

MeDi: Metadata-Guided Diffusion Models for Mitigating Biases in Tumor Classification

David Jacob Drexlin^{1,2}, Jonas Dippel^{1,3,4}, Julius Hense^{1,3}, Niklas Prenißl⁵, Grégoire Montavon^{3,6}, Frederick Klauschen^{3,5,7,8,9}, and Klaus-Robert Müller^{1,3,10,11}

¹ Machine Learning Group, Technische Universität Berlin, Germany

² The Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA), Germany

³ BIFOLD – Berlin Institute for the Foundations of Learning and Data, Germany

⁴ Aignostics, Germany

⁵ Institute of Pathology, Charité – Universitätsmedizin Berlin, Germany

⁶ Charité – Universitätsmedizin Berlin, Germany

⁷ Institute of Pathology, Ludwig-Maximilians-Universität München, Germany

⁸ German Cancer Research Center (DKFZ) & German Cancer Consortium (DKTK), Berlin & Munich Partner Sites

⁹ Bavarian Cancer Research Center (BZKF), Germany

¹⁰ Department of Artificial Intelligence, Korea University, Republic of Korea

¹¹ Max-Planck Institute for Informatics, Germany
{j.dippel, klaus-robert.mueller}@tu-berlin.de

Abstract. Deep learning models have made significant advances in histological prediction tasks in recent years. However, for adaptation in clinical practice, their lack of robustness to varying conditions such as staining, scanner, hospital, and demographics is still a limiting factor: if trained on overrepresented subpopulations, models regularly struggle with less frequent patterns, leading to shortcut learning and biased predictions. Large-scale foundation models have not fully eliminated this issue. Therefore, we propose a novel approach explicitly modeling such metadata into a **Metadata-guided generative Diffusion** model framework (MeDi). MeDi allows for a targeted augmentation of underrepresented subpopulations with synthetic data, which balances limited training data and mitigates biases in downstream models. We experimentally show that MeDi generates high-quality histopathology images for unseen subpopulations in TCGA, boosts the overall fidelity of the generated images, and enables improvements in performance for downstream classifiers on datasets with subpopulation shifts. Our work is a proof-of-concept towards better mitigating data biases with generative models.

Keywords: Generative Models · Histopathology · Shortcut Learning

1 Introduction

Histopathology is a cornerstone of clinical diagnostics and biomedical research, providing critical insights into disease mechanisms through tissue examination.

AI models have shown great potential to advance pathological workflows in recent years [17]. Research studies have shown that deep learning models can predict disease subtypes [3], disease progression and outcome [6], detect rare cancers [11] and even molecular properties such as mutations [8] and gene expressions [14] to a certain extent. Although models can reach or occasionally even surpass pathologist-grade performance in specific subtasks, they are not widely adopted in the clinical workflow yet [11, 17]. The varying conditions across medical centers such as different staining protocols, lab artifacts, scanner type and demographic differences pose unique challenges to machine learning algorithms [17]. If such tissue conditions are correlated with prediction targets, models use them as a shortcut in the training process and fail to generalize [21, 12, 13, 26]. Foundation models that are trained with self-supervised learning on large-scale histopathology datasets are proposed as a solution [5, 27, 10, 1]. However, as they not only model the biological features but also these metadata conditions, they can still fail catastrophically when the distribution is highly skewed [16, 18, 15, 26]. Therefore, additional mitigation strategies are needed to prevent such biases and ensure a safe deployment of machine learning models in the clinic.

We propose to explicitly model these conditions by training **metadata-guided generative diffusion models** (MeDi). Instead of conditioning a model only on the disease subclass, we also condition the generation process on readily available metadata attributes, such as the medical center, to be more sensitive to relevant statistical and morphological variations. Furthermore, the MeDi framework provides interpretable controls to balance out the training data distribution during downstream tasks. By contrast, when relying on a single class label, generative models may gravitate toward the “mode” of each class distribution, overlooking minority subpatterns/classes and failing to meaningfully increase diversity in the dataset. We demonstrate that (1) MeDi produces higher-quality images compared to class-only conditioning, and (2) can be used to improve generalization capabilities to unseen subpopulations in downstream tasks. The full MeDi pipeline (training, sampling, embedding, and evaluation scripts) is available on GitHub¹.

2 Related Work

Generative approaches for mitigating domain shifts in medical imaging have received considerable attention in recent years. Early attempts in histopathology often relied on stain normalization [22, 25] or style-transfer techniques (e.g., CycleGAN-based color normalization) to reduce color variability across laboratories, scanners, and staining protocols [7]. More recent advances in diffusion models have opened up new opportunities for controllable and high-fidelity image synthesis [9]. For example, Niehues et al. [23] demonstrated that both fine-tuned Stable Diffusion and GAN-based models can be used to augment datasets in colorectal cancer tasks, achieving improved performance on the CRC-100K dataset. Aversa et al. [2] extended histological image generation to larger regions

¹ <https://github.com/David-Drexlin/MeDi>

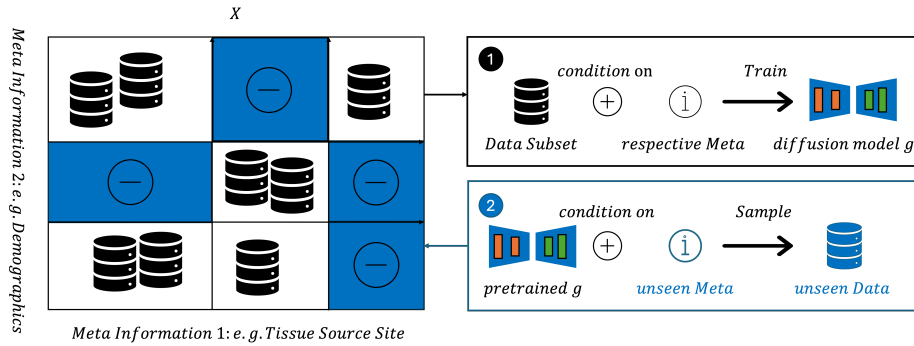


Fig. 1. MeDi Training and Inference Framework. During training (1), the model receives real images along with their class labels and metadata. At inference (2), users can condition on arbitrary combinations of class and metadata, enabling the generation of synthetic images for underrepresented subpopulations or unseen subpopulations altogether.

with conditioning on class-based segmentation masks, enabling the generation of synthetic whole slide images with pixelwise class annotations.

Beyond class annotations, other conditioning mechanisms in diffusion models have recently been explored. Osorio et al. [24] introduced a latent diffusion framework that leverages cluster-defined morphological features to enhance the fidelity of generated images. Carillo et al. [4] used bulk RNA sequencing data as a conditioning mechanism. Closest to our own approach, Ktena et al. [20] demonstrated that diffusion-based augmentation can improve model fairness and robustness in radiology, dermatology, and histopathology. However, their histopathology experiments were limited to breast cancer tissues from the Camelyon17 dataset which offers limited classes (2) and metadata variability (5 hospitals).

3 Method

In this work, we propose a framework for training metadata-conditioned diffusion models (MeDi) that integrates both class labels (e.g., cancer type) and relevant metadata (e.g., medical center, demographics) into the diffusion process. Conventional class conditioning may capture dominant variations within each class; however, it can fail to model underrepresented or subtle domain shifts. By explicitly modeling additional metadata, MeDi aims to mitigate batch effects in downstream tasks and enhance the diversity of synthetic samples.

Figure 1 illustrates the overall framework of MeDi. During training, each image is paired with: a class label (e.g., cancer subtype) and metadata attributes. The model learns the interactions between these attributes and the image features. At inference time, the user can condition the generation process on any valid combination of class and metadata, thereby synthesizing images for underrepresented or unseen subpopulations.

Essentially, our approach can interpolate within the metadata space to balance the distribution of existing meta-class combinations, or extrapolate to produce synthetic samples for combinations absent in the training set. This data-centric approach aims to "fill in" distribution gaps and help mitigate biases resulting from skewed datasets. For instance, consider two hospitals: Hospital A, which submitted lung adenocarcinoma and lung squamous cell carcinoma samples from overwhelmingly Caucasian patients, and Hospital B, which mostly has samples from African-American patients but exclusively submitted lung squamous cell carcinoma samples. By leveraging metadata-guided diffusion, MeDi can learn to generate realistic synthetic samples for patients with lung adenocarcinoma in Hospital B, and thus alleviate the imbalance across different unseen subpopulations that are known to bias classifiers [26].

By balancing these subpopulations, MeDi helps remove spurious correlations or learnable proxies between metadata and the disease label, reducing Clever Hans effects [21, 16] in downstream tasks.

3.1 Metadata-Conditioned Diffusion UNet

We build upon a standard 2D UNet architecture from the *diffusers* library², which progressively denoises an image across multiple residual downsampling and upsampling blocks.

Metadata Embeddings. Let $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ denote the set of metadata attributes associated with each patch. For each *categorical* attribute, we employ a learnable embedding layer that maps each discrete category to a fixed-dimensional vector of dimension d_e . For instance, if the attribute "medical center" has N_{site} possible values, then each site i is represented by an embedding vector $\mathbf{z}_{\text{site}}(i) \in \mathbb{R}^{d_e}$.

Metadata Combination. We concatenate the class label embedding with all metadata embeddings to form a single *conditioning vector*. Let $\mathbf{z}_{\text{class}} \in \mathbb{R}^{d_{\text{class}}}$ denote the class embedding and let $\mathbf{z}_{\text{meta},i} \in \mathbb{R}^{d_e}$ denote the embedding of the i th metadata attribute for $i = 1, \dots, k$. To ensure compatibility with the timestep embedding $\mathbf{z}_t \in \mathbb{R}^{d_t}$, we require that $d_{\text{class}} + k \cdot d_e = d_t$. The conditioning vector is then defined as

$$\mathbf{z}_{\text{cond}} = \text{concat}\left(\mathbf{z}_{\text{class}}, \mathbf{z}_{\text{meta},1}, \dots, \mathbf{z}_{\text{meta},k}\right) \in \mathbb{R}^{d_t}.$$

This vector is added to the timestep embedding:

$$\mathbf{z}_{\text{final}} = \mathbf{z}_t + \mathbf{z}_{\text{cond}},$$

which is provided to the UNet’s residual blocks at each down- and upsampling stage. In this way, the diffusion process is conditioned on both the class-specific morphology (e.g., tumor [sub]type) and the domain-specific variations (e.g., scanner artifacts, demographic effects).

² <https://github.com/huggingface/diffusers>

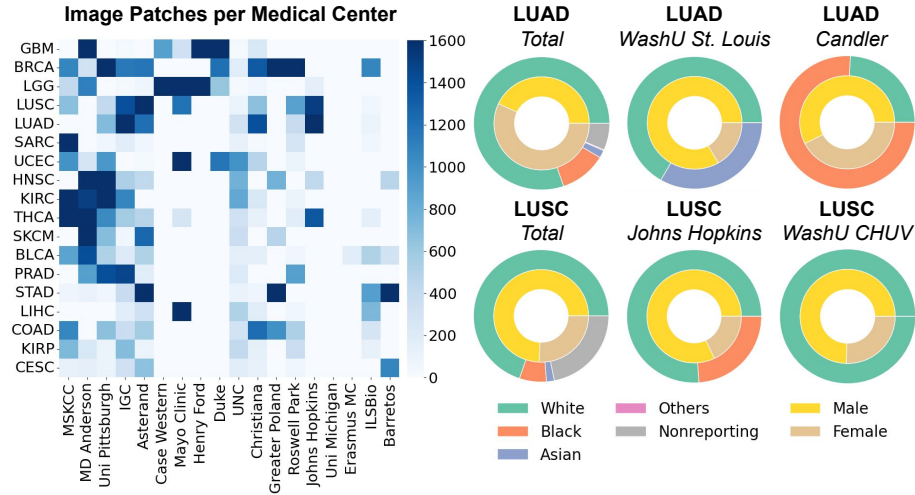


Fig. 2. Dataset statistics for TCGA-UT. *Left:* The number of patches for the 18 largest cancer types (vertical axis) and tissue source sites (horizontal axis), capped at 1,600. *Right:* Race and gender distributions for the patches of selected cancer types and tissue source sites. The dataset is highly imbalanced across classes, hospitals, and demographics, with many missing or underrepresented metadata combinations.

4 Experiments

We evaluate our metadata-guided diffusion framework (MeDi) on the TCGA-UT dataset³ [19], which provides histopathology tumor patches from 32 distinct cancer types within The Cancer Genome Atlas (TCGA). Each patch is associated with several metadata fields, including *tissue source site (TSS)*, *race*, *gender*, and *age*. In total, the dataset covers 8,726 patient-level records and 271,710 image patches. In terms of metadata dimensions, the dataset contains 626 TSS codes from 184 unique medical centers. This yields a high combinatorial space of $626 \times 32 = 20,032$, which is only partially represented. For instance, some cancer types contain fewer than 1,000 patches, while others exceed 20,000; certain TSS are overrepresented, whereas others appear only sporadically.

4.1 Dataset Splitting and Training Setup

To induce an out-of-distribution scenario for the classification tasks, we design a holdout strategy that excludes specific medical center and demographic combinations from diffusion model training. We remove certain tissue–demographic configurations from training such that no data leakage can occur through synthetic images. Specifically, for each cancer type (e.g., TCGA-CESC), we randomly

³ zenodo.org/records/5889558, License: Non-Commercial Use: CC-BY-NC-SA 4.0

select 30% of the medical center and patient race combinations for the holdout set. All patches matching those combinations are excluded from training, leaving 7,061 patient records in the training split and 1,665 in the holdout. This ensures that particular site–demographic subpopulations remain completely unseen during model training.

4.2 Diffusion Model Variants

We train two distinct diffusion models on the above training set, each for 800,000 optimization steps at a learning rate of 10^{-4} with a batch size of 64:

1. **Baseline: Class-Only (CLS).** A UNet conditioned solely on the cancer-type label.
2. **MeDi: Class + Tissue Source Site (CLS+TSS).** A UNet conditioned on the cancer-type label and the medical center (TSS code). Tissue source site is embedded via a learnable layer.

Both models share the same network capacity and training hyperparameters; they differ only in the presence or absence of the metadata embeddings. We picked the medical center as a proof of concept as it has shown to significantly bias classifiers in past studies [13, 18, 15].

4.3 Training Set Fidelity (FID)

We first assess how effectively each model reproduces the *training distribution* by measuring the Fréchet Inception Distance (FID) between real and synthetic images. Lower FID scores indicate that synthetic images better match the real data, however, they do not quantify the clinical realism of the synthetic data.

Setup. For each diffusion model, we generate 271,710 synthetic images—matching the exact size and class distribution of the training set - using the DDIM scheduler with 100 inference steps. For each cancer type, we synthesize patches according to the same frequencies of the cancer types (and metadata, when applicable). Thus, both models (CLS, MeDi) produce the same total number of images but differ in how conditioning is applied. Figure 3 breaks down the results per class.

- **CLS Baseline:** Synthetic patches are generated conditioned solely on the cancer type (CLS).
- **MeDi** Synthetic patches are generated conditioned on both class and the medical center (TSS code in TCGA).

We observe that MeDi achieves a lower average FID than the CLS-only baseline, suggesting that modeling the TSS information helps to produce more faithful synthetic patches. In total, 23 out of 32 cancer types show improved FID scores relative to the baseline, with the most notable improvements occurring in underrepresented cancers that have fewer real samples. This can be also seen qualitatively in Figure 4. MeDi captures individual staining patterns better.

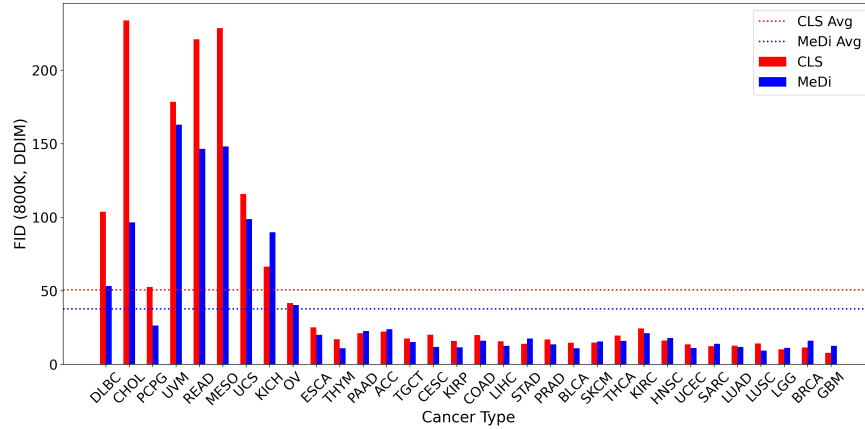


Fig. 3. FID per cancer type at 800K optimization steps sampled with DDIM. Depicted are the class-only baseline (CLS, red) and the metadata-conditioned model (MeDi, blue). Cancer types are ordered in descending order based on the number of images in the dataset. Dotted horizontal lines represent the average FID per model: CLS: 50.65, MeDi: 37.73.

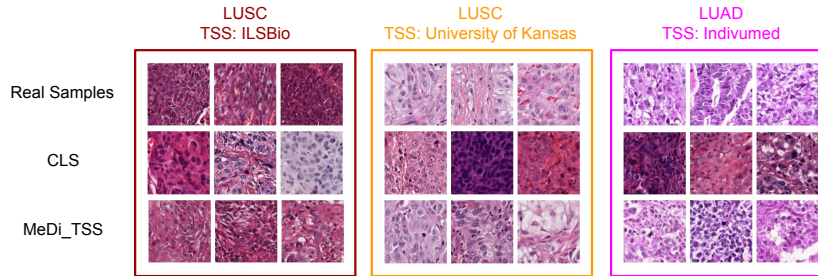


Fig. 4. LUSC and LUAD samples from different tissue source sites (TSS) along with generated images. The top row shows real data, while the two other rows show generated samples. MeDi enables capturing the staining mode of the different tissue source sites, while the CLS only model can not be conditioned on a specific TSS and, therefore, is unable to match the real data sample distribution.

4.4 Performance of Downstream Models on Subpopulation Shift

Next, we assess whether our generation model can improve downstream task performance by training tumor-subtype classifiers on three complementary subtyping problems in TCGA-UT (NSCLC, RCC and Uterine) spanning different degrees of imbalances and metadata-class correlations. In the training set, each cancer type is sourced from only one distinct TSS. This introduces a correlation

between TSS and class label, which mirrors many real-world scenarios where entire data cohorts may come from a single hospital. For both models (CLS and MeDi), we augment the training dataset by adding one synthetic example per real sample. For the CLS model, we generate samples only conditioning on the respective cancer type and sample uniformly over the classes. For MeDi, we generate synthetic samples by conditioning on the class and TSS: we uniformly sample over all possible combinations of selected cancer types and tissue source sites. This tailored sampling strategy aims to systematically break the class and metadata correlation. We train a linear classifier on top of the embeddings of 20 samples per class, extracted with the histopathology foundation model UNI [5]. The test set contains data from all TSS that the diffusion models have not seen during training, thereby creating a realistic subpopulation shift scenario. We repeat the model training for each possible combination of TSS from the training set as defined in Section 4.1.

Table 1. Linear probing results averaged over all possible tissue source site combinations in the training set. *Overall* indicates total balanced accuracy over the entire test set while for *TSS AVG*, the accuracy is computed for each TSS and averaged. The depicted variance shows standard error and best results are in bold.

Method	NSCLC		RCC		Uterine	
	Overall	TSS AVG	Overall	TSS AVG	Overall	TSS AVG
No syn. data	74.31 \pm 0.46	75.76 \pm 0.51	73.25 \pm 0.32	81.74 \pm 0.24	67.73 \pm 1.11	62.82 \pm 2.24
CLS only	79.99 \pm 0.31	81.72 \pm 0.33	78.21 \pm 0.27	84.64 \pm 0.14	67.77 \pm 0.89	70.80 \pm 1.37
MeDi	81.37 \pm 0.27	83.04 \pm 0.30	74.87 \pm 0.24	84.17 \pm 0.14	74.86 \pm 0.84	71.21 \pm 1.58

Table 1 shows the average balanced accuracy and the tissue source site (TSS)-averaged accuracy across all experimental runs. We find that synthetic data augmentation significantly improves performance to unseen tissue source sites in low data regimes. MeDi can further improve accuracy for NCLSC and Uterine classification while being roughly on par for RCC subtyping when considering the average TSS performance. This showcases that targeted data generation with MeDi can lead to better classifiers. In future work, we want to investigate if similar results can be achieved on larger datasets and in more general settings to confirm our findings.

5 Conclusion

Our experiments show that explicitly conditioning diffusion models on relevant metadata—such as hospital site yields two key benefits in computational histopathology: (1) improved fidelity of synthetic images to real-world distributions, and (2) that generative data augmentation can complement discriminative foundation models by selectively balancing subpopulation shifts. In future work, we will explore further metadata factors (e.g., scanner type, patient age, gender,

and race), larger-scale training regimes, extend the evaluation from linear probing to full fine-tuning, apply classifier-free guidance to improve the quality and downstream utility, and conduct a study to validate the relationship between FID and pathologist-rated image quality and realism.

Acknowledgments. This work was in part supported by the German Ministry for Education and Research (BMBF) under Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A, 031L0207D, and 01IS18037A. K.R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation). D.J.D. is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme, sponsored by the Federal Ministry of Education and Research.

The results shown here are in whole or part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

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

References

1. Alber, M., Tietz, S., Dippel, J., Milbich, T., Lesort, T., Korfiatis, P., Krügener, M., Cancer, B.P., Shah, N., Möllers, A., et al.: Atlas: A novel pathology foundation model by mayo clinic, charité, and aignostics. arXiv preprint arXiv:2501.05409 (2025)
2. Aversa, M., Nobis, G., Hägele, M., Standvoss, K., Chirica, M., Murray-Smith, R., Alaa, A.M., Ruff, L., Ivanova, D., Samek, W., et al.: Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology. *Advances in Neural Information Processing Systems* **36** (2024)
3. Campanella, G., Hanna, M.G., Geneslaw, L., Miralflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25**(8), 1301–1309 (2019)
4. Carrillo-Perez, F., Pizurica, M., Zheng, Y., Nandi, T.N., Madduri, R., Shen, J., Gevaert, O.: Rna-to-image multi-cancer synthesis using cascaded diffusion models. *bioRxiv* pp. 2023–01 (2023)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**(3), 850–862 (2024)
6. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al.: Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine* **25**(10), 1519–1525 (2019)
7. de Bel, T., Bokhorst, J.M., van der Laak, J., Litjens, G.: Residual cycleGAN for robust domain transformation of histopathological tissue slides. *Medical Image Analysis* **70**, 102004 (2021). <https://doi.org/https://doi.org/10.1016/j.media.2021.102004>, <https://www.sciencedirect.com/science/article/pii/S1361841521000505>

8. Dernbach, G., Kazdal, D., Ruff, L., Alber, M., Romanovsky, E., Schallenberg, S., Christopoulos, P., Weis, C.A., Muley, T., Schneider, M.A., et al.: Dissecting ai-based mutation prediction in lung adenocarcinoma: A comprehensive real-world study. *European journal of cancer* **211**, 114292 (2024)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021)
10. Dippel, J., Feulner, B., Winterhoff, T., Milbich, T., Tietz, S., Schallenberg, S., Dernbach, G., Kunt, A., Heinke, S., Eich, M.L., et al.: Rudolfv: a foundation model by pathologists for pathologists. *arXiv preprint arXiv:2401.04079* (2024)
11. Dippel, J., Prenißl, N., Hense, J., Liznerski, P., Winterhoff, T., Schallenberg, S., Kloft, M., Buchstab, O., Horst, D., Alber, M., et al.: Ai-based anomaly detection for clinical-grade histopathological diagnostics. *NEJM AI* **1**(11), AIoa2400468 (2024)
12. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
13. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications* **12**(1), 4423 (2021)
14. Jaume, G., Doucet, P., Song, A.H., Lu, M.Y., Almagro-Pérez, C., Wagner, S.J., Vaidya, A.J., Chen, R.J., Williamson, D.F., Kim, A., et al.: Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *arXiv preprint arXiv:2406.16192* (2024)
15. de Jong, E.D., Marcus, E., Teuwen, J.: Current pathology foundation models are unrobust to medical center differences (2025), <https://arxiv.org/abs/2501.18055>
16. Kauffmann, J., Dippel, J., Ruff, L., Samek, W., Müller, K.R., Montavon, G.: Explainable ai reveals clever hans effects in unsupervised learning models. *Nature Machine Intelligence* pp. 1–11 (2025)
17. Klauschen, F., Dippel, J., Keyl, P., Jurmeister, P., Bockmayr, M., Mock, A., Buchstab, O., Alber, M., Ruff, L., Montavon, G., Müller, K.R.: Toward explainable artificial intelligence for precision pathology. *Annual Review of Pathology: Mechanisms of Disease* **19**(Volume 19, 2024), 541–570 (2024), <https://www.annualreviews.org/content/journals/10.1146/annurev-pathmechdis-051222-113147>
18. Kömen, J., Marienwald, H., Dippel, J., Hense, J.: Do histopathological foundation models eliminate batch effects? a comparative study. *arXiv preprint arXiv:2411.05489* (2024)
19. Komura, D., Kawabe, A., Fukuta, K., Sano, K., Umezaki, T., Koda, H., Suzuki, R., Tominaga, K., Ochi, M., Konishi, H., Masakado, F., Saito, N., Sato, Y., Onoyama, T., Nishida, S., Furuya, G., Katoh, H., Yamashita, H., Kakimi, K., Seto, Y., Ushiku, T., Fukayama, M., Ishikawa, S.: Universal encoding of pan-cancer histology by deep texture representations. *Cell Reports* **38**(9), 110424 (2022). <https://doi.org/https://doi.org/10.1016/j.celrep.2022.110424>
20. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al.: Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine* pp. 1–8 (2024)
21. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. *Nature communications* **10**(1), 1096 (2019)

22. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. pp. 1107–1110. IEEE (2009)
23. Niehues, J.M., Müller-Franzes, G., Schirris, Y., Wagner, S.J., Jendrusch, M., Kloor, M., Pearson, A.T., Muti, H.S., Hewitt, K.J., Veldhuizen, G.P., et al.: Using histopathology latent diffusion models as privacy-preserving dataset augmenters improves downstream classification performance. *Computers in Biology and Medicine* **175**, 108410 (2024)
24. Osorio, P., Jimenez-Perez, G., Montalt-Tordera, J., Hooge, J., Duran-Ballester, G., Singh, S., Radbruch, M., Bach, U., Schroeder, S., Siudak, K., et al.: Latent diffusion models with image-derived annotations for enhanced ai-assisted cancer diagnosis in histopathology. *Diagnostics* **14**(13), 1442 (2024)
25. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. *IEEE Computer graphics and applications* **21**(5), 34–41 (2001)
26. Vaidya, A., Chen, R.J., Williamson, D.F., Song, A.H., Jaume, G., Yang, Y., Hartvigsen, T., Dyer, E.C., Lu, M.Y., Lipkova, J., et al.: Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine* **30**(4), 1174–1190 (2024)
27. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine* pp. 1–12 (2024)