Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Out-of-distribution detection in digital pathology: Do foundation models bring the end to reconstruction-based approaches?☆

Milda Pocevičiūtė [a,b,*], Yifan Ding [a], Ruben Bromée [a], Gabriel Eilertsen [a,b]

[a] Department of Science and Technology, Linköping University, Campus Norrköping, Norrköping, SE-601 74, Sweden
[b] Center for Medical Imaging and Visualization, Linköping University, University Hospital, Linkoping, SE-581 85, Sweden

ARTICLE INFO

Keywords:
Latent diffusion model
Foundation models
Out-of-distribution detection
Digital pathology
Deep learning

ABSTRACT

Artificial intelligence (AI) has shown promising results for computational pathology tasks. However, one of the limitations in clinical practice is that these algorithms are optimised for the distribution represented by the training data. For out-of-distribution (OOD) data, they often deliver predictions with equal confidence, even though these often are incorrect. In the pursuit of OOD detection in digital pathology, this study evaluates the state-of-the-art (SOTA) in computational pathology OOD detection, based on diffusion probabilistic models, specifically by adapting the latent diffusion model (LDM) for this purpose (AnoLDM). We compare this against post-hoc methods based on the latent space of foundation models, which are SOTA in general computer vision research. The approaches are not only evaluated on data from the same medical centres as the training set, but also on several datasets with data distribution shifts. The results show that AnoLDM performs similarly well or better than diffusion model based approaches published in previous studies in computational pathology but with reduced computational costs. However, our optimal configuration of an approach based on foundation models (*kang_residual*) outperforms AnoLDM on OOD detection on data not experiencing any covariate shifts, with an AUROC of 96.17 versus 91.86. Interestingly, AnoLDM is more successful at handling the data distribution shifts investigated in this study. However, both AnoLDM and *kang_residual* suffer substantial loss in the performance under the data distribution shifts, hence future work should focus on improving the generalisation of OOD detection for computational pathology applications.

1. Introduction

Pathology is an important discipline in clinical medicine for tissue-based diagnostics as it provides insights into the cellular and molecular mechanisms of diseases. Despite the valuable contributions of molecular and 'omics' data to histological evaluations, the microscopic examination of morphological alterations continues to be a fundamental aspect of pathology. As a result, the majority of computational pathology initiatives are centred around the analysis of haematoxylin and eosin (H&E) stained whole slide images (WSIs), the digitalised microscopy slides [1]. This discipline leverages artificial intelligence (AI), especially deep learning (DL) methods, to aid in the diagnosis, characterisation, and comprehension of diseases. Although AI has demonstrated promising proof-of-concepts and applications [2–5], achieving generalisation and robustness remains a challenge in clinical practice [6,7].

Two main distributional shifts can seriously affect the performance of AI models in pathology applications: semantic and covariate shift. Semantic shift refers to situations when labels in inference data do not match the label space in development data. An example of semantic shift is a diffuse large B-cell lymphoma encountered by an AI that is analysing lymph node sections for colon cancer metastasis detection. If the model has not been trained to detect lymphomas, by default it will not be able to warn the care providers that this patient has an additional disease. A way to address semantic shift is to implement out-of-distribution (OOD) detection methods. These target data samples that have previously, i.e., during the development of the AI, unseen characteristics, hence they are able to detect input with a semantic shift. There has been an increasing interest in the research community to propose methods for OOD detection in computational pathology [8–10].

☆ This research was supported by Swedish e-Science Research Center, VINNOVA, Sweden, The Zenith career development program at Linköping University, Sweden, and Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, Sweden.

* Corresponding author.
E-mail addresses: milda.poceviciute@liu.se (M. Pocevičiūtė), yifan.ding@liu.se (Y. Ding), rubbr232@student.liu.se (R. Bromée), gabriel.eilertsen@liu.se (G. Eilertsen).

https://doi.org/10.1016/j.compbiomed.2024.109327
Received 27 June 2024; Received in revised form 20 September 2024; Accepted 22 October 2024
0010-4825/© 2024 Published by Elsevier Ltd.

Please cite this article as: Milda Pocevičiūtė et al., *Computers in Biology and Medicine*, https://doi.org/10.1016/j.compbiomed.2024.109327

A covariate shift happens when there is a change in data features but the label space remains the same. A common cause for covariate shift is difference in fixation, staining or scanning protocols or variations in morphology between different populations which do not affect the label space. Furthermore, in the clinical setting a combination of semantic and distributional shifts may occur at the same time, e.g., if an AI system developed and validated on medical centres in the US is deployed in a hospital in Sweden and encounters a previously unseen disease. Multiple strategies to handle the covariate shift have been proposed in computational pathology: one can alter the training of the AI to make it more robust [11,12] or detect when the shift is occurring to warn the responsible parties [13,14]. A key question is how OOD detection methods are affected by covariate shifts that are likely to happen in a clinical setting.

A major paradigm change in computational pathology is fuelled by the advent of foundation models: large DL models that learn meaningful universal features from extremely large datasets without expert annotations. Several works have shown an incredible adaptability of these models for various tasks in pathology: from rare cancer subtyping to prognostic factors prediction [15–17]. However, foundation models may not only benefit in diagnostic AI model development, but also provide new means of OOD detection via utilisation of their latent space in existing OOD detection methods [18]. The hypothesis is that the richness of the learnt features by the foundation models would enable a more precise OOD detection than using other AI models.

In this work, we evaluate OOD detection on pathology data with a focus of comparing diffusion models with other state-of-the-art (SOTA) techniques based on latent space of foundation models. We are building on work by Graham et al. [19] and Linmans et al. [9] that have shown that reconstruction based OOD detection works best using diffusion model: it outperforms substantially generative adversarial networks (GANs) and variational auto-encoders (VAEs) and simple distance based approaches. However, diffusion models are computationally costly, preventing hospitals from adapting the technology and contributing to high carbon emission footprints [20]. To increase the efficiency of diffusion models for OOD detection, we utilise the latent diffusion model (LDM) [21]. Furthermore, we evaluate if good results could be achieved by using fewer sampling steps in the diffusion process in combination with a more modern sampler. Continuing the naming convention in [8,9] we name our proposed diffusion-based approach as *AnoLDM*.

Currently, SOTA approaches for OOD detection in general computer vision are based on the latent features (sometimes in combination with the logits) from supervised classifiers. However, most of those methods cannot be directly used in computational pathology due to the need of a supervised classifier trained on in-distribution (ID) data. Still, methods such as [22,23] combined with foundation models [18] can be adapted to ID data with no labels, and we name such combinations *foundation-based methods*.

Previous work in computational pathology has not evaluated how reconstruction-based approaches compare to the foundation-based methods. A key difference in these techniques is what data the AI has been exposed to during the training. In the reconstruction based methods, the generative model only sees ID samples during the training which, in theory, should enable it to detect anything that is different from the training distribution. In contrast, it is often infeasible to ensure the same constraint on the foundation models as they are usually developed by third parties due to immense costs associated with their development. Nevertheless, given their promise and growing adaptation, we believe it is important to evaluate whether relying on foundation models alone could alleviate the need to train an additional AI method (such as a diffusion model) for a successful OOD detection. We evaluate three publicly available foundation models: one trained on natural images and two specifically tailored for pathology applications. We hypothesise that different foundation models may be suitable for different tasks, even if they have been trained for the same purpose, such as computational pathology. Following [8,9], we define benign lymph node tissue as ID and tumour tissue as OOD. This setup maximises the distance between the two distributions, hence serves well as the initial evaluation step of OOD detection systems.

We argue that it is crucial for an OOD detection method to be universal, meaning that it should detect various types of OOD samples if the ID distribution is unchanged. [24] have shown that an AI algorithm, trained to detect breast cancer metastasis to lymph nodes, cannot handle well other types of cancer metastases, i.e. from colon or head and neck cancers. This motivated us to investigate whether an OOD detection framework, developed to detect breast cancer metastases as OOD samples, can generalise successfully to other types of cancer metastases to lymph nodes. In this case the ID distribution, i.e. the benign lymph node tissue, remains the same whether we are handling breast or colon cancer patients. Therefore, there is ground to assume that the OOD detection system should work well. Finally, a critical unanswered question is how resilient OOD detection frameworks are against covariate shifts that are commonly encountered in clinics. To answer this question, we have developed the OOD detection methods using data from 5 medical centres in the Netherlands and tested them on datasets not only from those centres but also from a medical centre in Sweden. Our contributions can be summarised as:

1. we compare SOTA OOD detection methods based on reconstruction with diffusion models against post-hoc methods using the latent space of foundation models,
2. we successfully adapt reconstruction-based OOD detection with diffusion models to LDMs, by optimising for the best sampler, sampling steps, noise steps, normalisation strategy, and distance measures in the latent space of the LDM,
3. we show what role the choice of the foundation model plays in OOD detection by utilising three recent publicly available foundation models,
4. we evaluate if minor changes to OOD sample definition, i.e. from breast cancer to colon cancer metastasis, have an impact on the success of detection,
5. we investigate whether the OOD detection methods are affected by a clinically realistic covariate shift.

2. Related work

Types of OOD detection: OOD detection in medical imaging is often referred to as anomaly detection, and there has been some confusion in how these differ and are defined [25,26]. Here, we consider OOD and anomaly detection as similar problem formulations, but which could be targeting different types of applications (i.e., OOD detection could be used to detect anomalies in a specific application). Furthermore, we consider OOD detection in digital pathology as a near-OOD problem [26], since ID and OOD samples are similar in composition and with only subtle differences between the ID and OOD domains.

Most previous work consider OOD detection under semantic shifts, such as label-space differences (i.e., OOD data contains images with class labels not part of the ID data). However, it is also possible to define covariate shifts, where images contain the same semantics but differ in their distributions due to other factors (e.g., differences in colour, contrast, composition, etc.). Generalising to these shifts is referred to as full-spectrum OOD detection, which is a significantly more challenging scenario [26]. A limited number of previous work has focused on the full-spectrum setting [27], also in digital pathology [28]. We recognise full-spectrum OOD detection to be especially important in digital pathology. For example, we are interested in evaluating if an OOD detection method can successfully generalise to data from different hospitals or from different types of tissue, which could be of critical importance for clinical deployment.

OOD detection for natural images: The OOD detection problem in imaging has been tackled with a range of different approaches. Likelihood-based deep generative models (DGMs) can be directly applied, as the likelihood score provides a direct way of assessing if a given sample should be considered ID or OOD [29,30]. However, it has been observed that such models often obtain higher likelihood on OOD data [31,32], likely due to focusing on low-level image features. Another way of utilising DGMs is in reconstruction-based OOD detection, where the closest match of a given test image is generated by the DGM. Since the DGM has only been exposed to ID data during training, the difference between the test image and the reconstructed image is larger for OOD data compared to ID data, such that this difference can be used as a score for flagging OOD. Reconstruction-based OOD detection has been tested with different generative models, such as GANs [8,33,34] and diffusion models [9,19,35].

For natural images, OOD detection has shifted towards post-hoc methods [22,23] and training time regularisation [36]. Post-hoc methods utilise large pre-trained, or foundational, models, by formulating a score function based on extracted latent representations and/or logits. SOTA methods, such as Virtual logit Matching (ViM) [23], ReAct [37], and Generalised Entropy [38], require a supervised model for extracting logits, i.e. the ID data need to be labelled. Some recent methods, such as KNN [22], Residual [23], and VGLR [39], can also be applied to unlabelled images, e.g., utilising self-supervised encoders for extracting representations.

OOD detection in digital pathology: In digital pathology, current SOTA is represented by diffusion models for reconstruction-based OOD detection. Linmans et al. [9] use DDPM [40], where different levels of noise are added to a test image followed by DDPM denoising. The denoised image is compared against the test image using measures such as MSE, SSIM, and LPIPS, or combinations thereof. The results showed promising improvements over other reconstruction-based methods, e.g., based on GANs [8,33,34] or autoencoders. However, the denoising is computationally demanding, even with an efficient sampler.

Additionally, researchers have explored estimating predictive uncertainty as a means of OOD detection. By providing calibrated uncertainty estimates, deep learning models can identify abnormalities and unseen classes such as lymphoma cases when looking for cancer metastasis to the lymph nodes [41,42]. However, many commonly used uncertainty estimates are negatively affected by domain shift, hence their utilisation for OOD detection may be unreliable in a clinical setting [43].

Although full-spectrum OOD detection is important in digital pathology, we are only aware of one previous work that has explored this setting. Bozorgtabar et al. [28] proposed to use clustering of a classifier's features to detect OOD samples in colorectal data. Their results indicate that the domain shift (type of covariate shift) negatively affects effectiveness of the OOD detection performance. However, they do not consider foundation models and the current SOTA OOD detection techniques from computer vision.

In this work, we accelerate reconstruction-based OOD detection with diffusion models by using the latent diffusion model (LDM) [21]. We show how metrics such as PSNR and SSIM provides better OOD scores if applied in the latent space of the LDM, and compare the results for different samplers that could further accelerate the denoising process. Then, we test the performance of SOTA post-hoc OOD detection methods, which have been successful for natural images but not tested in histopathology. We show how the pre-trained encoder used to extract representations is of critical importance and need to be trained on pathology data to allow for good performance. Finally, we show how the different methods compare in the full-spectrum setting, by investigating the generalisation capabilities when testing OOD detection on data from different hospitals and organs.

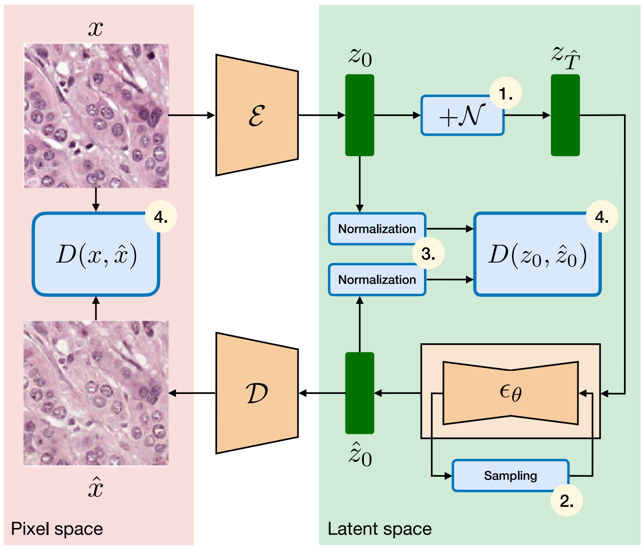

Fig. 1. AnoLDM performs reconstruction-based OOD detection in the latent space of the LDM [21]. We formulate the OOD detection in this setting by considering: (1) the amount of noise added to the latent representation; (2) the sampler and number of sampling steps for best quality-efficiency trade-off; (3) normalisation of the latent representations for comparison; (4) the measures used to formulate a reconstruction-based anomaly score in latent and/or pixel space.

3. Methods and data

In this work, we evaluate two SOTA approaches for OOD detection on digital pathology data. The established technique in computational pathology is based on the reconstruction error of diffusion probabilistic models. We adapt this strategy to LDMs, by considering different options in terms of added noise, sampling strategy, and distance measures (Section 3.1). In general computer vision, current SOTA results are achieved by foundation-based methods, extracting features and/or logits from a pre-trained encoder model. We test two different foundation-based methods in Section 3.2 that are applicable in our scenario for digital pathology where no labels are available; Residual Score (Section 3.2.2) and Deep Nearest Neighbour (Section 3.2.3).

3.1. AnoLDM

Our AnoLDM is formulated in a similar fashion as previous AnoDDPM [9], i.e. by adding a certain level of noise to a test sample followed be denoising and comparison to the original test sample. The main difference is that the diffusion process run in the latent space of the LDM instead of directly in pixel-space, as illustrated in Fig. 1.

3.1.1. The latent diffusion model

In LDM, the image x is encoded by $\mathcal{E}$ to a latent representation, $z = \mathcal{E}(x)$. A diffusion process adds noise to the latent representation, $z_t = z_0 + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma_t)$, while the UNet denoiser ϵ_θ is trained to iteratively remove noise, $z_{t-1} = z_t - \epsilon_\theta(z_t, t)$. Here, the UNet is trained to predict the noise at step t, by means of the loss

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta(z_t,t)\right\|_2^2\right]. \quad (1)$$

After denoising, the latent representation $\hat{z}_0$ can be decoded to image space through the decoder, $\hat{x} = \mathcal{D}(\hat{z}_0)$.

3.1.2. OOD detection with AnoLDM

In reconstruction-based OOD detection with diffusion models, the forward process adds a certain level of noise, $z_{\hat{T}} = z_0 + \epsilon_{\hat{T}}$, which results in a noisy representation that still maintains the structure of the image. If ϵ_θ is trained only on ID data, the rationale is that this partial noising-denoising will remove features that are specific to OOD data, essentially mapping OOD data to be closer to ID data. Thus, by measuring the distance $D(x, \hat{x})$, we expect a larger discrepancy for OOD data as compared to ID data, so that D can be used as an OOD score. Since we use the LDM, we also have the option to measure the distance directly in the latent space, $D(z_0, \hat{z}_0)$. We expect this higher-level representation to better capture the semantics of the images, so that the distance is less influenced by small discrepancies between the input and reconstructed samples (e.g., a small shift in the position of a cell can have less impact if we consider latent representations instead of images).

We train the denoising UNet, ϵ_θ, on ID data, while the autoencoder, $\mathcal{E}$ and $\mathcal{D}$, is pre-trained on natural images and fixed. Through initial experiments, we verified that the autoencoder generalises well to pathology images.

3.1.3. Calibration of AnoLDM

For AnoLDM, we make a number of considerations to improve the performance and reduce the computational demand, as denoted by the 4 points in Fig. 1:

Noise level (1): For the purpose of efficiency, we consider only one noise level instead of averaging over denoised samples from multiple noise levels as in previous OOD detection methods using diffusion models [9,19]. By searching for the optimal noise level, we can provide high performance while limiting the inference time to one reconstruction per test image.

Sampling (2): We optimise for the best combination of sampler and number of sampling steps to determine the best trade-off between quality and efficiency. To this end, we include the original DDIM sampler [44] used in the LDM [21], as well as more SOTA methods, including PLMS [45], DPM [46], and UniPC [47]. While the more recent samplers have demonstrated how it is possible to reduce the number of sampling steps without sacrificing quality, it is not clear how this translates to OOD detection with diffusion models. In this case, it is more important to reconstruct an image that is faithful to the test image as opposed to being of highest visual quality. For the different samplers, we use the original implementations and test different variations of the number of sampling steps.

Latent space normalisation (3): Since the latent space of the LDM is unbounded, it could be important to normalise representations before measuring the distance $D(z_0, \hat{z}_0)$, e.g., to account for the influence of extreme values. We compare 3 different strategies for this purpose, which are formulated to restrict z to be in the confined interval $[0, 1]$:

1. Normalisation using the per-image max and min values, $\tilde{z} = (z - z_{min})/(z_{max} - z_{min})$, where $z_{max} = \max(\max(z_0), \max(\hat{z}_0))$ and $z_{min} = \min(\min(z_0), \min(\hat{z}_0))$.
2. Normalisation as above, but using the min and max values over all images in the calibration dataset (see data setup below).
3. Normalisation as above, but using the 1% and 99% percentiles of the calibration data, followed by clamping, $\min(1, \max(0, \tilde{z}))$. We test this to see if removing extreme values can improve the distance measure.

Distance measures (4): We compare image distances both in pixel-space, $D(x, \hat{x})$, and in the latent space of the LDM, $D(z_0, \hat{z}_0)$. We include 5 different metrics D: MSE, SSIM, MS-SSIM, LPIPS using AlexNet as feature extractor, and LPIPS using VGG as feature extractor. The distance metrics were chosen based on which aspects of the images they are comparing. MSE is a widely-used metric that quantifies the average squared differences between corresponding pixel values, making it effective for measuring overall pixel-wise accuracy. Furthermore, we include SSIM and MS-SSIM as these metrics consider changes in structural information, luminance, and contrast, making it highly effective for assessing perceptual image quality [48,49]. Finally, we use LPIPS measure which leverages deep learning models to capture perceptual differences, providing a more nuanced and human-like assessment of image similarity [50]. While LPIPS is already performing a representation-based distance evaluation, we hypothesise that low-level metrics such as MSE and SSIM will perform better in the LDM latent space by capturing higher level image semantics instead of low-level statistics. This could be important in reconstruction-based OOD detection, where image features could be slightly modified or relocated in the noising-denoising process. Similar to Linmans et al. [9] we also test different combinations of metrics by averaging a Z-score,

$$Z = \frac{1}{M}\sum_{i=1}^{M}\frac{D_i(z_0, \hat{z}_0) - \mu_i}{\sigma_i}, \tag{2}$$

where μ_i and σ_i is the mean and standard deviation, respectively, for metric D_i over the calibration dataset. For SSIM and MS-SSIM, we invert the scores as 1−SSIM and 1−MS-SSIM, to align with the error-based MSE and LPIPS metrics. The Z-score can also be formulated by combining metrics in both pixel and latent spaces. LPIPS metric was computed using *lpips* python package [51] MSE was computed using *torch.nn.MSELoss* function in pytorch package [52] , and SSIM and MS-SSIM were computed using *pytorch-msssim* package [53].

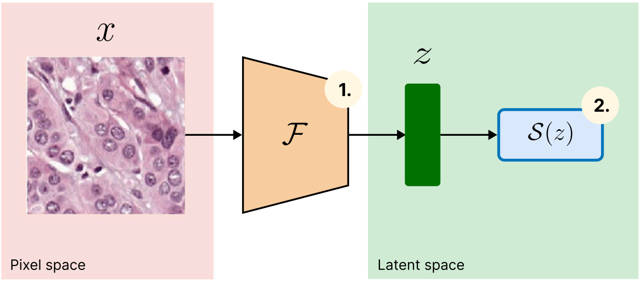

Fig. 2. OOD detection in the representation space relies on two essential components: (1) a robust foundation model, denoted as $\mathcal{F}$, and (2) an appropriate score function, $\mathcal{S}$, applied to the extracted image latent representations, the OOD detection is from the final latent score. The score functions used in this paper are KNN and Residual score.

3.2. OOD detection in representation space

The concept behind these methods is to develop a scoring algorithm that operates in the image representation space rather than the original image space. The effectiveness of OOD detection is heavily influenced by the choice of image encoder used to extract these representations, as illustrated in Fig. 2.

3.2.1. Image encoder

Let input image be x. The representation z are extracted using a pre-trained foundation model image encoder $\mathcal{F}$, where $z = \mathcal{F}(x)$. Then a OOD detection score $\mathcal{S}(z)$ is calculated on the representation z. We deploy foundation models trained on both natural image and medical image, details of those models are in Section 4.2.

3.2.2. Residual score

We have adapted Residual Score [23] OOD detection approach to detect OOD samples in the latent space of foundation models in computational pathology. This method is based on an assumption that features generally lie in low-dimensional manifolds compared to its original space. Therefore, the Residual Score [23] is calculated on a low-dimensional subspace spanned by PCA. Given training set $\mathbf{Z}$, a matrix containing all feature vectors extracted by a foundation model, the eigen decomposition on $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ is

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}, \tag{3}$$

where eigenvalues $\mathbf{\Lambda}$ are sorted decreasingly. Define residual $z^{p^{\perp}}$ as the projection of any given feature z onto subspace $P^{\perp}$. Let a new matrix $\mathbf{R}$ be the last L columns of $\mathbf{Q}$, then $z^{p^{\perp}} = \mathbf{R}\mathbf{R}^{\mathbf{T}}z$. Thus, the residual score is defined as the l2-norm of $z^{p^{\perp}}$,

$$Residual(z) = \|z^{p^{\perp}}\|. \tag{4}$$

This score reflects the variable changes in the low eigen value principle subspace, showing different patterns between ID and OOD data. Please note that we have no access to the last penultimate layer since the foundation model is not a supervised classifier. Thus, the offset $\mathbf{o} = -(\mathbf{W}^{\mathbf{T}})^{+}b$ in the original implementation is not calculated, and we simply set $\mathbf{o} = 0$.

3.2.3. Deep nearest neighbour

Another simple but practical post-hoc method that can be directly used is deep k-Nearest Neighbour (KNN) for OOD detection [22], hence we investigate its success when combined with the foundation models in our setting. Here, the feature z is extracted by the foundation model, and then normalised as $z = z/\|z\|_2$. Searching for the k-Nearest Neighbour in the training set for a given test sample based on Euclidean distance, the distance score is calculated as

$$R_k(z) = \|z - \hat{z}_k\|_2, \tag{5}$$

where z is a test sample and $\hat{z}_k$ is the corresponding k-Nearest Neighbour in the training set.

3.3. Performance evaluation metrics

Following [9,19], we use the area under receiver operating characteristic curve (AUROC) as our primary evaluation metric. ROC-AUC measures the overall ability of an approach to distinguish between ID and OOD classes, with a higher AUC indicating better performance. For completeness, we also report False Positive Rate at 95% True Positive Rate (FPR @ 95% TPR). This metric is commonly used in general computer vision research as it reveals the probability that an ID sample will be removed when the majority, i.e. 95%, of OOD samples are rejected. In contrast to ROC-AUC, the FPR @ 95% TPR focuses on a performance at a high true positive rate.

3.4. Data

In this work we use Camelyon17 dataset [54] for training, calibration (*calibration* dataset), and evaluation without covariate distribution shifts (*ndl_breast*) of the OOD detection approaches. The dataset contains H&E stained WSIs from sentinel lymph nodes of breast cancer patients collected at 5 medical centres in the Netherlands. For slides digitalisation, 3DHistech Pannoramic Flash II 250, Hamamatsu NanoZoomer-XR C12000-01, and Philips Ultrafast scanners were used at a resolution around 0.24 μm. For the training, we extracted 120k patches from 60 WSIs that were diagnosed as metastasis-free, i.e., 12 WSIs from each of the medical centres. The training dataset did not contain OOD samples. An OOD sample in this works is defined as a patch that has a central pixel belonging to the tumour class. All patches used in the experiments were extracted at approximately 0.5 μm resolution.

When preparing the training, *calibration*, and *ndl_breast* datasets, we ensured that there is no overlap of patients used in the three datasets. For the *calibration*, we randomly selected 18 WSIs covering all 5 medical centres and extracted patches which resulted in around 13k patches for each class: OOD and ID data. The patches for *ndl_breast* were extracted from 34 WSIs also covering all 5 medical centres. This resulted in around 46k OOD and 59k ID patches.

To study the effects of data distribution shifts, we use 3 additional test datasets:

- *swe_breast* data is extracted from 28 WSIs in AIDA BRLN dataset [55] which resulted in around 21k OOD and 24k ID patches. WSIs were scanned at a resolution of approximately 0.5 μml with Aperio ScanScope AT, Hamamatsu NanoZoomer XR, Hamamatsu NanoZoomer S360, and Hamamatsu NanoZoomer S60 scanners. This dataset is further split into two sub-parts depending on the carcinoma subtype, no special type (NST) and lobular carcinoma, to understand the effects of subtypes on the OOD detection performance. Please refer to A.3 for an explanation of the subtypes. This data represents a covariate shift as it is collected in a medical centre in Sweden.
- *ndl_colon* data is extracted from a subset of 59 WSIs from colon lymph node datasets used by [24]. The data comes from Rijnstate Hospital in Arnhem, the Netherlands. We extracted 6k OOD and 10k ID patches for *ndl_colon*. The WSIs were digitised using 3DHistech Pannoramic Flash II 250 at a resolution of 0.24 μm. It represents a shift of cancer metastasis origin without the domain shift, i.e. change of the hospital origin of the data.
- *swe_colon* data is extracted from 48 WSIs in AIDA LNCO dataset [56]. This resulted in around 9k OOD and 14k ID patches. This data represents two distribution shifts: the domain shift as it is collected in a medical centre in Sweden as well as the shift due to an unseen cancer metastasis origin.

4. Experiments

We begin by determining what setups work best with each type of the OOD detection approaches on the *calibration* dataset that does not experience any covariate shifts. To this end, we evaluate the AUROC for different combinations of design choices and select the setup with the highest performance on the *calibration* dataset. Furthermore, we design two experiments to investigate: (a) if the AnoLDM and foundation model-based OOD detection approaches achieve similar performance on the test data without covariate shift; (b) if they have any limitations to generalise to covariate shifts and new types of OOD samples (full-spectrum OOD detection).

4.1. OOD detection with AnoLDM

There are numerous decisions that have an impact on the performance of AnoLDM. For example, as explained in Section 3.1, we need to determine after which noising level to stop the noising process and begin the reconstruction of the image, what metrics should be included in the Z-score, what sampler to use, and how many sampling steps it should take. To determine the optimal configuration, we use the *calibration* dataset which is not utilised during the training of the LDM. In this work, we evaluate:

- Stopping the noising process at 50, 100, 150, …, 800 steps. With an increased amount of noise steps, less signal from the original image is kept, with only noise remaining after some point. Therefore, it is crucial to determine the point where the balance between signal and noise is the most appropriate for OOD detection.
- Number of sampling steps that should be used by the sampler: 5, 10, 50 or 100 sampling steps. This represents a trade-off between computational complexity and reconstruction quality.
- Three image normalisation methods, as explained in Section 3.1: (1) min and max of the two compared images (original input and the reconstructed), (2) min and max values over the whole calibration dataset, and (3) the 1% and 99% percentiles of the calibration data, followed by clamping.
- Five metrics applied in the pixel and latent spaces: MSE, SSIM, MS-SSIM, LPIPS based on AlexNet, and LPIPS based on VGG network.

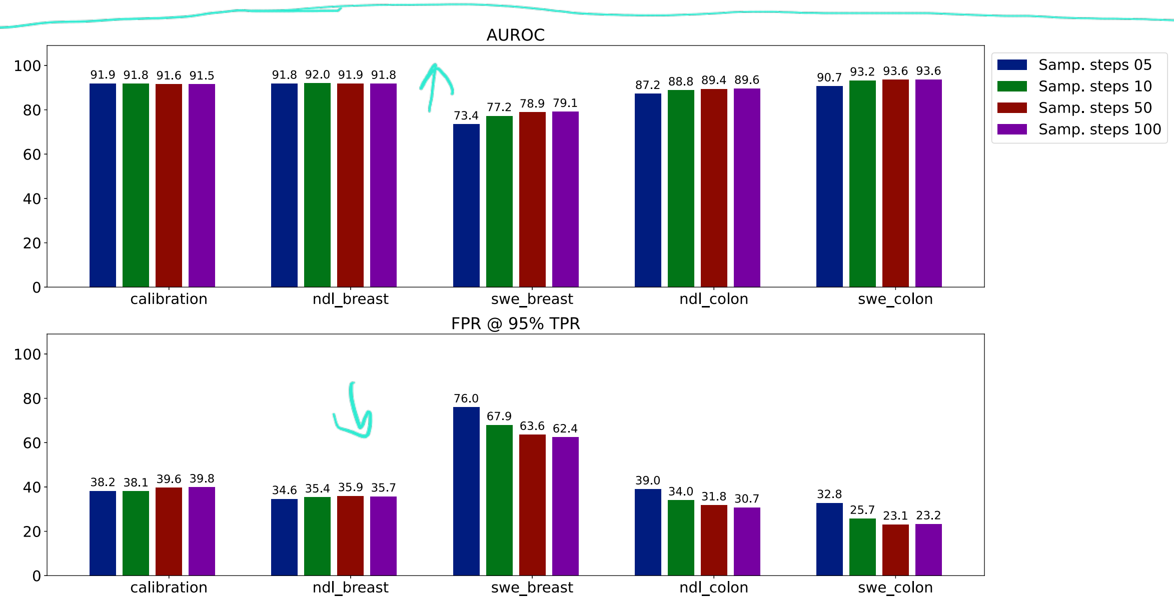

Fig. 3. Results on calibration data and the four test datasets by AnoLDM with different number sampling steps used by the DDIM sampler. The presented values are averages of 2000 bootstrapped samples.

In the initial experiments, we utilise DDIM sampler [44] as it was originally used by [21]. Since reducing the computational complexity of the OOD detection is important, we check if combining fewer sampling steps with some newer samplers, i.e. PLMS [45], UniPC [47] and DPM [46], improves the performance as indicated in the general computer vision research.

4.2. OOD detection with foundation models

We evaluate two commonly used OOD detection models in the computer vision literature: Residual [23] and Deep k-Nearest Neighbours (KNN) [22], as explained in Sections 3.2.2 and 3.2.3, respectively. We test them with 3 publicly available foundation models that all are based on the DINO [57] or DINOv2 framework [58] but utilise different datasets during the training:

- Original DINOv2 model from [58] trained on LVD-142M dataset (142 million natural images). The network architecture used in our experiment is ViT-B/14, the output *cls* token has 768 dimension. We refer to this as *dinov2_residual* or *dinov2_knn* depending on the tested OOD detection method.
- Kang et al. [59] foundation model trained on 32.6 million patches extracted from 46,000 H&E stained WSIs from multiple medical centres. We select the ViT-S/16 model trained by DINO, since this setting produced the best result in Kang et al.'s experiments [59]. Here, the output *cls* token has 384 dimension. We refer to this as *kang_residual* or *kang_knn*.
- Uni foundation model [17] trained on 100 million patches extracted from 100,000 H&E stained WSIs collected at multiple medical centres. We use the pre-trained ViT-H/14 by DINOv2 from [17] in our experiments, where the output *cls* token has 1024 dimension. We refer to this as *uni_residual* or *uni_knn*.

4.3. Performance without covariate shifts

In this experiment, we investigate the performance of the OOD detection methods on the test data that is not affected by any covariate shift. The patches are extracted from Camelyon17 dataset and comes from the 5 medical centres in the Netherlands (see Section 3.4). The ID data is defined as benign lymph node tissue from breast cancer patients and OOD samples are breast cancer metastasis to the lymph nodes.

4.4. Effects of data distribution shifts

In this part, we focus on understanding what effects do data distribution shifts have on OOD detection. Domain shift is a commonly discussed generalisation problem in computational pathology, hence we include lymph nodes of breast cancer patients from another country, i.e. Linköping University Hospital in Sweden. Also, we evaluate the approaches on lymph nodes from colon cancer patients from the hospital in the Netherlands. It is one of the hospitals that contributed data to the training set. The features of the benign class should not change between these patient groups, hence this uncovers how the detection performance works when there are minor differences in the OOD definition. However, the findings by Bejnordi et al. [2] indicate that this may cause generalisation issues for DL models. Finally, we investigate if the combination of domain shift and the change in OOD definition would have a larger negative effect by evaluating the approaches on the lymph nodes of colon cancer patients from the Linköping University Hospital (Sweden). A robust approach should not have substantial problems to successfully handle any of these variations.

5. Results

5.1. AnoLDM

The LDM was trained using the official implementation[1] and with default settings. We used the VQ-VAE autoencoder with a downsampling factor of 4, and trained with batch size 48 and an initial learning rate of $2 \cdot 10^{-6}$. We did not use any augmentations of the training images, as we want the LDM learning the data distribution without alterations imposed by augmentations. We trained the LDM in 59 epochs on the training dataset, which took approximately 65 h on an Nvidia A100 GPU. After training of the LDM, we determined that the optimal setup of LDM for OOD detection is: DDIM sampler with 50 sampling steps, the noising process stopped at 350 diffusion time steps, and the Z-score combining SSIM in the LDM latent space and LPIPS based on VGG network applied in the pixel space. UniPC and DPM produced inferior results, while PLMS performed on par with DDIM. Increasing the number of sampling steps increased the OOD detection

[1] https://github.com/CompVis/latent-diffusion.

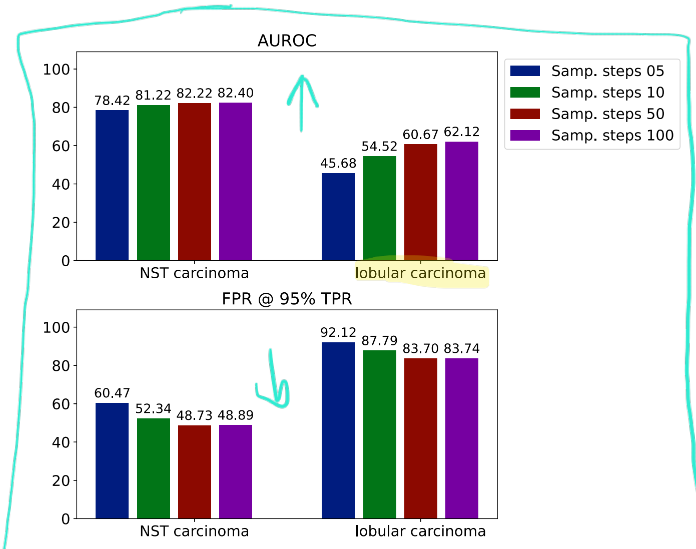

Fig. 4. AnoLDM: AUROC and FPR @ 95% TPR scores computed on No Special Type (NST) and lobular carcinoma subtypes separately in Breast, SWE dataset. The reported values are averages of 2000 bootstrapped samples.

performance up to 50 steps, while more steps gave diminishing returns, see Fig. 3. For the latent space metrics, normalisation using the min and max values over the whole calibration dataset produced slightly better results compared to the two other tested normalisation approaches. While LPIPS performs better in pixel space, other metrics are better suited for the latent space of the LDM, and the best Z-score combines the best pixel space metric (LPIPS using VGG) and the best latent space metric (SSIM). The two metrics applied in pixel and latent spaces, respectively, complement each other well, providing a significant boost in OOD detection performance when combined. Therefore, unless otherwise mentioned, we use these settings in our results (DDIM sampler with 50 steps, 350 noise steps, Z-score based on pixel space VGG and latent space SSIM, with a latent space normalised by min/max over the calibration dataset). For a comparison between samplers, distance measures, and normalisation methods, we refer to Appendix A.1.

Fig. 3 shows the performance of the proposed AnoLDM approach on all datasets used in the evaluation. The first noticeable result is that the performance on calibration and test data from the Netherlands (*calibration* and *ndl_breast* in the figure) is very similar for all sampling steps. However, once the approach is exposed to a data distribution shift (dataset *swe_breast*, *ndl_colon*, and *swe_colon*), the sampling steps need to be increased to 50 to achieve a better performance in terms of both AUROC and FPR @ 95% TPR. There was no substantial improvement observed after further increasing the sampling steps to 100.

An interesting finding is that the AnoLDM approach has most problems in detecting OOD samples in the *swe_breast* dataset. This dataset has an increased amount of lobular carcinoma cases which could be an explanation (around 25% compared to 10% encountered in clinical practice). Fig. 4 shows the performance on two subsets of the *swe_breast*: no special type (NST) and lobular carcinomas. We conclude that AnoLDM clearly struggles to handle lobular carcinoma as an OOD class, however the performance on NST carcinomas is still lower than on the other test datasets.

5.2. Foundation-based approaches

The checkpoints of pre-trained foundation models are obtained from their official implementations. These models are not fine-tuned and are used solely for representation extraction purposes, extraction time for all training images is approximate 2 min on an Nvidia RTX 3080 GPU. The optimal configuration for Residual and KNN approaches was determined to be:

Table 1
AUROC and FPR@95%TPR scores by kang_residual and AnoLDM with the optimal Z-score on 4 test datasets. Reported mean and standard deviation of 2000 bootstrap iterations.

Method	Dataset	AUROC	FPR @ 95%TPR
kang_residual	ndl_breast	96.17 (0.09)	12.41 (0.18)
AnoLDM	ndl_breast	91.86 (0.08)	35.94 (0.42)
kang_residual	swe_breast	85.56 (0.24)	81.91 (0.74)
AnoLDM	swe_breast	78.86 (0.29)	63.59 (0.81)
kang_residual	ndl_colon	87.00 (0.28)	50.7 (1.27)
AnoLDM	ndl_colon	89.43 (0.25)	31.78 (0.9)
kang_residual	swe_colon	91.38 (0.15)	23.72 (0.41)
AnoLDM	swe_colon	93.56 (0.11)	23.07 (0.52)

Residual: In practice, we use the original implementation in [23] as a negative Residual score. We search for the optimal number L of Matrix $\mathbf{R}$ in Eq. (3) on the calibration dataset, for features from different foundation models. The optimal dimension L for *kang_residual*, *uni_residual*, and *dinov2_residual* are 366, 2, and 768, respectively. ?

KNN: The original implementation of [22] also uses a negative distance value. Following [22], we search for the optimal k from {1, 20, 50, 100, 200, 500, 1000, 3000, 5000} on the calibration dataset. The optimal k for *kang_knn*, *uni_knn*, and *dinov2_knn* are 20, 1, and 1, respectively. ?

Fig. 5 shows the results for the foundation-based approaches on the calibration and the four test datasets. There is a clear trend that the *kang* foundation model developed by [59] outperforms other foundation models by a large margin on all datasets. Residual method seems to perform better than KNN method on most datasets, hence we deem it to be the optimal choice.

We observe a similar trend as in the LDM approach: all tested configurations are having most problem at successfully detecting OOD samples in the *swe_breast*, especially when evaluating with the FPR @ 95% TPR metric. When investigating how foundation-based approaches handle NST versus lobular carcinoma in the*swe_breast* dataset (see Fig. 6), we see that *kang_residual* and *kang_knn* have similar performance: a lot smaller drop in AUROC compared to a substantial increase in FPR @ 95% TPR (which indicates a bad performance at the high TPR threshold).

5.3. Comparison of the approaches

Table 1 shows the results by the optimal configuration of the AnoLDM and Residual method combined with the *kang* foundation model. We conclude that even though *kang_residual* outperforms the AnoLDM on the test data that has no data distribution shifts (*ndl_breast*), the changes in data distribution affects both approaches. Notably, AnoLDM performs substantially better on the *swe_colon* than on *ndl_breast* dataset, while *kang_residual* is negatively impacted by this shift.

None of the compared approaches can successfully handle the covariate shift in *swe_breast* dataset. Even though kang_residual has a slightly smaller drop in AUROC on *swe_breast* dataset, i.e. 10.6 compared to 13 percentage points, it has a lot worse FPR @ 95% TPR score. Fig. 7 shows the histogram of OOD scores on *swe_breast* by the two methods. We can see that neither method has good separation between the OOD and ID classes. Interestingly, *kang_residual* has three peaks for ID samples. The peaks do not seem to correlate with the carcinoma subtypes (see Appendix A.2).

6. Discussion

The aim of this work is to determine if foundation-based OOD detection outperforms reconstruction-based OOD detection, which currently

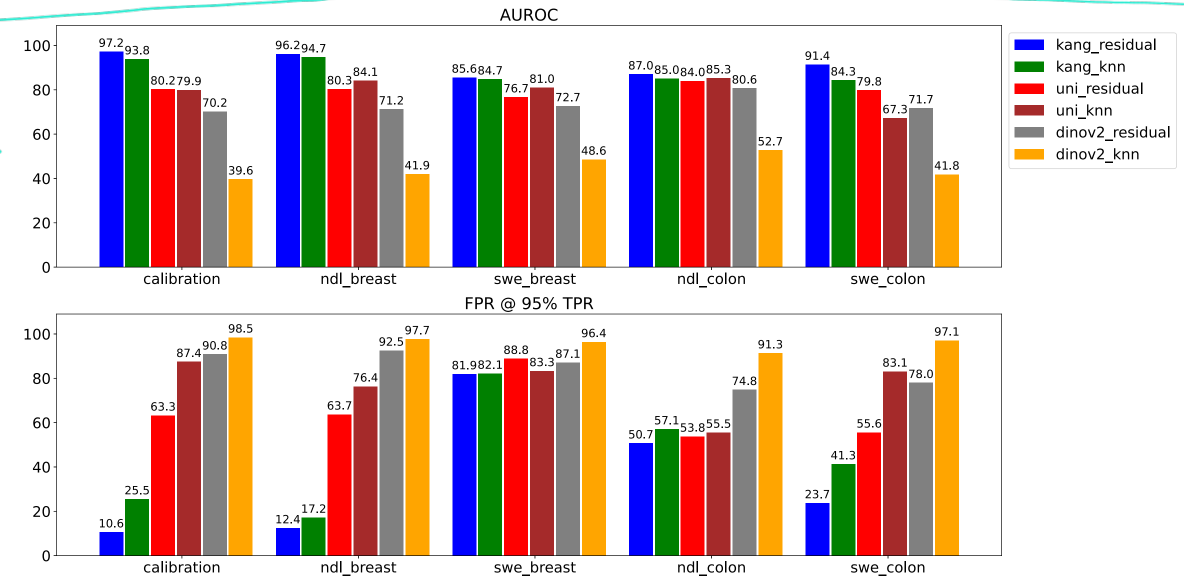

Fig. 5. Results on calibration data and the four test datasets by Residual and KNN approaches combined with different foundation models. The presented values are averages of 2000 bootstrapped samples.

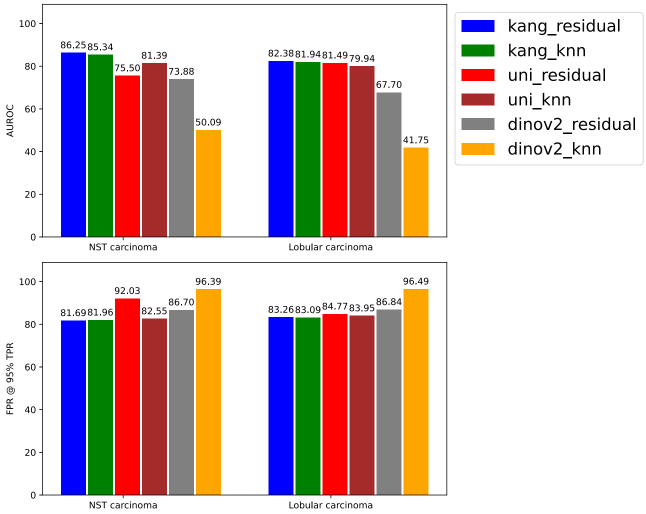

Fig. 6. Residual and KNN: AUROC and FPR @ 95% TPR scores computed on No Special Type (NST) and lobular carcinoma subtypes separately in Breast, SWE dataset. The reported values are averages of 2000 bootstrapped samples.

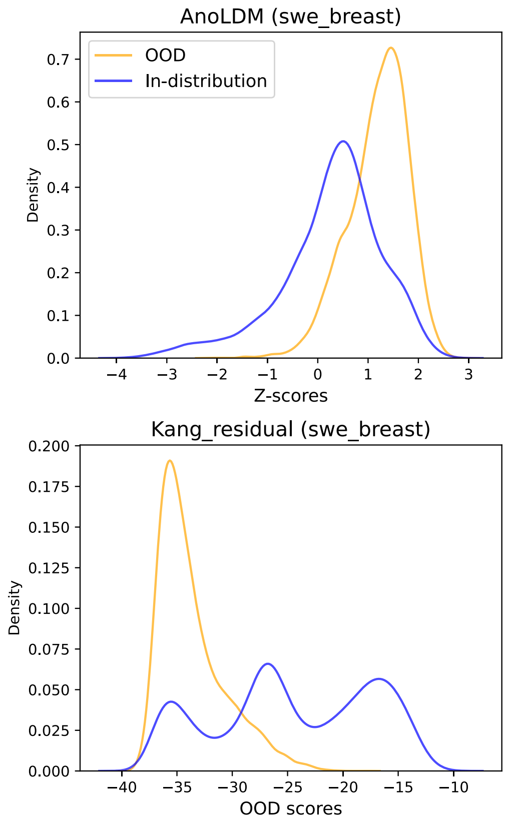

Fig. 7. Histogram of OOD scores computed by AnoLDM and kang_residual approaches on swe_breast dataset. An OOD sample is indicated by having a high AnoLDM Z-score but a low kang_residual score.

is the SOTA in the computational pathology community. In this section we discuss the results and their implications as well as the limitations of our work and the future research directions.

We chose to use the latent diffusion model (LDM) in the reconstruction-based approach in contrast to the DDPM used in the previous work [9,19]. AnoLDM results are comparable or exceed the SOTA performance achieved with DDPM in computational pathology in [9]. Noteworthy is that [9] have considered two different OOD definitions in their experiments: (1) patches that are 100% made of tumour cells (their approach achieved AUROC of 93.33 in this case) and (2) patches with tumour cell in the centre pixel (achieved AUROC of 85.87). We used their more difficult definition of an OOD sample: a patch that has a centre pixel belonging to a tumour metastasis class. In our experiments, the proposed AnoLDM achieved an AUROC of 91.86 on the test data unaffected by the distribution shifts. We used the Camelyon17 dataset in our experiments while [9] used an older Camelyon16 version, hence the results show an indication but cannot be compared directly. Nevertheless, we conclude that one can safely replace computationally expensive DDPM with a cheaper LDM in the reconstruction-based OOD detection on digital pathology data. This is consistent with the findings in the computer vision research [21].

We evaluated two commonly used OOD detection approaches with foundation models in general computer vision: Residual part from ViM [23] and KNN [22]. Due to the design of our experiments, it is infeasible to train a classifier on our ID dataset because it only has one class. Therefore, we are not able to use most SOTA methods, such as MSP [60], ODIN [61], Virtual-logit Matching [23], and Generalised Entropy [38], as they require a multi-class training setup. We also tested three pre-trained foundation models, *kang* [59], *uni* [17], and *dinov2* [58]. Our conclusion is that it is significantly more important which foundation model is used to extract the features than which OOD detection technique is chosen to convert the extracted features to OOD scores (see Fig. 5). A somewhat unexpected finding is that the *kang* foundation model which was trained on fewer WSIs outperformed the newer *uni* foundation model that was trained on a double amount of WSIs. This highlights that the amount of data is not the only factor that determines the success and usefulness of a foundation model. Interestingly, a similar conclusion is made by the authors of *uni* [17]. As expected, a *dinov2* foundation model trained on a broad range of natural images performed the worst in our experiments. The unique features that are important in performing pathology tasks do not seem to be learnt by a general-purpose foundation model.

Has Residual with *kang* foundation model (kang_residual) surpassed the performance of AnoLDM? There does not seem to be a straightforward answer to this. If data distribution shifts did not exist in a clinical practice, *kang_residual* would be a clear winner as it achieved 4.3 percentage points higher AUROC and an impressive reduction by almost a factor of 3 in FPR @ 95% TPR score (see Table 1). However, the experiments reveal that AnoLDM is more resilient to minor changes in OOD definition as it performs better on both colon datasets. In fact, AnoLDM achieved its best performance on the *swe_colon* when comparing to other test datasets. An explanation could be that colon cancer is predominantly an adenocarcinoma which originates from the same type of cells as the NST carcinomas in the breast cancer. One could argue that AnoLDM has learnt to differentiate adenocarcinomas better than *kang_residual*. However, further experiments need to be conducted to understand why the same benefits are not observed on the colon data from the same hospital as the training data, i.e. *ndl_colon*.

The *swe_breast* dataset has caused most problems for both approaches. This indicates that the covariate shift is a real challenge to the current SOTA approaches. A better AUROC score is achieved by *kang_residual* approach than by AnoLDM on this dataset: 85.56 compared to 78.86 (see Table 1), though it is still a significant drop in performance compared to the 96.17 AUROC on the *ndl_breast*. However, *kang_residual* has a much worse FPR @ 95% TPR score compared to AnoLDM (81.91 versus 63.59). This indicates that the *kang_residual* is better at assigning high OOD-likelihood to OOD samples with a better degree of separation across most thresholds, but at the threshold that results in a very high TPR, *kang_residual* is also incorrectly classifying a larger number of ID instances as OOD.

The sharp drop in performance on *swe_breast* dataset could be a result of two factors: (1) increased frequency of lobular carcinoma subtype in the *swe_breast*, and (2) superparamagnetic iron oxide tracer used in the lymph node dissections in Sweden. In *swe_breast* there is an increased quantity of lobular carcinoma subtype which here constitutes 25% of the cases compared to the 10% of cases normally encountered in clinical practice. Lobular carcinoma is more challenging to diagnose for both pathologists and AI due to its morphology [62], and this also seems to be true for the OOD detection approaches. For a more in-depth explanation of the differences between NST and lobular carcinoma that could contribute to this challenge, we refer to Appendix A.3.

The clinical routine to use superparamagnetic iron oxide tracer in the sentinel lymph node dissections in Sweden is likely another cause of the strong domain shift observed in our study. It enables to detect to which lymph node the breast tumour is draining. However, it is known to add some brown staining to the resulting WSIs [63]. Such tracer is not used in the surgical removal of the sentinel lymph nodes in the *ndl_breast* dataset. It is somewhat disappointing that the foundation models that are trained on large datasets from multiple centres still cannot handle the covariate shift in this dataset.

The histograms of OOD scores produced by AnoLDM and *kang_residual* on the *swe_breast* dataset reveal that the separation between the OOD and ID samples is somewhat better for *kang_residual* approach which is consistent with the higher AUROC score achieved by this approach (see Fig. 7). However, we could not find a viable explanation in context of pathology why the histogram for ID samples by *kang_residual* has the three peaks. Results in Appendix A.2 negate the hypothesis that this could be related to the cancer subtypes. We conclude that neither of the approaches can reliably handle the OOD samples in the data from this hospital.

One of the limitations of this work is the choice of OOD definition. It serves well to get an indication of the OOD detection approaches' capabilities and uncover their limitations, but the next step should be to design experiments that handle clinically important OOD scenarios. For example, some conditions are difficult to diagnose on breast core needle biopsies of suspected breast cancer patients [64]. Even though AI has shown some promising results [65], achieving reliable performance on all rare conditions can still be challenging. OOD detection could be one way to address this problem. Finally, if the success of the proposed OOD detection is verified in different scenarios in a research setting, a clinical validation study should be conducted to establish the real impact that this technique could have on the reliability of an AI solution in a clinical setup.

In this work, we evaluated three publicly available foundation models including the recent SOTA *uni* model. However, research on foundation models for computational pathology is an active area and inevitably new models are soon to come. We cannot exclude the option that some of these may handle data shifts better. For example, if organ-specific foundation models become available, they may have capacity to learn the nuances of distribution shifts better than general foundation models for pathology. Furthermore, due to our experimental setup, we only had one ID class, hence we were not able to test many of the existing OOD detection methods using foundation models that are available in the computer vision literature, as these require multiple ID classes. Future work could focus on multi-class ID setups that would enable to evaluate other SOTA approaches.

7. Conclusion

In conclusion, in this work we proposed and evaluated OOD detection approaches for computational pathology based on reconstruction of the images as well as foundation-based models in feature space. For efficient use of diffusion models for the purpose of reconstruction-based OOD detection, we adapted the latent diffusion model (LDM) by considering options such as noise level, sampler, and combinations of distance measures in both latent and pixel spaces. We compared this against post-hoc methods applied in the feature space of foundation models. While these have been demonstrated to achieve SOTA on natural images, we provided a first evaluation in OOD detection for digital pathology by considering different choices of foundation models for extracting representations. We tested the different approaches under different variations of both semantic and covariate shifts, providing an evaluation of OOD detection in the full-spectrum setting for digital pathology.

We found that on the test data without any covariate shifts, foundation-based approaches outperform the diffusion-based approach. However, the latter seems to be handling the covariate shifts slightly better than the foundation-based methods, providing a more robust alternative for full-spectrum OOD detection in digital pathology.

CRediT authorship contribution statement

Milda Pocevičiūtė: Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yifan Ding:** Writing – review & editing, Software, Methodology. **Ruben Bromée:** Software, Investigation. **Gabriel Eilertsen:** Writing – review & editing, Software, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding

The funding used in this study had no role in deciding the study design, collection, analysis and interpretation of data, writing of the report and decision to submit the article for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109327.

Data availability

All datasets used in this work are fully anonymised. Camelyon17, AIDA LNCO, and AIDA BRLN datasets are publicly available. Colon dataset from Rijnstate Hospital in Arnhem, the Netherlands was obtained by data transfer agreement, but we do not have permission to share this data. Therefore, in compliance with the relevant laws and institutional guidelines ethical approval was not required.

References

[1] A.H. Song, G. Jaume, D.F. Williamson, M.Y. Lu, A. Vaidya, T.R. Miller, F. Mahmood, Artificial intelligence for digital and computational pathology, Nat. Rev. Bioeng. 1 (12) (2023) 930–949.

[2] B.E. Bejnordi, M. Veta, P.J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J.A. Van Der Laak, M. Hermsen, Q.F. Manson, M. Balkenhol, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, Jama 318 (22) (2017) 2199–2210.

[3] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K.J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Med. 25 (8) (2019) 1301–1309.

[4] J. Sandbank, G. Bataillon, A. Nudelman, I. Krasnitsky, R. Mikulinsky, L. Bien, L. Thibault, A. Albrecht Shach, G. Sebag, D.P. Clark, et al., Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies, NPJ Breast Cancer 8 (1) (2022) 129.

[5] P. Raciti, J. Sue, J.A. Retamero, R. Ceballos, R. Godrich, J.D. Kunz, A. Casson, D. Thiagarajan, Z. Ebrahimzadeh, J. Viret, et al., Clinical validation of artificial intelligence–augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection, Arch. Pathol. Lab. Med. 147 (10) (2023) 1178–1185.

[6] M. Pocevičiūtė, Generalisation and Reliability of Deep Learning for Digital Pathology in a Clinical Setting (Ph.D. thesis), Linköping University Electronic Press, 2023.

[7] M. Jahanifar, M. Raza, K. Xu, T. Vuong, R. Jewsbury, A. Shephard, N. Zamanitajeddin, J.T. Kwak, S.E.A. Raza, F. Minhas, et al., Domain generalization in computational pathology: survey and guidelines, 2023, arXiv preprint arXiv:2310.19656.

[8] M. Pocevičiūtė, G. Eilertsen, C. Lundström, Unsupervised anomaly detection in digital pathology using GANs, in: 2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI, IEEE, 2021, pp. 1878–1882.

[9] J. Linmans, G. Raya, J. van der Laak, G. Litjens, Diffusion models for out-of-distribution detection in digital pathology, Med. Image Anal. 93 (2024) 103088.

[10] P. Abolfath Beygi Dezfouli, Uncertainty Estimation of Weakly Supervised Predictive Models for Out-Of-Distribution Detection in Digital Pathology (Ph.D. thesis), University of British Columbia, 2024.

[11] R. Yamashita, J. Long, S. Banda, J. Shen, D.L. Rubin, Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation, IEEE Trans. Med. Imaging 40 (12) (2021) 3945–3954.

[12] M.J. Hetz, T.-C. Bucher, T.J. Brinker, Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images, Med. Image Anal. 94 (2024) 103149.

[13] M. Pocevičiūtė, G. Eilertsen, S. Garvin, C. Lundström, Detecting domain shift in multiple instance learning for digital pathology using fréchet domain distance, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 157–167.

[14] K. Stacke, G. Eilertsen, J. Unger, C. Lundström, Measuring domain shift for deep learning in histopathology, IEEE J. Biomed. Health Inform. 25 (2) (2020) 325–336.

[15] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, S. Liu, P. Mathieu, A. van Eck, D. Lee, J. Viret, et al., Virchow: A million-slide digital pathology foundation model, 2023, arXiv preprint arXiv:2309.07778.

[16] J. Dippel, B. Feulner, T. Winterhoff, S. Schallenberg, G. Dernbach, A. Kunft, S. Tietz, P. Jurmeister, D. Horst, L. Ruff, et al., Rudolfv: A foundation model by pathologists for pathologists, 2024, arXiv preprint arXiv:2401.04079.

[17] R.J. Chen, T. Ding, M.Y. Lu, D.F. Williamson, G. Jaume, A.H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al., Towards a general-purpose foundation model for computational pathology, Nature Med. 30 (3) (2024) 850–862.

[18] F.C. Borlino, L. Lu, T. Tommasi, Foundation models and fine-tuning: A benchmark for out of distribution detection, IEEE Access (2024).

[19] M.S. Graham, W.H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, J. Cardoso, Denoising diffusion models for out-of-distribution detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2947–2956.

[20] A.V. Sadr, R. Bülow, S. von Stillfried, N.E. Schmitz, P. Pilva, D.L. Hölscher, P.P. Ha, M. Schweiker, P. Boor, Operational greenhouse-gas emissions of deep learning in digital pathology: a modelling study, Lancet Digit. Health 6 (1) (2024) e58–e69.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[22] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-distribution detection with deep nearest neighbors, in: International Conference on Machine Learning, PMLR, 2022, pp. 20827–20840.

[23] H. Wang, Z. Li, L. Feng, W. Zhang, Vim: Out-of-distribution with virtual-logit matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4921–4930.

[24] P. Bándi, M. Balkenhol, M. Van Dijk, M. Kok, B. van Ginneken, J. van der Laak, G. Litjens, Continual learning strategies for cancer-independent detection of lymph node metastases, Med. Image Anal. 85 (2023) 102755.

[25] Z. Hong, Y. Yue, Y. Chen, H. Lin, Y. Luo, M.H. Wang, W. Wang, J. Xu, X. Yang, Z. Li, et al., Out-of-distribution detection in medical image analysis: A survey, 2024, arXiv preprint arXiv:2404.18279.

[26] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, K. Zhou, W. Zhang, Y. Li, Z. Liu, Y. Chen, H. Li, OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection, in: NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models, 2024, URL: https://openreview.net/forum?id=vTapqwaTSi.

[27] J. Yang, K. Zhou, Z. Liu, Full-spectrum out-of-distribution detection, Int. J. Comput. Vis. 131 (10) (2023) 2607–2622.

[28] B. Bozorgtabar, G. Vray, D. Mahapatra, J.-P. Thiran, Sood: Self-supervised out-of-distribution detection under domain shift for multi-class colorectal cancer tissue types, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3324–3333.

[29] J. Ren, P.J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, B. Lakshminarayanan, Likelihood ratios for out-of-distribution detection, Adv. Neural Inf. Process. Syst. 32 (2019).

[30] E. Nalisnick, A. Matsukawa, Y.W. Teh, B. Lakshminarayanan, Detecting out-of-distribution inputs to deep generative models using typicality, 2019, arXiv preprint arXiv:1906.02994.

[31] E. Nalisnick, A. Matsukawa, Y.W. Teh, D. Gorur, B. Lakshminarayanan, Do deep generative models know what they don't know? in: International Conference on Learning Representations, 2019, URL: https://openreview.net/forum?id=H1xwNhCcYm.

[32] P. Kirichenko, P. Izmailov, A.G. Wilson, Why normalizing flows fail to detect out-of-distribution data, Adv. Neural Inf. Process. Syst. 33 (2020) 20578–20589.

[33] T. Schlegl, P. Seeböck, S.M. Waldstein, G. Langs, U. Schmidt-Erfurth, F-anogan: Fast unsupervised anomaly detection with generative adversarial networks, Med. Image Anal. 54 (2019) 30–44.

[34] A. Berg, M. Felsberg, J. Ahlberg, Unsupervised adversarial learning of anomaly detection in the wild, in: ECAI 2020, IOS Press, 2020, pp. 1002–1008.

[35] J. Wyatt, A. Leach, S.M. Schmon, C.G. Willcocks, Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 650–656.

[36] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, Y. Li, Mitigating neural network overconfidence with logit normalization, in: International Conference on Machine Learning, PMLR, 2022, pp. 23631–23644.

[37] Y. Sun, C. Guo, Y. Li, React: Out-of-distribution detection with rectified activations, Adv. Neural Inf. Process. Syst. 34 (2021) 144–157.

[38] X. Liu, Y. Lochman, C. Zach, Gen: Pushing the limits of softmax-based out-of-distribution detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23946–23955.

[39] A. Ahmadian, Y. Ding, G. Eilertsen, F. Lindsten, Unsupervised novelty detection in pretrained representation space with locally adapted likelihood ratio, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 874–882.

[40] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[41] J. Linmans, J. van der Laak, G. Litjens, Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks, in: MIDL, 2020, pp. 465–478.

[42] J. Linmans, S. Elfwing, J. van der Laak, G. Litjens, Predictive uncertainty estimation for out-of-distribution detection in digital pathology, Med. Image Anal. 83 (2023) 102655.

[43] M. Pocevičiūtė, G. Eilertsen, S. Jarkman, C. Lundström, Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology, Sci. Rep. 12 (1) (2022) 8329.

[44] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, 2020, arXiv preprint arXiv:2010.02502.

[45] L. Liu, Y. Ren, Z. Lin, Z. Zhao, Pseudo numerical methods for diffusion models on manifolds, 2022, arXiv preprint arXiv:2202.09778.

[46] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu, Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, Adv. Neural Inf. Process. Syst. 35 (2022) 5775–5787.

[47] W. Zhao, L. Bai, Y. Rao, J. Zhou, J. Lu, Unipc: A unified predictor-corrector framework for fast sampling of diffusion models, Adv. Neural Inf. Process. Syst. 36 (2024).

[48] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2, Ieee, 2003, pp. 1398–1402.

[49] Z. Wang, A.C. Bovik, H.R. Sheikh, Structural similarity based image quality assessment, in: Digital Video Image Quality and Perceptual Coding, CRC Press, 2017, pp. 225–242.

[50] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

[51] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, 2019, arXiv:1912.01703.

[53] pytorch-msssim python package, 2019, https://github.com/jorge-pessoa/pytorch-msssim. (Accessed 03 September 2024).

[54] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. Van de Loo, R. Vogels, et al., 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset, GigaScience 7 (6) (2018) giy065.

[55] S. Jarkman, M. Lindvall, J. Hedlund, D. Treanor, C. Lundström, J. van der Laak, Axillary lymph nodes in breast cancer cases, 2019.

[56] G. Maras, M. Lindvall, C. Lundström, Regional lymph node metastasis in colon adenocarcinoma, second collection series, 2020.

[57] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[58] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, 2023, arXiv preprint arXiv:2304.07193.

[59] M. Kang, H. Song, S. Park, D. Yoo, S. Pereira, Benchmarking self-supervised learning on diverse pathology datasets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3344–3354.

[60] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2016, arXiv preprint arXiv:1610.02136.

[61] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, 2017, arXiv preprint arXiv:1706.02690.

[62] S. Jarkman, et al., Generalization of deep learning in digital pathology: Experience in breast cancer metastasis detection, Cancers 14 (21) (2022) 5424.

[63] N. Mirzaei, F. Wärnberg, P. Zaar, H. Leonhardt, R. Olofsson Bagge, Ultra-low dose of superparamagnetic iron oxide nanoparticles for sentinel lymph node detection in patients with breast cancer, Ann. Surg. Oncol. 30 (9) (2023) 5685–5689.

[64] C. Quinn, A. Maguire, E. Rakha, Pitfalls in breast pathology, Histopathology 82 (1) (2023) 140–161.

[65] N. Cadavid-Fernández, I. Carretero-Barrio, E. Moreno-Moreno, A. Rodríguez-Villena, J. Palacios, B. Pérez-Mies, The role of core needle biopsy in diagnostic breast pathology, Rev. Senol. Patología Mamaria 35 (2022) S3–S12.