

MaRS: Robust Out-of-Distribution Detection via Mahalanobis Residual Scoring

Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig

xAILab Bamberg, University of Bamberg, Germany
`francesco.di-salvo@uni-bamberg.de`

Abstract. Foundation models provide highly descriptive representations for medical images, yet their reliability degrades under distribution shifts arising from changes in patients, devices, or acquisition conditions. Reliable out-of-distribution (OOD) detection is therefore essential for safe deployment. Recent post-hoc detectors efficiently exploit frozen embeddings (*e.g.*, kNN), whereas reconstruction-based OOD detection in latent feature space has seen limited adoption due to inconsistent performance. In this work, we show that the limitation of reconstruction-based methods in latent space does not stem from poor reconstruction quality, but from how reconstruction errors are scored. Standard L_2 residual norms collapse the anisotropic residual structure, thereby suppressing informative deviations. To address this limitation, we introduce **MaRS** (Mahalanobis Residual Scoring), a label-free OOD detector that learns an in-distribution manifold using a lightweight autoencoder and measures deviation via a Mahalanobis distance on reconstruction residuals, yielding variance-aware OOD scores. Across three imaging modalities, multiple types of distribution shift, and different model families and scales, **MaRS** outperforms established confidence-, distance-, and reconstruction-based baselines, while remaining fully post-hoc and lightweight. The code is available at github.com/francescodisalvo05/mars.

Keywords: OOD detection · Autoencoder · Residual · Mahalanobis.

1 Introduction

Deep learning has achieved remarkable performance across a wide range of medical imaging tasks. However, despite this progress, models remain vulnerable to distribution shifts caused by changes in patient populations, imaging devices, acquisition protocols, or artifacts [10,31]. Such shifts can silently degrade performance and compromise clinical reliability. Out-of-distribution (OOD) detection [30,7] therefore plays a central role in ensuring reliable and safe use of AI systems in medicine. Existing OOD detection methods differ primarily in how deviation from the in-distribution is measured. *Confidence-based* methods [6,13] use classifier uncertainty as an OOD signal, but can be fragile in medical settings, where class imbalance, weak supervision, and miscalibration can reduce the reliability of classifier confidences. *Distance-based* methods, including Deep kNN [23] and

Mahalanobis-based approaches [11,2,16], measure deviation in feature space and benefit from the rise of foundation models [4,18,22,28,9]. However, their performance can degrade in high-dimensional embeddings, where nearest-neighbor distances become less informative and covariance estimates become harder to regularize reliably. *Reconstruction-* and *subspace-based* methods offer a label-free alternative by learning structure from in-distribution features and flagging samples with large reconstruction or projection error [15,12]. The Residual baseline [17], for instance, models OOD as deviation from an in-distribution subspace (via PCA or kernel PCA), while ViM [26] combines such linear residuals with classifier logits and thus relies on a well-defined classification head. Critically, Residual PCA [17] aggregates residuals with a uniform L_2 norm, implicitly assuming isotropic residual structure. In latent space, however, residuals can be strongly anisotropic. Consequently, ignoring their covariance structure obscures the distinction between expected reconstruction noise and OOD deviations.

In this work, we show that a key limitation of reconstruction- and residual-based OOD detection lies not in reconstruction quality, but in how reconstruction residuals are scored. We introduce **MaRS** (*Mahalanobis Residual Scoring*), a lightweight and fully post-hoc OOD detector that applies a covariance-aware Mahalanobis distance to autoencoder reconstruction residuals. By explicitly modeling anisotropy in the residual distribution, **MaRS** avoids reliance on class labels, classifier heads, or predefined subspace assumptions, while generalizing prior residual-based approaches. Taking advantage of the strong representational power of domain-agnostic foundation models [21,14], **MaRS** operates directly on frozen backbone features, ensuring the method is both computationally efficient and adaptable to various imaging modalities. This design further eliminates any reliance on labeled data for backbone fine-tuning, thereby enabling a label-free OOD detection framework.

Extensive experiments across three imaging modalities, multiple types of distribution shift, two backbone families, and different model scales demonstrate that **MaRS** consistently outperforms confidence-, distance-, and reconstruction-based baselines. Our results highlight the effectiveness of variance-aware residual scoring as a simple yet powerful principle for post-hoc OOD detection on frozen representations. Our contributions are as follows:

- We introduce **MaRS** (see Figure 1), a plug-and-play, label-free OOD detection method that applies Mahalanobis distance to autoencoder reconstruction residuals, consistently outperforming strong baselines across modalities, shift types, backbone families, and model sizes.
- We identify residual anisotropy as a key but overlooked signal for OOD detection, and show that variance-aware Mahalanobis scoring in residual space substantially outperforms uniform L_2 aggregation, supported by targeted spectral analyses and ablations.
- We demonstrate that **MaRS** remains robust across backbone architectures and scales, and benefits from operating on pre-normalization features, where the intrinsic anisotropic variance structure exploited by residual-space Mahalanobis scoring is preserved.

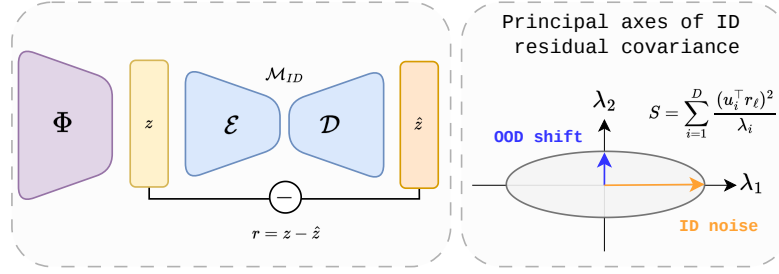


Fig. 1. **Left:** Given a frozen backbone Φ producing features z , an autoencoder $(\mathcal{E}, \mathcal{D})$ learns a projection onto the in-distribution (ID) manifold $\mathcal{M}_{\text{ID}}$, and the reconstruction residual $r = z - \hat{z}$ captures deviations from it. **Right:** Principal components of the ID residual covariance reveal that out-of-distribution (OOD) deviations concentrate in low-variance directions, which are amplified by variance-aware Mahalanobis scoring.

2 Method

Foundation models provide semantically rich representations but remain vulnerable to distribution shifts [24]. We consider the setting of label-free, post-hoc OOD detection on frozen features and introduce MaRS, a lightweight adapter that estimates deviation from the in-distribution (ID) feature manifold.

2.1 Autoencoder as a projection onto the ID manifold

Given a frozen backbone Φ , MaRS operates on features extracted from a fixed layer. Let $z = \Phi(x) \in \mathbb{R}^D$ denote the feature representation of an input x , where D is the latent dimension. An autoencoder $(\mathcal{E}, \mathcal{D})$ produces a reconstruction $\hat{z} = \mathcal{D}(\mathcal{E}(z))$ and is trained on ID features by minimizing

$$\mathcal{L}_{\text{AE}} = \|z - \hat{z}\|_2^2. \quad (1)$$

Geometrically, the autoencoder implements a data-driven projection

$$\Pi_{\mathcal{M}_{\text{ID}}}(z) = \mathcal{D}(\mathcal{E}(z)) = \hat{z} \quad (2)$$

onto the ID manifold $\mathcal{M}_{\text{ID}}$. ID samples lie close to this manifold and are reconstructed accurately, whereas OOD samples are mapped back toward $\mathcal{M}_{\text{ID}}$, inducing structured reconstruction errors that reflect deviation from the learned manifold.

2.2 Residual space as a drift descriptor

We define the reconstruction residual as $r(x) = z - \hat{z} \in \mathbb{R}^D$. Intuitively, $r(x)$ measures how much the feature representation must be “pulled back” to lie on the ID manifold. For ID samples, residuals are small and primarily reflect reconstruction imperfections. For OOD samples, residuals exhibit systematic components that are rare or absent under the ID residual distribution, encoding informative drift signals.

2.3 Mahalanobis scoring on residuals

A standard reconstruction score utilizes a uniform L_2 norm, $\|r\|_2^2 = \sum_{i=1}^D r_i^2$, which treats all residual directions equally. This is suboptimal when ID residuals are anisotropic: some directions correspond to high-variance reconstruction noise, while others are highly stable and thus more sensitive to distributional shifts. We capture this structure by estimating the ID residual covariance

$$\Sigma = \text{Cov}(r(x) \mid x \sim \text{ID}), \quad (3)$$

and by defining the **MaRS** score as a Mahalanobis distance in residual space:

$$S_{\text{MaRS}}(x) = r(x)^\top \Sigma^{-1} r(x). \quad (4)$$

Let $\Sigma = U\Lambda U^\top$ be the eigendecomposition of the residual covariance, where $U = [u_1, \dots, u_D]$ contains the principal directions and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_D)$ the corresponding eigenvalues. The score can then be written as

$$S_{\text{MaRS}}(x) = \sum_{i=1}^D \frac{(u_i^\top r(x))^2}{\lambda_i}, \quad (5)$$

which amplifies deviations along low-variance residual directions via λ_i^{-1} , while high-variance directions are downweighted. Unlike Mahalanobis scoring in embedding space based on class-conditional means [16], our formulation is label-free, does not assume well-separated class clusters, and relies on a single global covariance. Consequently, it avoids computing distances to all class centroids and scales independently of the number of classes. Subtracting the learned projection, $r = z - \Pi_{\mathcal{M}_{\text{ID}}}(z)$, suppresses class-semantic variation and isolates deviations orthogonal to the ID manifold, making a single global covariance Σ sufficient while avoiding reliance on potentially noisy class labels. OOD detection is performed by thresholding $S_{\text{MaRS}}(x)$ on ID validation data (*e.g.*, at the 95th percentile).

3 Experimental results

3.1 OOD Detection

Methods We compare **MaRS** against a diverse set of established baselines. Confidence-based methods include MSP [6], ODIN [13], and ViM [26]. Distance-based baselines include Deep kNN [23], one-class SVM (OCSVM) [20], and Mahalanobis++ (Maha) [16], all operating directly in latent feature space. Notably, the latter exhibits recent state-of-the-art performance [16], relying on class-conditional Gaussian modeling and feature normalization. We further include *Residual*, which measures the deviation from a low-variance linear subspace estimated via PCA, following [17]. Finally, we evaluate reconstruction-based methods using a deterministic autoencoder (AE), where OOD scores are computed as the L_2 norm of reconstruction residuals [17]. **MaRS** builds upon this reconstruction framework by replacing the uniform L_2 with a variance-aware Mahalanobis scoring of residuals.

Datasets We evaluate MaRS on multiple medical imaging datasets covering diverse modalities and shift types. First, we include the established MIDOG histopathology benchmark [3] (CC BY 4.0), using the splits of OpenMIBOOD [5]. ID data consist of mitotic and non-mitotic cell crops extracted from H&E-stained whole-slide images. Covariate shifts (CS) arise from changes in imaging devices, while near-OOD datasets introduce semantic shifts across cancer types, species, and acquisition settings. Far-OOD samples originate from unrelated medical applications, including cervical cancer cell images [1] (Apache 2.0) and breast FNAC cytology [19] (license not disclosed). We further consider two additional imaging modalities. For chest X-ray imaging, adult chest X-rays [27] are treated as ID data, while pediatric chest X-rays [8] serve as covariate-shifted data. For dermatoscopy, we use the HAM10000 dataset [25], treating non-melanocytic skin lesions as in-distribution (ID) data and melanocytic lesions as out-of-distribution (OOD) samples. This choice induces a low-data, imbalanced ID regime, with training restricted to a small and heterogeneous lesion set ($N=1,535$). For dermatoscopy (CC BY-NC 4.0) and chest X-ray (CC BY 4.0), we adopt the data collection of Yang et al. [29], using the training split as ID data and the corresponding test splits for evaluation. Due to the limited size of the pediatric chest X-ray test set, its training split is used as OOD data.

Implementation details We adopt off-the-shelf backbones and avoid any reliance on labeled data for backbone fine-tuning. All methods operate on last-layer embeddings extracted from a supervised ViT/S backbone [4]. MSP, ODIN, and ViM train a lightweight linear head for 50 epochs using SGD with a learning rate of 10^{-3} . ODIN and Deep kNN use the default hyperparameters of [32]. We additionally include Mahalanobis++ [16], which applies feature normalization prior to distance computation. Reconstruction-based methods (AE, MaRS) employ a two-layer MLP encoder and decoder (hidden ratio 0.5, bottleneck dimension 128) trained for 50 epochs with SGD (learning rate 10^{-3}). Unless stated otherwise, results are reported as mean AUROC over three runs. While latent normalization is a prerequisite for standard distance-based baselines to prevent dimensional dominance, it collapses the anisotropic variance that characterizes meaningful OOD signals. MaRS preserves this signal by operating on pre-normalization features, using the Mahalanobis metric to statistically whiten the residual space.

Results As shown in Table 1, confidence-based detectors (MSP, ODIN, ViM) underperform on MIDOG and X-Ray, while ViM is the second-best method on dermatoscopy. Among label-free latent-space methods, Residual and Deep kNN achieve competitive performance, confirming the utility of residual-based signals. However, extending reconstruction to a nonlinear autoencoder and scoring via a uniform L_2 norm leads to a clear performance drop. Mahalanobis++ ranks second overall, but relies on class-conditional statistics and exhibits reduced robustness in the low-data dermatoscopy setting. In contrast, MaRS significantly outperforms Mahalanobis++ across 3 runs and 5 dataset splits, in both AUROC and FPR@95 (paired Wilcoxon signed-rank test, $p < 0.05$), while remaining fully label-free. As Mahalanobis++ is deterministic, its results are aligned with the three-seed evaluation protocol of MaRS to enable paired statistical testing.

Table 1. OOD detection performance measured by average AUROC ($\uparrow$). MIDOG results are reported under covariate (CS), near-OOD, and far-OOD shifts. Chest X-ray and dermatoscopy correspond to covariate and semantic shifts, respectively. The Average AUROC ($\uparrow$) column reports the mean AUROC, and the Average FPR@95 ($\downarrow$) column reports the mean false positive rate at 95% true positive rate, both averaged across all benchmarks. Results are further averaged over three seed runs, and the **best two methods** per column are bolded. Overall, **MaRS** significantly outperforms Mahalanobis++ ($\dagger$, paired Wilcoxon signed-rank test, $p < 0.05$).

<i>Unsup</i>		MIDOG			X-Ray	Derma	Average	
		CS	near-OOD	far-OOD	CS	Semantic	AUROC $\uparrow$	FPR@95 $\downarrow$
$\times$	MSP	46.71	48.93	62.73	71.63	48.99	55.80	88.02
$\times$	ODIN	45.53	48.98	63.44	71.63	55.20	56.96	87.33
$\times$	ViM	45.84	47.02	76.27	59.26	58.43	57.36	81.05
$\checkmark$	Residual	92.09	82.92	100.00	92.90	57.25	85.03	45.31
$\checkmark$	OCSVM	87.68	81.21	99.89	91.80	49.27	81.97	47.95
$\checkmark$	Deep kNN	90.92	80.10	99.95	90.55	57.42	83.79	47.88
$\checkmark$	AE	86.48	77.80	99.87	92.73	50.90	81.56	50.72
$\times$	Maha	92.56	83.80	99.99	93.09	57.24	85.34	43.59
$\checkmark$	MaRS	93.43	83.43	99.98	94.00	59.24	86.01†	42.72†

3.2 A closer look at the residual space

This ablation investigates *why* Mahalanobis scoring in residual space outperforms a uniform L_2 norm. We estimate the residual covariance on the MIDOG training set, evaluate in-distribution (ID) behavior on the MIDOG test set, and use FNAC2019 as an OOD dataset. We analyze the principal components (PCs) of the ID residual covariance and measure the mean squared residual deviation $\mathbb{E}_x[(u_i^\top r)^2]$ of ID and OOD samples along each principal axis u_i .

Results Figure 2 reveals two key findings. (i) The residual covariance is highly anisotropic: over 90% of the variance is concentrated in fewer than one-third of the principal components, with the remaining components exhibiting very low variance (small eigenvalues). (ii) When projecting ID and OOD residuals onto these axes, the relative separation between the OOD deviation (blue line) and the ID variance (orange line) is predominantly observed in the low-variance directions. Quantitatively, for each PC u_i , we define a normalized residual energy ratio $\rho_i = \mathbb{E}_{\text{OOD}}[(u_i^\top r)^2] / \mathbb{E}_{\text{ID}}[(u_i^\top r)^2]$. Averaging ρ_i over high-variance components (explaining the first 90% of ID variance) and low-variance components separately, we observe a $1.64\times$ larger ratio in the low-variance subspace, indicating that OOD deviations concentrate along stable residual directions. To exploit this property, we employ Mahalanobis scoring, which reweights each PC by the inverse of its variance (λ_i^{-1}). This ensures that OOD deviations along the low-variance directions are most heavily penalized (since λ_i^{-1} is large), thereby maximizing the separation between the ID residual noise and the OOD signal.

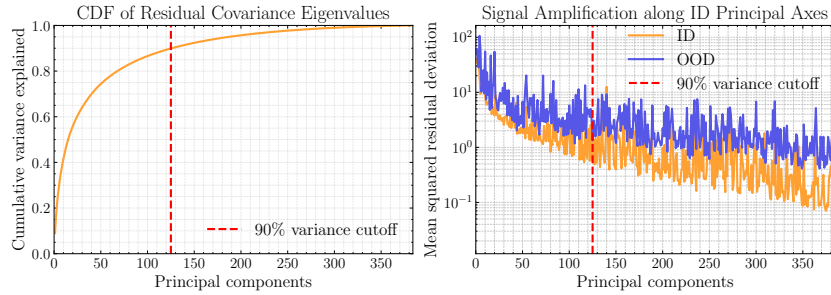


Fig. 2. Analysis of residual space for MIDOG+FNAC. **Left:** CDF of eigenvalues of the ID residual covariance, indicating a highly anisotropic residual distribution. **Right:** Mean squared residual deviation $\mathbb{E}_x[(u_i^\top r)^2]$ of ID and OOD samples along each principal component u_i . OOD residuals exhibit larger and more persistent deviations precisely in the low-variance directions, motivating variance-aware Mahalanobis scoring.

3.3 Backbone size and normalization sensitivity

Backbone size We assess the generalizability of MaRS across supervised (ViT [4]) and self-supervised (DINOv3 [22]) models at small (S) and base (B) scales. This evaluates the detector’s robustness to the increased feature dimensionality ($D = 384$ vs. 768) and thereby induced data sparsity. Using Mahalanobis++ as a competitive latent-space baseline, we isolate whether scoring residuals remains superior as representational complexity grows. We focus this analysis on chest X-ray (covariate shift) and dermatoscopy (semantic shift).

Results As shown in Figure 3 (left), MaRS outperforms Mahalanobis++ across almost all configurations. These improvements are consistent across both supervised (ViT) and self-supervised (DINOv3) models, indicating that the benefits of residual-space scoring are largely independent of the pretraining paradigm. Crucially, while Mahalanobis++ generally exhibits a notable performance decay when scaling from small to base models, likely due to the unreliable estimation of high-dimensional global covariance, MaRS remains remarkably robust. This suggests that reconstruction residuals act as an implicit filter, suppressing high-dimensional variability while preserving directions most sensitive to distributional shifts, consistent with the anisotropic structure of the residual space.

Pre- vs. post-normalization We study the impact of the backbone’s final LayerNorm on OOD detection performance across all datasets (average AUROC). We evaluate all methods under both *pre-normalization* and *post-normalization* features. Latent-space methods such as Mahalanobis++ and PCA-based residual baselines are conventionally applied to normalized features, as normalization stabilizes distances and improves numerical conditioning. In contrast, MaRS benefits from un-normalized features, preserving variance anisotropy.

Results Figure 3 (right) reports average AUROC across datasets for the top-performing methods under pre- and post-normalization. As expected, while normalization improves the performance of latent-space baselines, it degrades **MaRS**. This behavior is consistent with our analysis: normalization flattens the variance spectrum and suppresses the low-variance directions where OOD deviations are most informative. By preserving this anisotropy, **MaRS** benefits from pre-normalized features, outperforming reference methods. In contrast, Mahalanobis++ estimates covariance directly in embedding space, where dominant variance directions are more strongly tied to semantic and class-dependent variation than to OOD-specific residual structure.

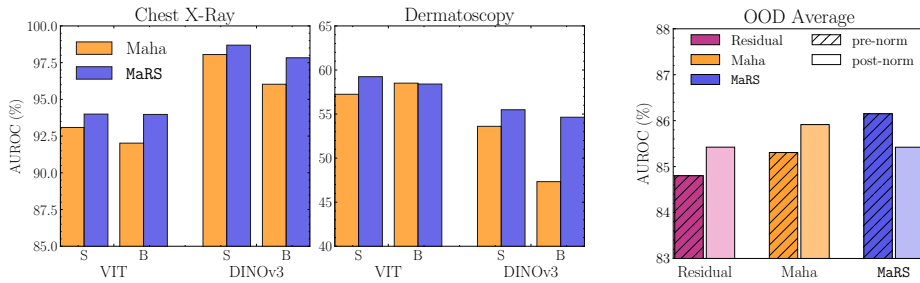


Fig. 3. **Left:** AUROC on chest X-ray and dermatoscopy for ViT and DINOv3 backbones at small (S) and base (B) scales. **MaRS** consistently outperforms Mahalanobis++ and is less sensitive to increased feature dimensionality. **Right:** Average AUROC of the top three methods under pre- and post-normalization. While normalization improves latent-space baselines, it suppresses variance anisotropy and degrades **MaRS**, confirming that residual-based Mahalanobis scoring benefits from pre-normalization features.

4 Conclusions

Limitations Although **MaRS** is highly effective for the subtle covariate and semantic shifts commonly observed in medical image analysis, its performance is inherently tied to the quality of the underlying manifold estimation. Future work will explore whether specialized medical backbones, rather than domain-agnostic ones, further refine the residual space anisotropy. Additionally, while we focus on single-modality OOD, extending **MaRS** to multimodal data presents a promising direction.

Discussion Our results show that a central limitation of existing reconstruction- and residual-based OOD detectors lies in the implicit assumption of isotropic reconstruction errors. **MaRS** addresses this limitation by replacing uniform L_2 aggregation with covariance-aware Mahalanobis scoring in residual space.

Importantly, the gain does not stem from Mahalanobis scoring alone, but from estimating covariance on reconstruction residuals rather than raw embeddings. After subtracting the learned ID projection, OOD information concentrates in anisotropic low-variance residual directions, making covariance-aware scoring substantially more effective. This interpretation is supported by our analysis, which shows that OOD deviations are most pronounced along low-variance residual directions that are largely ignored by standard L_2 norms but explicitly amplified by the Mahalanobis metric. As a result, MaRS consistently outperforms latent-space and reconstruction-based baselines, while remaining label-free, post-hoc, and robust to high-dimensional feature representations and data sparsity.

Acknowledgments. This study was funded through the Hightech Agenda Bayern (HTA) of the Free State of Bavaria, Germany.

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

References

1. Amorim, J.G.A., Macarini, L.A.B., Matias, A.V., Cerentini, A., Onofre, F.B.D.M., Onofre, A.S.C., Von Wangenheim, A.: A novel approach on segmentation of agnored-stained cytology images using deep learning. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE (2020)
2. Anthony, H., Kamnitsas, K.: On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. pp. 136–146. Springer (2023)
3. Aubreville, M., Wilm, F., Stathonikos, N., Breininger, K., Donovan, T.A., Jabari, S., Veta, M., Ganz, J., Ammeling, J., van Diest, P.J., et al.: A comprehensive multi-domain dataset for mitotic figure detection. *Scientific data* **10**(1), 484 (2023)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021)
5. Gutbrod, M., Rauber, D., Nunes, D.W., Palm, C.: Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25874–25886 (2025)
6. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
7. 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:2404.18279* (2024)
8. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9 (2018)

9. Koch, V., Wagner, S.J., Kazemina, S., Sancar, E., Hehr, M., Schnabel, J.A., Peng, T., Marr, C.: Dinobloom: a foundation model for generalizable cell embeddings in hematology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 520–530. Springer (2024)
10. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al.: Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine* **30**(4) (2024)
11. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NIPS* (2018)
12. Li, X., Lu, Y., Desrosiers, C., Liu, X.: Out-of-distribution detection for skin lesion images with deep isolation forest. In: International Workshop on Machine Learning in Medical Imaging. pp. 91–100. Springer (2020)
13. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018)
14. Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Zhang, Y., Li, J., et al.: Does dinov3 set a new medical vision standard? *arXiv preprint arXiv:2509.06467* (2025)
15. Lu, Y., Xu, P.: Anomaly detection for skin disease images using variational autoencoder. *arXiv:1807.01349* (2018)
16. Müller, M., Hein, M.: Mahalanobis++: Improving OOD detection via feature normalization. In: Forty-second International Conference on Machine Learning (2025)
17. Ndiour, I., Ahuja, N., Tickoo, O.: Out-of-distribution detection with subspace techniques and probabilistic modeling of features. *arXiv:2012.04250* (2020)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*
19. Saikia, A.R., Bora, K., Mahanta, L.B., Das, A.K.: Comparative assessment of cnn architectures for classification of breast fnac images. *Tissue and Cell* **57** (2019)
20. Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. *Advances in neural information processing systems* **12** (1999)
21. Schulthess, N., Konukoglu, E.: Anomaly detection by clustering dino embeddings using a dirichlet process mixture. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2025)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025)
23. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International conference on machine learning. PMLR (2022)
24. Teterwak, P., Saito, K., Tsiligkaridis, T., Plummer, B.A., Saenko, K.: Is large-scale pretraining the secret to good domain generalization? In: The Thirteenth International Conference on Learning Representations (2025)
25. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**(1), 1–9 (2018)
26. 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)

27. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3462–3471 (2017)
28. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**(8015), 181–188 (2024)
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 Neural Information Processing Systems* **35** (2022)
31. Yoon, J.S., Oh, K., Shin, Y., Mazurowski, M.A., Suk, H.I.: Domain generalization for medical image analysis: A review. *Proceedings of the IEEE* (2024)
32. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y., Li, H.: OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research* (2024), dataset Certification