute `requires_grad=True` on the input tensor. The obtained gradients are flattened and the L_2 norm is applied to each batch of images. Finally, the norms are averaged across all batches to obtain the score approximate. The hyperparameter $\alpha = 2$ is chosen to yield gradients of comparable magnitudes to those appearing with the log-likelihood. During testing, the scores are obtained per data instance. Similarly, we have found that computing the score may exhibit training instability depending on the model implementation [12]. However, in our experiments it is found that suitable data standardization already addresses this issue.

In all of our experiments, we train different GLOW models on benign images from the training set and use the benign validation set to select the best model. Finally, we test the selected models on the official test set, which contains both benign and malignant cases, if available. The models are evaluated using the (1) standard log-likelihood (LL) method, the (2) Typicality Test in Latent space (TTL) [24] and the (3) proposed gradient score-based typicality test. Implementation (1) and (2) are trained minimizing the standard negative log-likelihood, while (3) is trained with the proposed penalized loss.

4 Results and Discussion

Quantitative evaluation: The empirical distributions of likelihoods and gradient scores, obtained by the baseline and proposed method respectively, are visualized in Figure 1. It can be observed that the baselines based on likelihood are severely limited in their ability to distinguish ID and OOD data, which is indicated by the substantial overlap of the empirical distributions, where in some

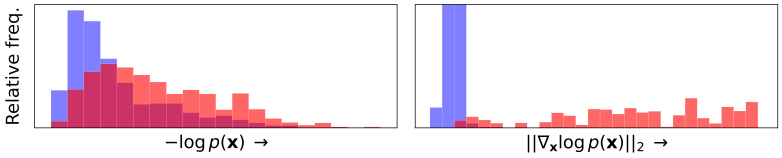
*Code available at: `https://github.com/lemarabd/typicality-MOOD`

Table 1: AUROC results of the proposed method compared to SOTA models. Separate indication for semi-supervised models (†) and re-implementation (∗).

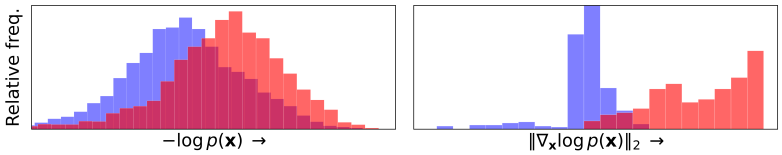
Model	Dataset	AUROC ↑
AE-FLOW [37]	ISIC	0.878 ± - - -
GLOW (LL)*		0.788 ± 0.041
GLOW (TTL)*		0.675 ± 0.039
Proposed method		**0.950 ± 0.059**
MorphAEus [1]	COVID-19	0.860 ± 0.070
GLOW (LL)*		0.654 ± 0.027
GLOW (TTL)*		0.648 ± 0.007
Proposed method		**0.937 ± 0.080**
MorphAEus [1]	Pneumonia	0.836 ± 0.080
GLOW (LL)*		0.577 ± 0.025
GLOW (TTL)*		0.581 ± 0.032
Proposed method		**0.904 ± 0.056**
DevNet† [25]	HeadCT	**0.982 ± 0.009**
GLOW (LL)*		0.538 ± 0.052
GLOW (TTL)*		0.494 ± 0.055
Proposed method		0.922 ± 0.086

instances, the model erroneously assigns higher likelihood to OOD data. With high contrast, the proposed approach (right column) is significantly better in distinguishing the individual histograms across all datasets. We present this quantitative improvement with the AUROC scores, averaged across five random seeds, in Table 1. With the exception of the HeadCT dataset, the proposed methodology significantly improves upon the dataset-specific SOTA methods, LL-based GLOW evaluation and evaluation using TTL. Nonetheless, the HeadCT SOTA method is a semi-supervised approach.

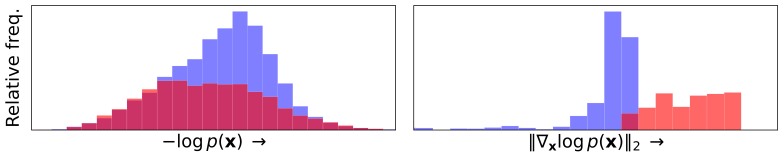
Qualitative evaluation: The most likely images are usually not representative for the majority of the dataset. In theory, typicality should be more accurate in reflecting the diversity of the training distribution. This is verified for the ISIC datasets in Figure 2a, where the diversity of likely samples are minimal. The most likely samples exhibit consistent characteristics such as smooth skin texture, uniform pigmentation, and lesion size. In contrast, the typical images in Figure 2c are much more varied, thereby accurately reflecting the dataset. While the likely samples are most certainly benign, the visualization exposes the intrinsic bias of the model towards simple textural content. This is especially emphasized in Figure 2b. It is quite clear that the likelihood-based model is biased towards assigning low likelihoods, due to extreme textures such as hairs, which is a limitation elucidated in previous work [30]. The proposed methodology does not suffer from this and correctly classifies malignant samples, regardless of their



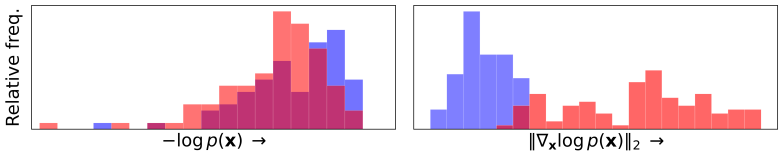
(a) ISIC



(b) COVID-19

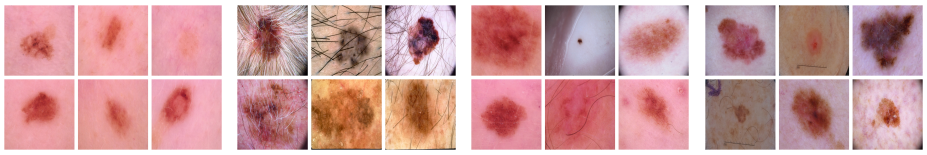


(c) Pneumonia



(d) HeadCT

Fig. 1: Negative log-likelihood histograms obtained by the vanilla GLOW model (left figure of each subfigure) against the gradient score histograms obtained using the proposed method (right figure of each subfigure). Training ID data is depicted in blue and the OOD data in red.



(a) Most likely (b) Least likely (c) Most typical (d) Least typical

Fig. 2: Comparison of most typical and most likely samples versus least likely and least typical samples. Most likely samples seem to visualize standard cases, while typical samples have a broader variation.

graphical contents, as seen in Figure 2d. Notably, it can be observed that the presence of hairs on benign images does not inhibit OOD detection.

Data imbalance: Balanced class data-distributions, usually employed during model development, often do not reflect the real clinical incidence rate, leading to a large number of false positives predictions when deployed [20]. To demonstrate the impact of severe data imbalance on the performance of supervised models in contrast to the proposed method, we have conducted several ablation experiments. Each supervised model [13,29,33] is independently trained on varying ratios, starting with a 1:1 ratio of normal to abnormal data, then with decreasing abnormal data. Each unsupervised model, trained exclusively on normal data, uses progressively smaller subsets of normal data. It should be noted that the HeadCT dataset has been excluded in these experiments because of the relatively limited number of samples. The supervised models are pre-trained on ImageNet1K [8], following the same training settings as in [13], with smaller batch sizes for increasing data ratios. The models are plotted against performance for each data ratio, as can be seen in Figure 3. While supervised models tend to achieve high AUROC scores under balanced conditions, our analysis reveals a clear decline in performance when subjected to imbalanced data. However, the proposed method is consistent in performance even with 0.1% of the dataset. Additionally, unlike the GLOW baseline, the proposed method demonstrates performance akin to supervised models. The excessive susceptibility of supervised models underscores the practical value of our contributions in diverse clinical scenarios. Firstly, data imbalance has no impact on the proposed model and secondly, the model continues to exhibit SOTA performance, even under scarcity of training samples.

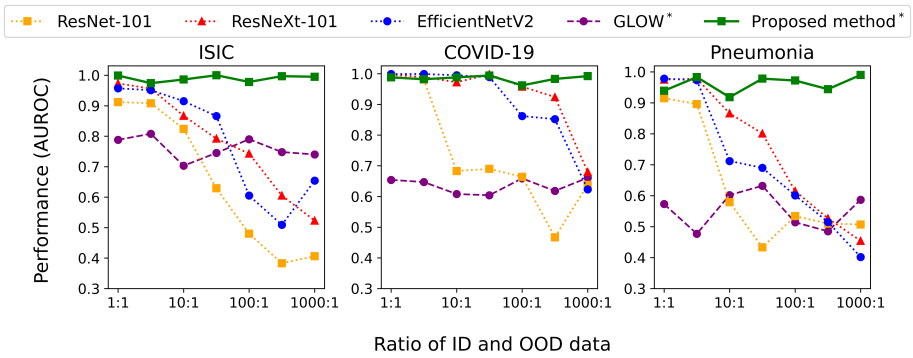


Fig. 3: Impact of data imbalance on the performance of unsupervised methods, indicated with (∗), and various supervised methods.

5 Conclusion

This study has explored the effectiveness of leveraging typicality for OOD detection in medical imaging. The obtained results demonstrate that ID medical images are more typical rather than more likely. The proposed method shows competitive performance even against supervised models, while exceeding them under significant data imbalances. Furthermore, the proposed model achieves SOTA in unsupervised pathology detection across different medical imaging modalities. The proposed method improves AUROC scores by 10%, 15%, and 14% for ISIC, COVID-19, and Pneumonia, respectively, compared to the SOTA models for each dataset. While our focus lies on image-level semantic OOD detection, we advocate for future research to explore the utility of typicality for other OOD sub-tasks in the medical domain, such as sensory anomaly detection.

References

1. Bercea, C.I., Rueckert, D., Schnabel, J.A.: What do aes learn? challenging common assumptions in unsupervised anomaly detection. In: MICCAI. pp. 304–314. Springer (2023)
2. Caterini, A.L., Loaiza-Ganem, G.: Entropic Issues in Likelihood-Based OOD Detection. In: I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021. pp. 21–26. PMLR (Feb 2022), iSSN: 2640-3498
3. Chali, S., Kucher, I., Duranton, M., Klein, J.O.: Improving Normalizing Flows with the Approximate Mass for Out-of-Distribution Detection. In: CVPRW. pp. 750–758. IEEE, Vancouver, BC, Canada (Jun 2023)
4. Choi, H., Jang, E., Alemi, A.A.: WAIC, but Why? Generative Ensembles for Robust Anomaly Detection (May 2019), arXiv:1810.01392 [cs, stat]
5. 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)
6. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 isbi, hosted by the isic. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
7. Cremer, C., Li, X., Duvenaud, D.: Inference suboptimality in variational autoencoders. In: ICML. pp. 1078–1086. PMLR (2018)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
9. 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)
10. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. arXiv preprint arXiv:1605.08803 (2016)
11. 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. 2947–2956 (2023)
12. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Ho, J., Chen, X., Srinivas, A., Duan, Y., Abbeel, P.: Flow++: Improving flow-based generative models with variational dequantization and architecture design. In: ICML. pp. 2722–2730. PMLR (2019)
15. Jeong, H., Byun, H., Kang, D.U., Lee, J.: Blindharmony. In: ICCV. pp. 21072–21082. IEEE Computer Society (2023)
16. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. NIPS **31** (2018)
17. 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)

18. Kitamura, F.C.: Head ct - hemorrhage (2018). https://doi.org/10.34740/KAGGLE/DSV/152137, https://www.kaggle.com/dsv/152137
19. 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: CVPR. pp. 11578–11589 (2023)
20. 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)
21. Nakao, T.e.a.: Unsupervised deep anomaly detection in chest radiographs. Journal of Digital Imaging **34**, 418–427 (2021)
22. Nalisnick, E., Matsukawa, A., Teh, Y.W., Lakshminarayanan, B.: Detecting out-of-distribution inputs to deep generative models using typicality. arXiv preprint arXiv:1906.02994 (2019)
23. 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)
24. Osada, G., Takahashi, T., Nishide, T.: Understanding likelihood of normalizing flow and image complexity through the lens of out-of-distribution detection. Proceedings of the AAAI Conference on Artificial Intelligence **38**(19), 21492–21500 (Mar 2024). https://doi.org/10.1609/aaai.v38i19.30146
25. Pang, G., Ding, C., Shen, C., Hengel, A.v.d.: Explainable deep few-shot anomaly detection with deviation networks. arXiv preprint arXiv:2108.00462 (2021)
26. 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)
27. Ren, J.e.a.: Likelihood ratios for out-of-distribution detection. NIPS **32** (2019)
28. Shih, G.e.a.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
29. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: ICML. pp. 10096–10106. PMLR (2021)
30. 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)
31. Vershynin, R.: High-dimensional probability: An introduction with applications in data science, vol. 47. Cambridge university press (2018)
32. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: MICCAI. pp. 35–45. Springer (2022)
33. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. pp. 1492–1500 (2017)
34. Zhang, J.e.a.: Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE transactions on medical imaging **40**(3), 879–890 (2020)
35. Zhang, L., Goldstein, M., Ranganath, R.: Understanding failures in out-of-distribution detection with deep generative models. In: International Conference on Machine Learning. pp. 12427–12436. PMLR (2021)
36. Zhao, S., Song, J., Ermon, S.: Infovae: Information maximizing variational autoencoders. arXiv preprint arXiv:1706.02262 (2017)
37. Zhao, Y., Ding, Q., Zhang, X.: Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection. In: The Eleventh ICLR (2022)
38. Zhou, Y.: Rethinking reconstruction autoencoder-based out-of-distribution detection. In: CVPR. pp. 7379–7387 (2022)