

Group Equivariant Diffusion for Anomaly Detection in Computational Cytology

Swarnadip Chatterjee¹, Ssharvien Kumar Sivakumar^{2,3}, and Anirban Mukhopadhyay²

¹ Uppsala University, Uppsala, Sweden
swarnadip.chatterjee@it.uu.se

² Technical University of Darmstadt, Darmstadt, Germany

³ Carl Zeiss AG, Munich, Germany

Abstract. Computational cytology on whole-slide images is challenging because malignant cells are rare, heterogeneous, and annotated slides are scarce. Anomaly detection frameworks can be trained on normal slide-negative patches and then applied at test time to flag abnormal patches in held-out slides. Most unsupervised anomaly detection approaches including generative ones (GAN-based and diffusion-based), are tuned to organ-level imaging and require large curated datasets. In cytology the signal is cell-centric: rotating or flipping a single-cell patch does not change its diagnostic class, yet standard diffusion models treat transformed views as distinct inputs, leading to transformation-dependent reconstructions and unstable anomaly scores. We propose a D_4 -equivariant diffusion framework that enforces rotation and reflection symmetry both architecturally, via a D_4 -equivariant U-Net, and at inference, via equivariant noise coupling and (optionally) frame averaging. This alignment with biological invariance yields transformation-consistent pseudo-healthy reconstructions and more stable anomaly ranking under symmetry. On two publicly available cytology datasets of bone marrow and peripheral blood smears, our D_4 -equivariant diffusion models achieve higher AUC and retrieve more abnormal cells in the top K predictions than non-equivariant generative baselines, a deep one-class, and a multiple instance learning based method, while substantially reducing score variance across rotations and flips. Code is available at <https://swchmida.github.io/D4diffCyto/>.

Keywords: Computational cytology · D_4 -equivariant diffusion · Unsupervised anomaly detection

1 Introduction

Cytology [1] is a core component of pathology in which decisions hinge on subtle single-cell morphology, while truly malignant or severely dysplastic cells can be orders of magnitude rarer than normal or reactive cells, even on malignant

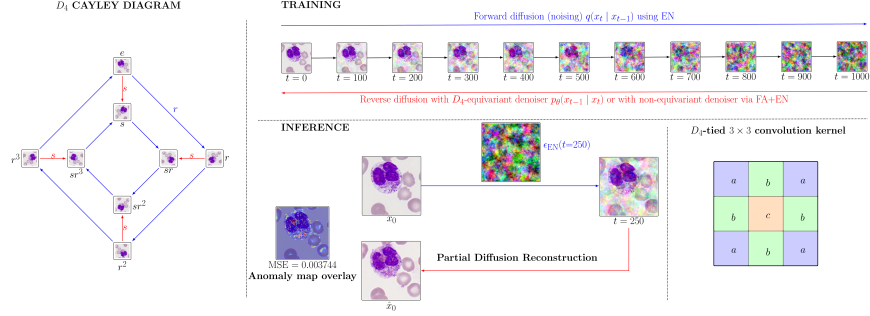


Fig. 1. Overview of the proposed D_4 -equivariant diffusion framework. Left: Cayley diagram of the dihedral group D_4 acting on a single-cell patch. Top-right: training with a D_4 -equivariant U-Net with equivariant noise (EN) coupling. Bottom-right: Inference using partial diffusion reconstruction at test time and an example D_4 -tied convolution kernel.

slides, leading to extreme class imbalance. Digital cytology and whole-slide scanners now enable high-resolution digitization of entire slides [14,17], making exhaustive manual review unrealistic when most slide-negative (and many slide-positive) cases are dominated by non-malignant morphology [3,8,10,15]. Self- and weakly supervised multiple instance learning (MIL) based slide classifiers [13,18,20,21,23,29,31] can predict slide labels but often fail to reliably surface the few key malignant instances, motivating unsupervised anomaly-detection methods [11,27] trained on guaranteed-normal slide-negative patches to model normal-cell distributions and flag outliers as candidate abnormal cells [7].

A broad range of medical anomaly-detection methods [11] has been explored, from deep one-class models such as Deep SVDD [28] to generative approaches that reconstruct toward a normal manifold and use reconstruction-based scores (e.g., GAN-based [30] and diffusion-based [4,33]). Diffusion anomaly detectors built on partial diffusion (AnoDDPM-style) [19,34] and their refinements [2,5,6,12,26,33] achieve strong lesion localization in organ-centric imaging (brain MRI, retinal OCT, X-ray), where anomalies form contiguous lesions in a fixed anatomy. In cytology, by contrast, abnormalities are local nuclear or nuclear-to-cytoplasmic changes at the single-cell level, so orientation does not change the diagnostic class, yet standard generative detectors treat rotated/flipped views as distinct inputs, increasing sample complexity and inducing orientation-dependent score variation. Under extreme class imbalance and budgeted review, cytologists inspect only a small top- K set; if rotations or flips change the rank ordering of anomaly scores, the system fails operationally even when global AUC is high. We therefore aim for *consistent ranking under D_4 transforms*, encoding group symmetry to remove nuisance variance from the generative model and stabilize anomaly scores under extreme imbalance.

We propose a dihedral-equivariant diffusion framework that enforces D_4 symmetry throughout partial-diffusion reconstruction [22,35] (see Fig. 1). Concretely, we equip the U-Net denoiser with a D_4 -equivariant architecture [9,16], sharing parameters across rotated and flipped views to reduce hypothesis space and redundant variability, and we enforce symmetry at inference via *equivariant noise coupling* (EN) and *frame averaging* (FA) [22,35], which together promote path-wise equivariant trajectories and symmetrized reconstructions. These complementary routes: architectural equivariance and inference-induced equivariance (FA+EN), directly target D_4 -consistent anomaly scores, and on two publicly available cytology datasets of single-cell patches from bone marrow and peripheral blood smears our D_4 -equivariant diffusion yields more stable scores across orientations and higher test AUC than non-equivariant generative detectors, a deep one-class baseline, and a MIL-based framework.

2 Method

In this section, we use diffusion models for anomaly detection on cytology patches and show how we inject D_4 -equivariance. A denoising diffusion model is trained on slide-negative single-cell patches to learn a normal appearance prior. At test time, we run partial diffusion: each query patch is noised to an intermediate step and then denoised back; the pixelwise residual is the anomaly map and its mean-squared energy is the anomaly score. Standard data augmentation is not enough here: the stochastic reconstruction process can still make scores change under rotations/flips, so our goal is *ranking stability under symmetry*, where patches differing only by D_4 transforms receive comparable scores.

We pursue two routes to D_4 -consistent scoring, both covered by Algorithm 1. First, we use a standard (non-equivariant) U-Net trained with the usual diffusion loss and impose D_4 structure only at inference via *equivariant noise* (EN) and *frame averaging* (FA) [22] over rotated and flipped views. Second, we use a D_4 -equivariant U-Net, built from group-equivariant convolutions and, in one variant, a D_4 -equivariant attention block with shared group-transformed relative positional biases; this backbone is trained with the same objective, and at test time Algorithm 1 is run with EN-only, letting architectural equivariance carry the D_4 structure through the sampling trajectory.

2.1 Diffusion anomaly detection via partial reconstruction

Let $x_0 \in \mathbb{R}^{C \times H \times W}$ denote a single-cell patch, with training samples drawn from the normal distribution p_n and test samples drawn from held-out slides. We use a standard denoising diffusion model: the forward process adds Gaussian noise in T steps via $q(x_t | x_0) = \mathcal{N}(\sqrt{\alpha_t} x_0, (1 - \alpha_t)\mathbf{I})$, and the denoiser $f_\theta(x_t, t)$ is trained with the usual DDPM noise-prediction loss $\mathcal{L}(\theta) = \mathbb{E}_{x_0, t, \epsilon} [\|\epsilon - f_\theta(x_t, t)\|_2^2]$ on slide-negative patches.

At test time we run *partial diffusion*: for a chosen step t^* , we diffuse x_0 to x_{t^*} using the forward process and then apply t^* reverse steps (Algorithm 1) to obtain

Algorithm 1 D_4 -aware partial-diffusion anomaly detection

```

1: Inputs: normal training set  $\{x_0^{(i)}\}_{i=1}^N$ , test patch  $x_0$ , steps  $T$ , partial step  $t^*$ , group
    $G = D_4$ ,
   backbone  $f_\theta \in \{\text{UNet}, \text{UNet}_{D_4}\}$ , flags EN, FA
2: Training
3: for minibatch  $\mathcal{B}$  do
4:   draw  $t \sim \mathcal{U}\{1, \dots, T\}$ ,  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ; set  $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$  for  $x_0 \in \mathcal{B}$ 
5:    $\hat{\epsilon} = f_\theta(x_t, t)$ ;  $\mathcal{L} = \|\epsilon - \hat{\epsilon}\|_2^2$ ; update  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ 
6: end for
7: Inference (score for  $x_0$ )
8: if FA then
9:    $\mathcal{O}(x_0) = \{g \cdot x_0 : g \in G\}$ ; draw  $\epsilon, \{\eta_t\}_{t=1}^{t^*}$ 
10:  for  $g \in G$  do
11:    if EN then  $\epsilon_g = g \cdot \epsilon$ ,  $\eta_{t,g} = g \cdot \eta_t$ ; else draw  $\epsilon_g, \eta_{t,g}$ 
12:     $x_{t^*}^g = \text{DiffForward}(g \cdot x_0, \epsilon_g, t^*)$ ;  $\hat{x}_0^g = \text{DiffReverse}(x_{t^*}^g, \eta_{1:t^*,g})$ 
13:     $A^g = |g \cdot x_0 - \hat{x}_0^g|$ ;  $S^g = \frac{1}{C_{HW}} \|g \cdot x_0 - \hat{x}_0^g\|_2^2$ 
14:  end for
15:   $A^{\text{FA}}(x_0) = \frac{1}{|G|} \sum_{g \in G} g^{-1} \cdot A^g$ ;  $S^{\text{FA}}(x_0) = \frac{1}{|G|} \sum_{g \in G} S^g$ ; return  $S^{\text{FA}}(x_0)$ 
16: else
17:  draw  $\epsilon, \{\eta_t\}_{t=1}^{t^*}$ ;  $x_{t^*} = \text{DiffForward}(x_0, \epsilon, t^*)$ ;  $\hat{x}_0 = \text{DiffReverse}(x_{t^*}, \eta_{1:t^*})$ 
18:   $A(x_0) = |x_0 - \hat{x}_0|$ ;  $S(x_0) = \frac{1}{C_{HW}} \|x_0 - \hat{x}_0\|_2^2$ ; return  $S(x_0)$ 
19: end if

```

a pseudo-normal reconstruction $\hat{x}_0 = \mathcal{R}(x_0; \epsilon, \eta_{1:t^*})$. We then define an anomaly map and scalar anomaly score as $A(x_0) = |x_0 - \hat{x}_0|$ and $S(x_0) = \frac{1}{C_{HW}} \|x_0 - \hat{x}_0\|_2^2$. All diffusion-based variants in this work use $S(x_0)$ as the final anomaly score; $A(x_0)$ is used only for qualitative visualization.

2.2 D_4 -equivariant sampling and architectures

We model in-plane rotations and flips of single-cell patches using the dihedral group $G = D_4$, acting on images by pixel re-indexing $(g \cdot x)(u) = x(g^{-1}u)$. Our goal is for the scalar score $S(\cdot)$ to be approximately invariant under G , so that patches differing only by D_4 transforms receive comparable anomaly scores.

To couple stochasticity across the orbit, we use *equivariant noise* (EN): a single base noise sample $(\epsilon, \{\eta_t\}_{t=1}^{t^*})$ is transformed as $\epsilon_g = g \cdot \epsilon$ and $\eta_{t,g} = g \cdot \eta_t$ for each $g \in G$, so the forward and reverse trajectories for $g \cdot x_0$ share the same underlying randomness as for x_0 , up to the group action. When desired, we further apply *frame averaging* (FA), averaging residuals and scores over the D_4 orbit to stabilize $A(x_0)$ and $S(x_0)$.

Architectural and sampling choices are combined in a unified way: Algorithm 1 toggles (i) the backbone (standard vs. D_4 -equivariant U-Net) and (ii) the EN/FA flags. For a generic U-Net we use EN+FA at inference to enforce D_4 consistency, while for the D_4 -equivariant U-Net (with group-tied convolutions and, in our second variant, D_4 -equivariant attention) we typically use EN-only,

relying on architectural equivariance to propagate group structure through the sampling trajectory.

2.3 Theoretical guarantees

In our setting, D_4 invariance refers to class invariance of centered single-cell patches: rotating or flipping a cell patch should not change its diagnostic class.

D_4 -averaging cannot increase DDPM risk We consider the idealized setting where $X_0 \sim p_n$ is D_4 -invariant and the injected noise $E \sim \mathcal{N}(0, \mathbf{I})$ is isotropic. Let X_t be the noised samples, E be the sampled equivariant noise and assume the training pairs are D_4 -invariant in distribution: $(X_t, E) \stackrel{d}{=} (g \cdot X_t, g \cdot E) \quad \forall g \in G$. Define the D_4 -averaged denoiser $f_\theta^{\text{avg}}(x_t, t) = \frac{1}{|G|} \sum_{g \in G} g^{-1} \cdot f_\theta(g \cdot x_t, t)$. For the squared-error loss $\ell(\hat{\epsilon}, \epsilon) = \|\hat{\epsilon} - \epsilon\|_2^2$, Jensen’s inequality together with D_4 pair invariance yields $\mathbb{E}[\ell(f_\theta^{\text{avg}}(X_t, t), E)] \leq \mathbb{E}[\ell(f_\theta(X_t, t), E)]$. Thus, symmetrizing the denoiser with respect to D_4 cannot increase the DDPM training risk and can, in principle, improve it by averaging out orientation-specific fluctuations.

Invariant detection under class-conditional symmetry: Let $Y \in \{0, 1\}$ denote normal/abnormal. Assume the class-conditionals are G -invariant: $p(x | y) = p(g \cdot x | y) \forall g \in G, y \in \{0, 1\}$. For any score $s(x)$, define its group-averaged version $\bar{s}(x) = \frac{1}{|G|} \sum_{g \in G} s(g \cdot x)$. For a false-positive budget α , the optimal miss-rate among threshold rules based on s is $\beta^*(\alpha; s) = \inf_\tau \Pr(s(X) \leq \tau | Y = 1)$ s.t. $\Pr(s(X) > \tau | Y = 0) \leq \alpha$. Under G -invariant class-conditionals, restricting to G -invariant scores (e.g., scores obtained by averaging over D_4 orbits) does not worsen this optimal trade-off. In other words, symmetry-respecting averaging is theoretically justified when the underlying data distribution exhibits the corresponding rotational and flip symmetries, aligning our D_4 -equivariant diffusion design with the structure of single-cell cytology images.

3 Experiments and Results

In this section, we describe the datasets and experimental setup, then report quantitative and qualitative results for our D_4 -equivariant diffusion models and competing baselines, followed by targeted ablation studies.

3.1 Datasets and splits

We evaluate on two cytology single-cell datasets. The MLL bone marrow dataset [25] contains 250×250 RGB patches from bone marrow smears. We treat typical lymphocytes (LYT) as the normal class, using 9,185 LYT patches for training, 7,873 LYT patches for testing, and merge all seven minority diagnostic subsets (e.g. blasts and other rare morphologies) into a single abnormal class with 396 test patches.

The AML LMU Cytomorphology dataset [24] contains 400×400 RGBA patches from peripheral blood smears. We group mature leukocytes as normal and immature leukocytes or *other* types as abnormal. From the mature group, we use 10,383 patches for training and 4,450 for testing; from the immature group we use all 3,532 patches for testing.

Table 1. Patch-level anomaly detection on MLL bone marrow and AML Cytomorphology LMU datasets (AUC and number of true positives among top- K ranked patches).

Method	MLL				AML LMU			
	AUC	True Positives			AUC	True Positives		
		TP ₄₀₀	TP ₂₀₀	TP ₁₀₀		TP ₄₀₀	TP ₂₀₀	TP ₁₀₀
Baselines								
ItS2CLR	0.521107	27	20	19	0.360515	92	38	13
Deep SVDD	0.613331	49	35	21	0.514742	227	122	56
f-AnoGAN	0.594601	37	24	9	0.538857	273	131	57
THOR (Gaussian)	0.540851	24	14	7	0.582212	247	137	68
THOR (Simplex)	0.404264	8	3	1	0.544415	207	104	58
BerDiff	0.550799	44	32	26	0.708558	313	157	64
AnoDDPM	0.632846	57	30	18	0.694320	311	162	83
D_4 -equivariant diffusion variants								
FA only (inference)	0.612543	46	32	19	0.680111	277	144	72
FA+EN (inference)	0.684121	78	53	34	0.678373	285	149	75
D_4 conv.	0.675404	74	53	27	0.699635	315	159	86
D_4 conv.+attn.	0.679251	67	44	33	0.720098	325	162	81

3.2 Setup and baselines

We benchmark *patch-level* anomaly detection. Each method assigns an anomaly score to every test patch and we report AUC (Area Under the Receiver Operating Characteristic Curve) as the primary metric. Since cytology screening is dominated by extreme class imbalance and review is budgeted, we treat each test set for both datasets as a single abnormal *slide* containing a mix of normal and abnormal single-cell patches. Thus, we also measure the number of true positives in the top- K most anomalous predictions (TP₄₀₀, TP₂₀₀, TP₁₀₀), reflecting realistic pre-screening where only a limited candidate set is inspected. For our AnoDDPM and D_4 -equivariant AnoDDPM variants, the anomaly score is the mean-squared error (MSE) between the input patch and its reconstruction.

As baselines we include: (i) a MIL-based instance scorer (ItS2CLR) [21]; (ii) a deep one-class method (Deep SVDD)[28]; and (iii) representative generative detectors: f-AnoGAN[30], THOR (with gaussian and simplex noise)[6], Masked Bernoulli Diffusion (BerDiff)[32], and a vanilla AnoDDPM detector[34] (non-equivariant U-Net, simplex noise, partial diffusion). All methods are trained on the same normal training splits and evaluated on exactly the same test sets.

Our D_4 -equivariant diffusion models fall into two families: those that impose group structure only at inference via Frame Averaging (FA) and Equivariant Noise (EN), and those that build D_4 structure into the U-Net backbone; the concrete instantiations are detailed in Section 3.4.

3.3 Quantitative results

Table 1 summarizes AUC and TP_K on both datasets. On MLL, the non-equivariant AnoDDPM baseline reaches $AUC = 0.632$ with $TP_{400} = 57$. Our best D_4 variant, FA+EN with a generic U-Net, improves AUC to 0.684 and TP_{400} to 78, while the D_4 -equivariant convolutional U-Net (no attention) yields $AUC = 0.679$ and $TP_{400} = 74$. This shows that enforcing D_4 structure, either architecturally or via inference-time averaging, leads to both better global separation and denser concentration of truly abnormal cells at the top of the ranked list.

On AML LMU, BerDiff and vanilla AnoDDPM are strong generative baselines ($AUC = 0.708$ and 0.694 , respectively). Here the best performance is obtained with the D_4 -equivariant U-Net with attention: $AUC = 0.720$, $TP_{400} = 325$, and $TP_{200} = 162$. This indicates that D_4 -equivariant diffusion remains beneficial in a different cytology modality (blood smears) as well. Figure 2 shows some qualitative examples on both datasets.

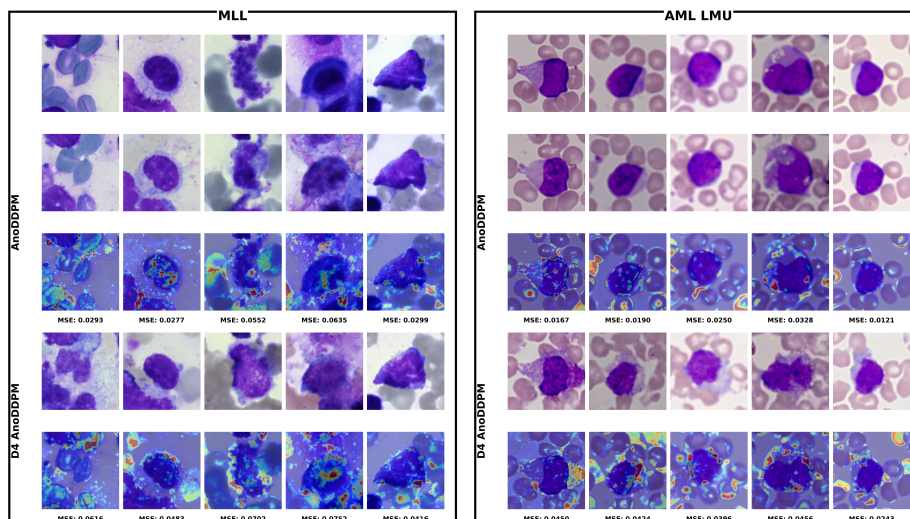


Fig. 2. Qualitative comparison of AnoDDPM vs. D_4 -equivariant AnoDDPM on MLL bone marrow (left panel) and AML LMU (right panel). For each method we show inputs (row 1), reconstructions (rows 2 and 4), and qualitative reconstruction-deviation overlays with patch-level anomaly-score MSE values (rows 3 and 5).

3.4 Ablation studies

Inference-induced vs. architectural D_4 variants. We ablate four D_4 variants (Table 1). On MLL, FA-only with a standard U-Net slightly degrades AUC, while FA+EN recovers and improves over vanilla AnoDDPM. The D_4 convolutional U-Net on the other hand, slightly outperforms the attention-augmented variant. On AML LMU, FA-only and FA+EN each improve top- K retrieval over the baseline, whereas the architectural D_4 models further strengthen both AUC and TP counts, with the attention-based variant giving the strongest overall diffusion performance.

Anomaly-score invariance under D_4 transforms. On MLL, we sample 100 normal and 100 abnormal patches and generate their rotations by 90° , 180° , 270° and horizontal flips, then measure variation of anomaly scores as (mean SD, median SD, mean range) per image, where SD is standard deviation. Vanilla AnoDDPM yields (0.00375, 0.00343, 0.01036), whereas the D_4 FA+EN variant reduces this to (**0.00206**, **0.00161**, **0.00515**) and the D_4 U-Net without attention to (0.00224, 0.00177, 0.00560), indicating substantially more rotation- and flip-consistent anomaly scores for the dihedral-aware diffusion models.

Uncertainty and statistical significance. For the MLL dataset, we used stratified bootstrap resampling of test patches and compared each D_4 -equivariant model against AnoDDPM via paired two-sided tests on AUC and TP_{400} . AnoDDPM obtained AUC 95% CI (0.603, 0.661) and TP_{400} CI (43, 71), while the D_4 FA+EN variant tightened these to (0.657, 0.710) and (61, 93), and the D_4 convolutional U-Net (no attention) to (0.647, 0.703) and (58, 89). Paired bootstrap tests gave $p < 10^{-3}$ for AUC and **$p = 0.006$ for TP_{400} for both D_4 models versus AnoDDPM**, making the proposed method clinically significant and indicating statistically reliable gains in ranking and top- K retrieval. On the AML LMU dataset, BerDiff showed AUC CI (0.697, 0.720) and TP_{400} CI (295, 328), whereas the D_4 -equivariant U-Net with attention achieved (0.709, 0.731) and (307, 340). Paired bootstrap tests yielded $p \approx 0.125$ for AUC and $p \approx 0.275$ for TP_{400} , so the D_4 model’s advantage is numerically consistent but statistically inconclusive on AML LMU; with the clearly significant MLL gains. Together, these results suggest that enforcing D_4 structure can improve anomaly ranking, with strong evidence on MLL and encouraging trends on AML LMU.

4 Conclusion

We addressed a structural mismatch between generative anomaly detection and computational cytology. In cell-centric imaging, diagnostic class is invariant to in-plane rotations and reflections, yet standard diffusion models yield transformation-dependent reconstructions and unstable anomaly scores. We introduced a D_4 -equivariant diffusion framework that enforces dihedral symmetry in both architecture and sampling, yielding rotation/flip-consistent reconstructions and more stable anomaly ranking under extreme class imbalance. Across two cytology datasets, enforcing D_4 structure improved global separation and clinically relevant top- K retrieval: on MLL, the D_4 FA+EN variant raised TP_{400} from 57

(AnoDDPM) to 78 with statistically significant gains in both AUC and TP_{400} ($p = 0.006$), while on AML LMU, the D_4 -equivariant U-Net with attention achieved the best AUC (0.720) and TP_{400} (325). Overall, aligning generative anomaly detection with cytology’s dihedral symmetries enhances ranking stability and improves practical abnormal-cell retrieval in realistic screening workflows.

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

Acknowledgements. This work was supported by the Liljewalch Foundation, Sweden, through a travel scholarship awarded to the first author for a research visit to Technical University of Darmstadt.

References

1. Al-Abbadi, M.A.: Basics of cytology. *Avicenna journal of medicine* **1**(01), 18–28 (2011)
2. Beizae, F., Lodygensky, G., Desrosiers, C., Dolz, J.: Mad-ad: Masked diffusion for unsupervised brain anomaly detection. In: IPMI. pp. 139–153. Springer (2025)
3. Bengtsson, E., Malm, P.: Screening for cervical cancer using automated analysis of pap-smears. *Computational and mathematical methods in medicine* **2014**(1), 842037 (2014)
4. Bercea, C.I., Cattin, P.C., Schnabel, J.A., Wolleb, J.: Denoising diffusion models for anomaly localization in medical images. *Machine Learning for Biomedical Imaging* (2025)
5. Bercea, C.I., Neumayr, M., Rueckert, D., Schnabel, J.A.: Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. *arXiv preprint arXiv:2305.19643* (2023)
6. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Diffusion models with implicit guidance for medical anomaly detection. In: MICCAI. pp. 211–220. Springer (2024)
7. Chatterjee, S., Göksel, O., Sladoje, N., Lindblad, J.: Detection of extremely sparse key instances in whole slide cytology images via self-supervised one-class representation learning. In: ICPR. pp. 408–421. Springer (2024)
8. Cheng, S., Liu, S., Yu, J., Rao, G., Xiao, Y., Han, W., Zhu, W., Lv, X., Li, N., Cai, J., et al.: Robust whole slide image analysis for cervical cancer screening using deep learning. *Nature communications* **12**(1), 5639 (2021)
9. Chidester, B., Ton, T.V., Tran, M.T., Ma, J., Do, M.N.: Enhanced rotation-equivariant u-net for nuclear segmentation. In: IEEE CVPR Workshops (2019)
10. Coustan-Smith, E., Song, G., Clark, C., Key, L., Liu, P., Mehrpooya, M., Stow, P., Su, X., Shurtleff, S., Pui, C.H., et al.: New markers for minimal residual disease detection in acute lymphoblastic leukemia. *Blood, The Journal of the American Society of Hematology* **117**(23), 6267–6276 (2011)
11. Fernando, T., Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Deep learning for medical anomaly detection—a survey. *ACM Computing Surveys (CSUR)* **54**(7), 1–37 (2021)

12. Fuchs, M., Sivakumar, S.K.R., Schöber, M., Woltering, N., Eich, M.L., Schweizer, L., Mukhopadhyay, A.: Harp: Unsupervised histopathology artifact restoration. In: *Medical Imaging with Deep Learning*. pp. 465–479. PMLR (2024)
13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *ICML*. pp. 2127–2136. PMLR (2018)
14. Jiang, H., Zhou, Y., Lin, Y., Chan, R.C., Liu, J., Chen, H.: Deep learning for computational cytology: A survey. *Medical Image Analysis* **84**, 102691 (2023)
15. Kruse, A., Abdel-Azim, N., Kim, H.N., Ruan, Y., Phan, V., Ogana, H., Wang, W., Lee, R., Gang, E.J., Khazal, S., et al.: Minimal residual disease detection in acute lymphoblastic leukemia. *International journal of molecular sciences* **21**(3), 1054 (2020)
16. Lafarge, M.W., Bekkers, E.J., Pluim, J.P., Duits, R., Veta, M.: Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis* **68**, 101849 (2021)
17. Landau, M.S., Pantanowitz, L.: Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. *Journal of the American Society of Cytopathology* **8**(4), 230–241 (2019)
18. Li, Y., Liu, Y., Xu, Y., Zhang, L., Xing, L., Huang, J.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: *IEEE ICCV*. pp. 14318–14327 (2021)
19. Linmans, J., Raya, G., van der Laak, J., Litjens, G.: Diffusion models for out-of-distribution detection in digital pathology. *Medical Image Analysis* **93**, 103088 (2024)
20. Liu, J., Mao, A., Niu, Y., Zhang, X., Gong, T., Li, C., Gao, Z.: Pamil: Prototype attention-based multiple instance learning for whole slide image classification. In: *MICCAI*. pp. 362–372. Springer (2024)
21. Liu, K., Zhu, W., Shen, Y., Liu, S., Razavian, N., Geras, K.J., Fernandez-Granda, C.: Multiple instance learning via iterative self-paced supervised contrastive learning. In: *IEEE CVPR*. pp. 3355–3365 (2023)
22. Lu, H., Szabados, S., Yu, Y.: Diffusion models under group transformations. In: *AISTATS* (2025)
23. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**(6), 555–570 (2021)
24. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T., Marr, C.: An expert-annotated dataset of bone marrow cytology in hematologic malignancies. *The Cancer Imaging Archive (TCIA) [Data set]* (2021)
25. Matek, C., Schwarz, S., Marr, C., Spiekermann, K.: A single-cell morphological dataset of leukocytes from aml patients and non-malignant controls. *The Cancer Imaging Archive (TCIA) [Data set]* (2019)
26. Naval Marimont, S., Siomos, V., Baugh, M., Tzelepis, C., Kainz, B., Tarroni, G.: Ensembled cold-diffusion restorations for unsupervised anomaly detection. In: *MICCAI*. pp. 243–253. Springer (2024)
27. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE* **109**(5), 756–795 (2021)
28. Ruff, L., Vandermeulen, R.A., Görnitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: *ICML*. pp. 4393–4402. PMLR (2018)

29. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J.: Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical image analysis* **79**, 102464 (2022)
30. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis* **54**, 30–44 (2019)
31. Torpey, D., Klein, R.: Deepset simclr: Self-supervised deep sets for improved pathology representation learning. *Pattern Recognition Letters* **186**, 64–70 (2024)
32. Wolleb, J., Bieder, F., Friedrich, P., Zhang, P., Durrer, A., Cattin, P.C.: Binary noise for binary tasks: Masked bernoulli diffusion for unsupervised anomaly detection. In: MICCAI. pp. 135–145. Springer (2024)
33. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: MICCAI. pp. 35–45. Springer (2022)
34. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: IEEE CVPR. pp. 650–656 (2022)
35. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. In: ICLR (2024), <https://openreview.net/forum?id=FKksTayvGo>