Syn4MD: Aiding Medical Diagnosis through Image Synthesis and Classification

Kanishk Choudhary[1]

Received December 21, 2025
Accepted July 2, 2026
Electronic access August 15, 2026

Medical trainees depend on diverse visual reference material to develop diagnostic pattern-recognition skills, yet such material is often limited in diversity and accessibility. This paper presents a classifier-in-the-loop generative framework for synthesising realistic histopathology images: a Stable Diffusion v1.5 model is fine-tuned using Low-Rank Adaptation (LoRA) on the PathMNIST dataset, consisting of nine colorectal tissue classes, and a separately trained classifier screens each generated image, regenerating it (up to five attempts) if it fails a confidence threshold. LoRA hyperparameters were selected by Bayesian optimisation rather than by manual sweeps. To control for selection bias, a second, architecturally distinct classifier (ResNet-18) is held entirely outside the screening loop and used as the independent evaluator throughout. Two configurations are evaluated on three paired random seeds: Pipeline-1 (no rejection) and Pipeline-2 (DenseNet-121 screening + ResNet-18 independent evaluator). Pipeline-2 achieves a mean ResNet-18 accuracy of 0.939 ± 0.003 versus 0.913 ± 0.009 for Pipeline-1, a gain of +2.6 percentage points driven primarily by the hardest class, cancer-associated stroma ($0.360 \pm 0.061 \rightarrow 0.507 \pm 0.012$, +14.7 pp). The across-seed standard deviation on stroma also falls by an order of magnitude ($0.061 \rightarrow 0.012$), indicating the loop stabilises outcomes rather than merely raising means. A paired Wilcoxon signed-rank test on per-class recall pooled across class and seed (n=27) yields $p=3.5 \times 10^{-3}$. Limitations include the absence of a conditional-GAN (cGAN) comparator and expert review; both are future work.

Keywords: histopathology synthesis; diffusion models; LoRA fine-tuning; classifier-in-the-loop generation; rejection sampling; Bayesian hyperparameter optimisation.

Introduction

Medical imaging is a foundational tool for clinical diagnosis and education. Yet, access to a diverse range of high-quality medical images is often limited. Trainees may rely on only a few reference cases and rarely encounter certain conditions, hindering their learning. This creates a need for systems that can generate accurate, representative medical images on demand to augment training datasets and support diagnostic reasoning [1,2]. A greater focus was placed on generating realistic images for human training rather than solely developing classification models because human medical practitioners remain central to clinical decision-making due to reasons such as interpretability, patient trust, regulatory requirements, and liability concerns. Generative models have shown promise in producing synthetic medical images, but the precision required in medicine means that any inaccuracy can be misleading or harmful. The goal is to produce synthetic images with both the realism and the *semantic accuracy* - defined here as prompt-label adherence, i.e. a generated image whose tissue class as judged by an independently-trained classifier matches the tissue class specified in its conditioning prompt - necessary to support data augmentation and medical education, not for direct clinical decision-making.

Generative Adversarial Networks (GANs) have been widely used for medical image synthesis and augmentation [1]. Goodfellow et al.'s GAN framework [2] trains a generator against a discriminator to create realistic images, and numerous studies have applied GANs to expand medical training data [1,3]. For example, Frid-Adar et al. improved liver lesion classification by adding GAN-generated images to a training set, boosting a classifier's performance [4]. However, GANs often require large specialised datasets and careful tuning [1], and they commonly suffer from instability and mode collapse on limited data [1].

Diffusion models have emerged as a compelling alternative, achieving high-resolution, semantically rich outputs that rival or surpass GANs [5]. Latent diffusion models (e.g., Stable Diffusion) generate images via iterative denoising in a latent space [6], demonstrating state-of-the-art results on natural image synthesis [5]. Yet, a model like Stable Diffusion, trained on broad datasets (e.g., LAION-5B), performs poorly on specialised medical imagery without fine-tuning, as has

[1] *Irvington High School, California, USA*

been shown across diverse imaging modalities including chest X-ray[7], general medical synthesis[8], and histopathology[9]. If prompted with medical terms out-of-the-box, it may produce anatomically incorrect results. Figure 1 illustrates this concretely on a colorectal-tissue prompt: the baseline Stable Diffusion output (left) is anatomically implausible, while the fine-tuned LoRA generator paired with the rejection-sampling validator (middle) produces an image visually comparable to the real PathMNIST sample (right). This gap has spurred broader efforts to adapt diffusion models to medical domains[10,11].

This work introduces a classifier-in-the-loop generative framework — a form of rejection sampling for medically-conditioned diffusion - to address both realism and accuracy in medical image synthesis. Stable Diffusion v1.5 was fine-tuned on a histopathology dataset (PathMNIST[12]) using LoRA[13] to specialise it for colorectal tissue images. Then, a DenseNet-121 classifier was trained on the same dataset to serve as an automated screening classifier, with an architecturally distinct ResNet-18[14] held entirely outside the loop as the independent evaluator. During generation, each output image is evaluated by the screening classifier; if the predicted class does not match the input prompt or the top-class confidence falls below 0.80, the image is discarded and the model tries again. This iterative loop continues until a generated image is classified correctly, up to a maximum of five attempts, ensuring that only images consistent with the target class are retained. This pipeline (generator + screening classifier + independent evaluator) thus enforces semantic accuracy in real time. The result is a set of synthetic images that are not only visually realistic but also labelled correctly, supporting data augmentation and educational use without claiming readiness for direct clinical deployment.

For clarity, two terms are used throughout this paper. *Classifier-in-the-loop generation* refers to any procedure in which an independently trained classifier evaluates each generated sample and decides whether to accept or reject it, prompting regeneration on rejection. *Rejection sampling* is the standard Monte Carlo construction that this procedure instantiates: a tractable proposal distribution (here, the fine-tuned diffusion model conditioned on a tissue-type prompt) is repeatedly sampled, and samples are accepted only when they satisfy a target criterion (here, the classifier's predicted class matches the prompt). This differs from classifier-guided and classifier-free guidance[5,15,16], in which a classifier's gradient or implicit signal steers the interior denoising trajectory at each timestep; in this work the classifier acts only at the output of the diffusion sampler, as a binary gate over completed images. The most direct antecedents are discriminator-based rejection schemes for GANs — Discriminator Rejection Sampling[17] and Metropolis–Hastings GANs[18] — which use a trained discriminator to accept or reject completed generator samples; this work transfers that idea to text-conditioned diffusion, with an independently-trained tissue classifier as the acceptance gate. The contribution is therefore not the rejection-sampling primitive itself — which is well-established in the generative-modelling literature[19] — but its application to histopathology synthesis on PathMNIST with explicit instrumentation of the rejection loop, multi-seed evaluation, and an independent (cross-architecture) classifier as the final evaluator.

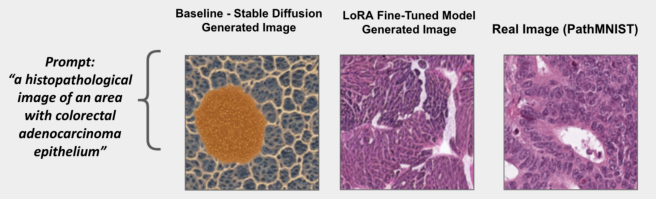

Figure. 1 Example of Generated vs. Real Images: Comparison between a real histopathology image (right), a synthetic image generated by the two-model system (middle), and an inaccurate image generated by a baseline Stable Diffusion model (left). These images all aim to represent the same tissue type.

Related Work

Combining image generation with automated validation in the medical domain is relatively new. Most prior studies focus on generating realistic images and evaluating them indirectly (e.g., by seeing if synthetic data improves a model's performance) rather than checking each image's accuracy. GAN-based methods have dominated early work in medical image synthesis[1]. Such work demonstrated that synthetic data can bolster limited datasets, but generation quality was typically evaluated only via downstream task performance rather than by verifying each generated image - the gap this work addresses.

More recently, Xue et al. introduced HistoGAN[20], a conditional GAN for histopathology that includes a selective augmentation step. HistoGAN generates images conditioned on class labels and then uses a classifier to filter out low-confidence outputs, adding only the most realistic samples to the training set; this is conceptually related to the broader literature on anomaly detection in medical imaging, where classifier confidence is used as a proxy for in-distribution membership[21]. This provides a form of post-hoc quality control and leads to improved histology image classification. However, HistoGAN's validation occurs after generation — to decide which images to keep for augmentation - rather than intervening during the image creation process.

Other approaches evaluate synthetic images via classifier metrics or expert review. For instance, some diffusion-based

pipelines for medical and histopathological imagery evaluate realism with pathologist surveys, FID-style distributional metrics, and downstream classification performance on models trained with the synthetic data[8,9]. These underline the importance of quality in medical image synthesis, yet they stop short of enforcing correctness during generation.

This work tightly integrates generation and validation. Unlike prior GAN or diffusion studies that optimise for visual realism or rely on indirect validation (e.g., improved downstream task performance), this system explicitly checks each output against the target label as it is produced. Indirect validation provides only aggregate feedback and cannot pinpoint specific failed images. In contrast, the classifier-driven rejection-sampling mechanism forces the generator to immediately regenerate or discard any output the classifier deems incorrect. Conceptually, this is similar to classifier-guided generation, but rather than using the classifier's gradients to steer the generation process (as in some diffusion models[5]), this system uses its predictions to iteratively accept or reject completed outputs - a form of rejection sampling over the joint distribution of diffusion outputs and classifier verdicts[19].

This work fine-tuned a state-of-the-art diffusion model on a specialised dataset and paired it with a domain-trained validator, creating an automated feedback loop for semantic accuracy in synthetic images. This is one of the first systems to actively enforce ground-truth consistency (each generated image's class matches its prompt) during image generation, rather than treating validation as an afterthought. It is further shown that this framework produces synthetic images that are not only visually plausible but also credibly representative of their intended class. Such images can serve as a reliable resource for data augmentation and medical training in scenarios where real data are scarce; we do not claim that such images are suitable as direct evidence in clinical diagnostic workflows.

Methods

Dataset

Experiments were conducted on the PathMNIST dataset (part of MedMNIST v2[12]), which consists of 107,180 haematoxylin and eosin stained pathology image patches categorised into nine tissue classes. These classes correspond to distinct colorectal tissue types (e.g., adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and adenocarcinoma epithelium) derived from the NCT-CRC-HE-100K collection[22]. Each image is a small region of a histology slide, resized to 224×224 pixels to standardise the input dimensions. We use the *predefined official PathMNIST splits* (train / validation / test) released with MedMNIST v2; the train split is used for fine-tuning the diffusion generator and for training both classifiers, the validation split is used for Bayesian hyperparameter optimisation, and the test split is used only for the real-image accuracy figures reported for ResNet-18 and DenseNet-121. Because the splits are inherited from MedMNIST, slide-level isolation between train, validation, and test is whatever is provided by the official release; the multi-seed evaluation reported in the Results section is conducted on synthetic images generated de novo at inference time and is therefore independent of the train/test slide assignment. The class distribution is approximately balanced - each tissue type is represented by on the order of $\sim$10,000 training examples - so no class reweighting is applied. For the classification model, data augmentation techniques (random rotations, zooms, contrast adjustments, etc.) were applied to the training images, following standard practice in computational pathology where augmentation and stain-colour normalisation are known to materially affect generalisation[23]. These augmentations improve the classifier's robustness to image variability without altering the underlying tissue label. Importantly, data augmentations were not applied when fine-tuning the generative model, since the diffusion model should learn each class's true visual characteristics without augmented distortions (ensuring that the text prompt "lymphocytes" always maps to authentically appearing lymphocyte patches, for example).

System Architecture

The system comprises a generative module and two classification modules that are combined into two evaluation pipelines (Fig. 2). The generative component is built on Stable Diffusion v1.5[6], a latent diffusion model capable of producing high-resolution images from text prompts. The model was fine-tuned on the PathMNIST training set using the LoRA technique[13], which introduces a small number of trainable low-rank adaptation parameters into the network's attention layers to efficiently specialise the model to the domain; LoRA is one of several text-to-image adaptation strategies that have been developed in parallel with subject-driven approaches such as DreamBooth[24].

Two classifiers are used in the system. ResNet-18[14] was trained on the PathMNIST training split and reaches accuracy on the official held-out test split consistent with the standard MedMNIST v2 benchmark ($\approx$0.91). A second classifier, DenseNet-121, was trained on the same PathMNIST split with matched augmentation and reaches comparable accuracy on real images; the two are deliberately chosen to be *architecturally distinct* so that agreement between them on synthetic images can be interpreted as evidence of robustness rather than shared inductive bias.

Two pipelines are then constructed from these components and compared head-to-head in the Results section. Pipeline-

1 (no rejection) samples from the LoRA-fine-tuned generator and classifies every generated image with both classifiers, with no regeneration. Pipeline-2 (classifier-in-the-loop) interposes DenseNet-121 as a screening classifier — a sample is accepted only if its top-class probability exceeds 0.80 *and* its predicted class matches the prompt; samples that fail screening are regenerated, up to a maximum of five attempts per image. Accepted images are then evaluated by ResNet-18, which is held entirely *outside* the screening loop to provide an evaluator that is independent of the rejection criterion and therefore controls for selection bias.

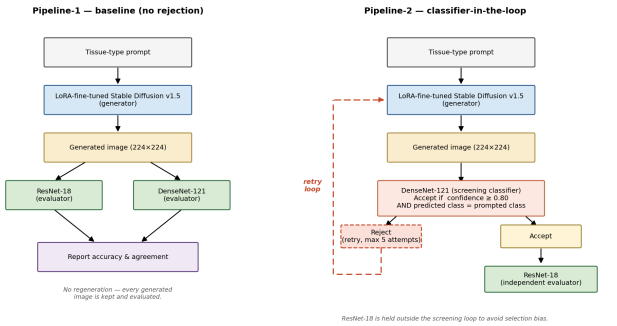

Figure. 2 The two evaluation pipelines compared in this work. Left: Pipeline-1 generates one image per prompt and evaluates it with both classifiers, with no regeneration. Right: Pipeline-2 uses DenseNet-121 as a screening classifier (accept if Prclass$\geq$0.80 and argmax=prompt, otherwise regenerate; up to five attempts), and uses ResNet-18 as an independent evaluator held outside the screening loop.

Experimentation Setup

The system was implemented in Python (using PyTorch) and all training and evaluation was conducted in a managed Jupyter environment (Google Colab). Two GPU types were leveraged during development: an NVIDIA A100 (40 GB) for computationally intensive tasks (e.g., diffusion model fine-tuning at higher resolution) and an NVIDIA T4 (16 GB) for lighter workloads (e.g., prototyping, classifier training, and image generation). The A100's larger memory enabled training with bigger batch sizes, while the T4 was sufficient for inference and for training the smaller ResNet-18 model.

To ensure reproducibility, version control and thorough experiment logging were employed throughout. All code, model configurations, and training logs were version-controlled, and the Bayesian hyperparameter optimisation study was managed with Optuna with per-trial training curves logged to Weights & Biases. Systematic naming/versioning (V1–V10 for the manual-pilot diffusion model variants, and Optuna trial numbers for the BO study) was also used to organise the experiments, and model checkpoints were saved at regular intervals (every few hundred steps) to enable rollback or analysis of intermediate results. Automated evaluation scripts were developed to compute key metrics (precision, recall, F1-score, and confusion matrices) on batches of generated images using the classifier, as well as a utility to generate a fixed set of prompts covering all tissue classes for consistent testing. These practices were chosen so that the workflow is reproducible and the reported results could be regenerated from the same code and settings.

Experimentation Details

Training Strategy and Model Versioning

Pilot Manual Exploration

Initial hyperparameter exploration was manual. The diffusion model was iteratively fine-tuned through ten versions (V1–V10), systematically adjusting training duration (1,000–3,000 steps), batch size (8, 16, 32), and learning rate (10^{-5}–5×10^{-5}, with gradual warm-up). Standard practices - a small noise offset during diffusion training and a cosine learning-rate decay schedule[25] - were also applied. Each intermediate version was scored on a small probe set (see "Model-selection evaluation" below). A steady trend in macro-F1 was observable over these versions, with V9 (1131 steps) reaching macro-F1 $\approx$0.67 on the probe set under the inference configuration that was later carried forward (Fig. 3). Although V9 was usable as a working configuration, manual sweeps cover a low-dimensional slice of the hyperparameter space and provide no principled control over the trade-off between trial budget and search precision; we therefore moved to Bayesian optimisation for the final hyperparameter selection.

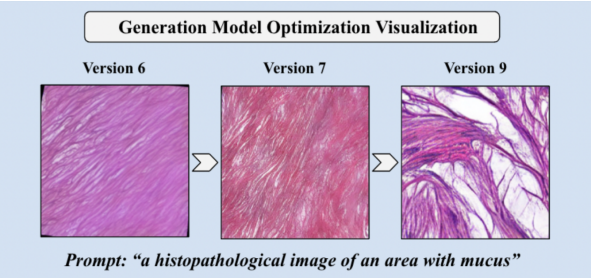

Figure. 3 Qualitative evolution of generation quality across the manual V1–V10 pilot, illustrated on a mucus prompt: outputs from V9 are much closer to the target tissue than those from V6 and V7.

Bayesian Hyperparameter Optimisation

We then ran a Bayesian optimisation study (Optuna, Tree-structured Parzen Estimator with Hyperband-style pruning) over a nine-dimensional search space: learning

rate $10^{-8}, 10^{-3}$ (log), noise offset 0,0.5, LoRA rank $\in\{2,4,8,16,32,64,128\}$, Adam weight decay $10^{-6}, 10^{-1}$ (log), LR scheduler $\in\{$cosine, linear, constant, cosine-with-restarts$\}$, SNR $-\gamma$ 0, 10, batch size $\in$4,8,16,32, Adam $\beta_1 \in$0.8,0.99, Adam $\beta_2 \in$0.9,0.9999. The objective was validation loss on a held-out PathMNIST validation set (minimisation). Fifty trials were budgeted; the pruner terminated 42 trials early on the strength of partial training trajectories, and 8 trials were trained to completion. The best trial (#10) reached a validation loss of 1.052, substantially below the next-best completed trial (1.058); the best configuration was lr = 9.0×10^{-4}, noise offset =0.36, LoRA rank =32, weight decay =1.2×10^{-6}, cosine LR scheduler, SNR $-\gamma$ = 6.2, batch size 16, Adam β_1=0.94, Adam β_2=0.96. Figure 4 shows the optimisation trajectory and Fig. 5 the parallel-coordinate view; the best-configuration trace is the dark line at low validation loss. Figure 6 reports a single-feature variance-explained estimate of hyperparameter importance; this is an indicative ablation rather than a full functional ANOVA (fANOVA) decomposition because eight completed trials are too few for reliable fANOVA, and we flag this limitation directly. Within those caveats, learning rate, Adam weight decay, and LoRA rank cluster as the top-three importance terms; the remaining parameters, led by noise offset, contribute progressively less. The best Bayesian configuration was used to retrain the production generator, which is the model used in all multi-seed evaluations reported in the Results section.

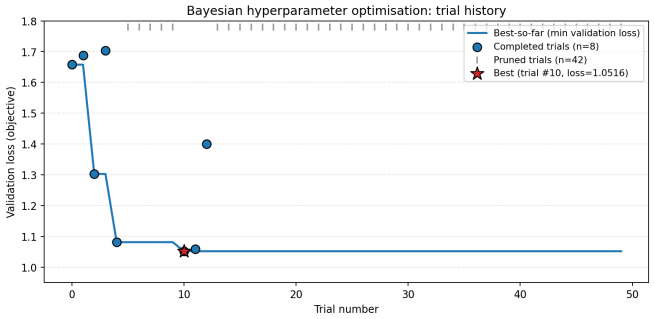

Figure. 4 Bayesian optimisation history. Validation loss (objective, minimised) of each completed trial, with the running best-so-far overlaid. Forty-two of fifty trials were terminated early by the pruner on partial trajectories (top tick marks). The best completed trial (#10, loss =1.052) was used to train the production generator.

Model-Selection Evaluation

At both the manual-pilot and Bayesian-optimisation stages, candidate configurations were scored on a small probe set: the diffusion model was prompted to generate 10 images per tissue class (90 images total), the ResNet-18 validator predicted a label for each image, and these predictions were compared to

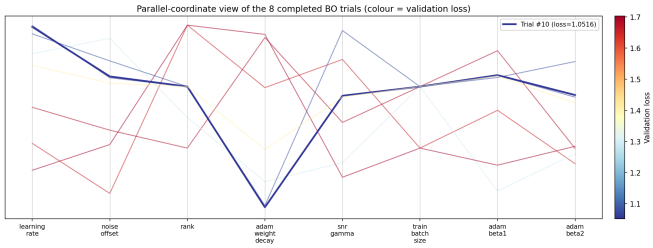

Figure. 5 Parallel-coordinate view of the eight completed BO trials. Each line is one trial across nine hyperparameter axes; line colour encodes validation loss (blue lower, red higher). The bold blue trace is the best trial (#10).

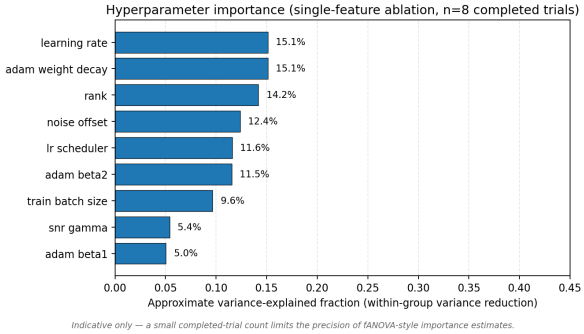

Figure. 6 Approximate hyperparameter importance from a single-feature within-group variance-reduction estimate over the eight completed trials. With this trial count fANOVA is not reliable, so this should be read as an indicative ranking rather than a precise decomposition. Learning rate, Adam weight decay, and LoRA rank are the most influential parameters under this estimate.

the prompt labels to compute precision, recall, and F1-score, macro-averaged across the nine classes. This lightweight protocol is sufficient for ranking configurations during hyperparameter search but is not the rigorous evaluation of the final pipeline. The rigorous evaluation — 100 images per class across three paired random seeds, with an independent cross-architecture evaluator — uses the production generator trained from the best Bayesian-optimisation configuration and is reported separately in the Results section. The production inference configuration carried forward is: DPMSolverMultistepScheduler, 40 inference steps, guidance scale 6.5, 224×224 resolution.

Results

Experimental Protocol

We evaluated two generation pipelines under matched conditions on three paired random seeds (1337, 42, 2024). Pipeline-

1 (baseline) samples from the LoRA-fine-tuned diffusion model with the production inference configuration (DPM-SolverMultistepScheduler, 40 inference steps, guidance scale 6.5, 224×224 resolution) and classifies each image with both ResNet-18 and DenseNet-121, with no rejection. Pipeline-2 (classifier-in-the-loop) interposes DenseNet-121 as a screening classifier with a confidence threshold of 0.80; samples that fail screening are regenerated, up to a maximum of five attempts per image. Accepted images are then evaluated by ResNet-18, which is held out of the screening loop to provide an evaluator that is independent of the rejection criterion. We frame Pipeline-2 as a form of classifier-in-the-loop rejection sampling for medically-conditioned diffusion[5,19]: the screening classifier is used to define an acceptance region in image space, and accepted samples are evaluated by an independent network to control for selection bias.

For each seed, both pipelines produce 100 images per class across the nine PathMNIST classes (900 images per seed per pipeline; 5,400 images in total across the multi-seed study). The same three seeds are applied to both pipelines so that the rejection-loop effect is isolated from sampler variance and a paired comparison by seed is statistically valid.

Overall Accuracy

Table 1 Overall accuracy and inter-model agreement, mean ± std over three paired seeds (1337, 42, 2024). ResNet-18 is the independent evaluator (held out of the screening loop); DenseNet-121 is the screening classifier in Pipeline-2 (starred numbers are selection-biased upper bounds). Acceptance rate and mean attempts are reported for Pipeline-2 only.

Pipeline	Metric	Mean	Std
Pipeline-1	ResNet-18 accuracy	0.913	0.009
Pipeline-1	DenseNet-121 accuracy	0.904	0.011
Pipeline-1	Inter-model agreement	0.908	0.009
Pipeline-2	ResNet-18 accuracy (independent)	0.939	0.003
Pipeline-2	DenseNet-121 accuracy (selected)*	0.995	0.003
Pipeline-2	Inter-model agreement	0.943	0.003
Pipeline-2	Acceptance rate	0.995	0.003
Pipeline-2	Mean attempts to acceptance	1.184	0.011

Table 1 reports overall accuracy of both classifiers on the synthetic images, together with inter-model agreement. The independent ResNet-18 evaluator measures Pipeline-1 accuracy at 0.913±0.009 and Pipeline-2 accuracy at 0.939±0.003, a gain of +2.6 percentage points (Fig. 7). DenseNet-121, which is the screening classifier inside Pipeline-2, measures Pipeline-1 accuracy at 0.904±0.011 and Pipeline-2 (selected) accuracy at 0.995±0.003; the latter is, by construction, a selection-biased upper bound and we therefore use the independent ResNet-18 number as the headline measure throughout. Inter-model agreement rises from 0.908±0.009 under

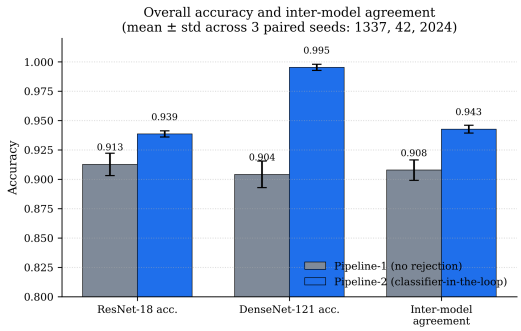

Figure. 7 Overall accuracy and inter-model agreement (mean ± std across three paired seeds). The independent ResNet-18 evaluator is the headline measure; the DenseNet-121 selected accuracy is flagged as a selection-biased upper bound.

Pipeline-1 to 0.943±0.003 under Pipeline-2 (Fig. 8) — classifier-in-the-loop sampling not only raises mean accuracy but also brings the two independently-trained classifiers into closer agreement on the synthetic samples that pass screening.

Pipeline-2 accepts 99.52%±0.26% of generated images within the five-attempt budget, at a mean of 1.184±0.011 attempts to acceptance; the additional generation cost is therefore modest in aggregate.

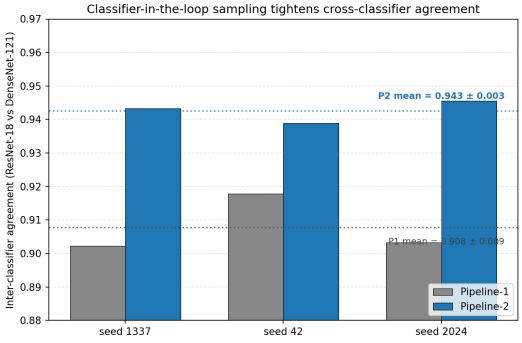

Figure. 8 Inter-model agreement (ResNet-18 vs. DenseNet-121) under both pipelines, shown per seed. Dashed lines mark the across-seed mean for each pipeline.

Per-Class Accuracy

Per-class accuracy reveals where the overall gain comes from (Fig. 9, Fig. 10, Table 2). Eight of nine classes are at or above 0.97 ResNet-18 accuracy under Pipeline-2 (seven of nine under Pipeline-1); the residual error is concentrated in cancer-associated stroma, which the diffusion model conflates with smooth muscle. Stroma accuracy under the independent evaluator rises from 0.360±0.061 (Pipeline-1) to 0.507±0.012 (Pipeline-2), a gain of +14.7 percentage points

Table 2 Per-class accuracy under the independent ResNet-18 evaluator, mean ± std over three paired seeds. Δ = Pipeline-2 - Pipeline-1 in percentage points. Rows sorted by Pipeline-2 accuracy ascending.

Class	P1 (mean±std)	P2 (mean±std)	Δ (pp)
Cancer-assoc. stroma	0.360 ± 0.061	0.507 ± 0.012	$+14.7$
Smooth muscle	0.970 ± 0.017	0.973 ± 0.015	$+0.3$
Debris	0.927 ± 0.031	0.977 ± 0.012	$+5.0$
Mucus	0.983 ± 0.006	0.993 ± 0.006	$+1.0$
Lymphocytes	1.000 ± 0.000	0.997 ± 0.006	-0.3
Adipose	0.990 ± 0.000	1.000 ± 0.000	$+1.0$
Background	1.000 ± 0.000	1.000 ± 0.000	$+0.0$
Normal mucosa	0.993 ± 0.012	1.000 ± 0.000	$+0.7$
Adenocarcinoma	0.990 ± 0.010	1.000 ± 0.000	$+1.0$

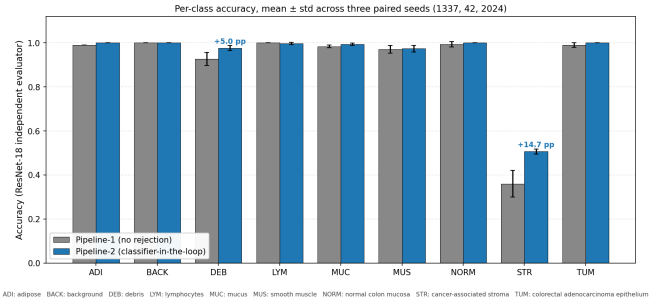

Figure. 9 Per-class accuracy under the independent ResNet-18 evaluator. Δ values (shown above each Pipeline-2 bar) are P2 - P1 in percentage points. Error bars show ± std across three seeds.

and the dominant contribution to the overall improvement. Notably, the across-seed standard deviation for stroma drops by an order of magnitude in Pipeline-2 (0.061→0.012), indicating the classifier-in-the-loop sampling not only raises the stroma mean but also stabilises the outcome across seeds.

The remaining classes show small, mostly positive deltas: debris (+5.0 pp), adipose (+1.0), mucus (+1.0), adenocarcinoma (+1.0), normal mucosa (+0.7), smooth muscle (+0.3), background (0.0), lymphocytes (-0.3). The single small negative is on a class that was already at 1.000 and reflects a single seed in which one of 100 lymphocyte samples was misclassified after screening.

Statistical Analysis

We use two paired tests, with the seed assignment as the pairing variable. As a primary test on per-class recall we pool across class and seed (n=27 paired observations) and apply the Wilcoxon signed-rank test (one-sided, P2 > P1), using Pratt's treatment of zero differences and the exact null distribution rather than the large-sample normal approximation, which is appropriate at this sample size. With the independent ResNet-18 evaluator this gives $p = 3.5 \times 10^{-3}$; the corresponding paired t-test gives t=2.65, two-sided p=0.014. Using DenseNet-121 (which is the screening classifier and thus selection-biased) the Wilcoxon test gives $p = 3.6 \times 10^{-5}$. As a secondary test on overall accuracy paired by seed alone (n=3) the paired t-test gives t=3.81, p=0.062; we note explicitly that the three-seed test is underpowered (Fig. 11 visualises the across-seed variability behind these tests) and report the pooled class × seed test as primary for this reason. We note that the nine per-class observations within a seed are not mutually independent — they derive from the same generator and screening run — so the pooled test trades strict independence for statistical power; pairing by seed controls for between-run variation, and the agreement of the pooled Wilcoxon test, the by-seed t-test, and the per-class sign test below guards against over-reliance on any single test's assumptions. A cluster-robust or hierarchical model treating class and seed as nested factors is left to future work.

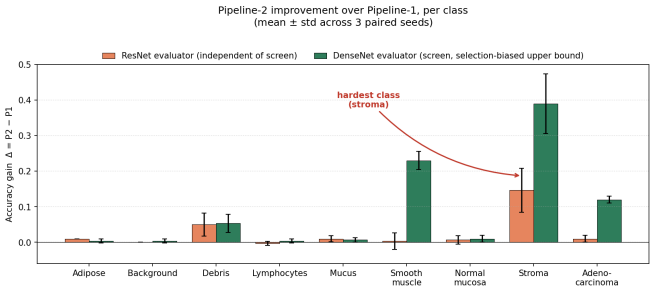

Figure. 10 Per-class accuracy gain Δ=P2-P1 under each evaluator. The independent ResNet-18 evaluator (orange) places the bulk of the improvement on cancer-associated stroma. The DenseNet-121 evaluator (green) reports larger gains on stroma, smooth muscle, and adenocarcinoma but is selection-biased because the same classifier defines the acceptance criterion in Pipeline-2; the ResNet-18 numbers are the headline measure throughout the paper.

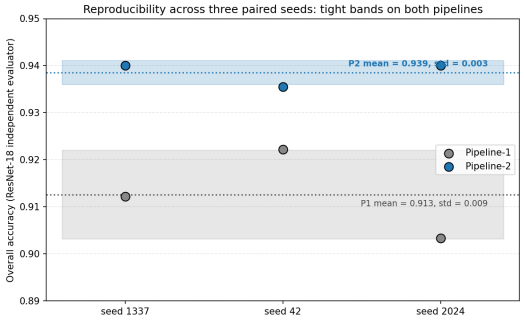

Figure. 11 Seed-stability plot. The plot shows the per-seed overall accuracy under the independent ResNet-18 evaluator, with the three seeds as individual points.

A sign test on the per-class mean improvement Δ=P2-P1 across the nine classes gives 7/9 classes improved, 1 tied, 1 marginally declined (one-sided binomial p=0.035); under the DenseNet evaluator all 9 classes improve ($p = 1.95 \times 10^{-3}$). Stroma alone, paired by seed (n=3), improves under both evaluators: ResNet Δ=+0.147, p=0.062; DenseNet Δ=+0.390, p=0.008.

Failure Analysis

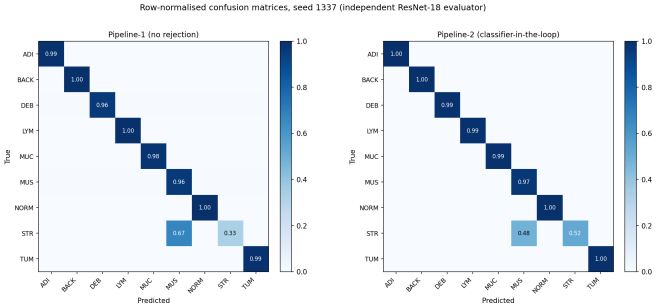

Figure. 12 Row-normalised confusion matrices on the synthetic images, evaluated by the independent ResNet-18, seed 1337. Left: Pipeline-1 (no rejection). Right: Pipeline-2 (classifier-in-the-loop).

Figure 12 shows row-normalised confusion matrices for seed 1337 (the other two seeds are qualitatively similar). The Pipeline-1 stroma row sends 67% of stroma samples to smooth muscle; under Pipeline-2 this drops to 48%, with the remainder correctly classified as stroma. No other class transition reaches 5% in either pipeline. Figure 13 summarises the stroma-specific picture across both classifiers, including the magnitude of selection bias on the DenseNet-121 screening view. The dominant residual failure mode is therefore stroma→smooth muscle confusion, consistent with the histological observation that these two tissues share overlapping morphological cues (elongated nuclei, fibrillar matrix) at the tile-level scale of PathMNIST. Prior work on the NCT-CRC-HE collection has shown that colorectal-tissue tiles at this resolution remain discriminative for downstream tasks[26], suggesting that the limit here is the generator's prompt-conditional mode coverage rather than the inherent separability of the two tissue types in the data.

Figure 14 reports per-class generation cost under Pipeline-2. Stroma requires 1.85±0.09 attempts on average to pass screening, compared with ≤1.5 for every other class, and is the only class with non-zero five-attempt failure rate (3.67%±2.1%). The diffusion generator is therefore the bottleneck for stroma, not the classifier - the rejection loop is correctly identifying poor stroma samples and the generator is unable to reliably produce alternatives within five attempts.

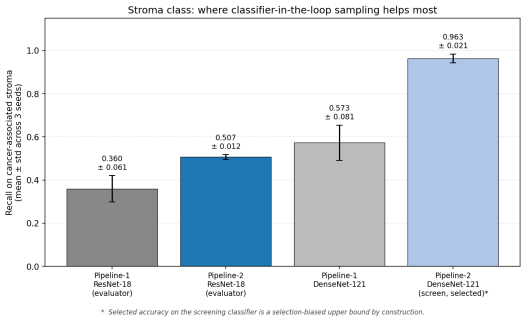

Figure. 13 Cancer-associated stroma deep-dive. Stroma recall under both pipelines, measured by the independent ResNet-18 evaluator and by the DenseNet-121 screening classifier (mean ± std across three paired seeds). The Pipeline-2 DenseNet-121 value is a selection-biased upper bound by construction.

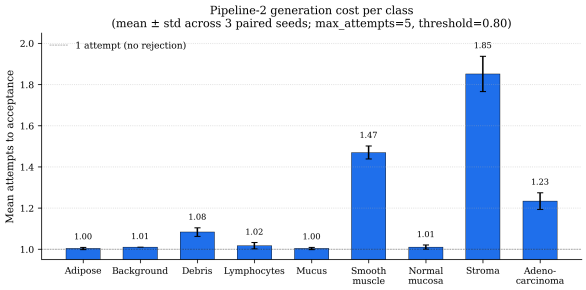

Figure. 14 Mean number of generation attempts to acceptance under Pipeline-2, per class (mean ± std across three paired seeds). The dashed reference line at 1.0 corresponds to Pipeline-1 (no rejection).

Discussion

We have presented Pipeline-2 as a form of classifier-in-the-loop rejection sampling for medically-conditioned diffusion: the screening classifier defines an acceptance region in image space and accepted samples are evaluated by an independently-trained network to control for selection bias. Framed this way, the contribution of the work is not a new generative paradigm but a careful integration and empirical evaluation of three existing components — LoRA-fine-tuned latent diffusion[6,13], classifier-guided/rejection sampling[5,15,19], and a held-out independent evaluator — applied end-to-end to a colorectal histopathology benchmark[22].

Two findings are worth highlighting. First, the per-class improvement is highly non-uniform: seven of nine PathMNIST classes are already near-saturated under the baseline diffusion sampler, and the majority of the gain — roughly two-thirds of the +2.6 pp — is concentrated on the single hardest class, cancer-associated stroma (+14.7 percentage points under the independent evaluator). This is the expected behaviour of rejection sampling in regions where the generator

places most of its mass on a competing mode — here, smooth muscle — and provides empirical support for the rejection-sampling interpretation. Second, the across-seed standard deviation on the stroma class drops by an order of magnitude under Pipeline-2 (0.061→0.012). The rejection loop is therefore not only raising the mean but also stabilising the outcome, which is the more interesting property for a downstream user who cares about consistency of synthetic-data quality across random seeds.

A separate, methodological observation concerns hyperparameter selection. Manual sweeps over V1–V10 covered only a low-dimensional slice of the hyperparameter space; replacing them with Bayesian optimisation (50-trial Optuna study with Hyperband-style pruning) produced a generator with lower validation loss than any manually-tuned variant, and the importance ranking it surfaced — learning rate, Adam weight decay, and LoRA rank as the top-three parameters — matches the intuition that learning rate and regularisation dominate parameter-efficient fine-tuning, though we did not attempt to confirm this against a formal external benchmark. We flag two caveats explicitly: only 8 of 50 trials completed (the remaining 42 were pruned early), and the per-parameter importance estimate is therefore single-feature rather than a full fANOVA decomposition. Within those caveats, BO replaces an ad-hoc grid with a budget-aware, reproducible search and is recommended for future LoRA-adaptation work on this benchmark.

We also considered but did not run a Pipeline-3 ablation in which the roles of DenseNet-121 and ResNet-18 are reversed (ResNet-18 as the screening classifier and DenseNet-121 as the independent evaluator). Such an ablation would isolate how much of the Pipeline-2 gain is attributable to the choice of screening classifier specifically, rather than to the existence of a screen at all. We omitted it from the present study to keep the experimental matrix tractable and because the DenseNet-as-screen / ResNet-as-evaluator direction is the one consistent with our per-class screening analysis (DenseNet-121 produced more permissive accept decisions on the hardest class, stroma, which is the property a screening classifier in this pipeline must have to be useful). The role-reversal direction is listed as a future-work ablation.

Limitations

Selection bias on the screening classifier. The DenseNet-121 selected accuracy of 0.995±0.003 is a selection-biased upper bound, because the same classifier defines acceptance and is used for evaluation. We therefore report the independent ResNet-18 number (0.939±0.003) as the headline measure, but the inflation of the DenseNet number across seeds illustrates the size of the selection effect that any work using its screening classifier as evaluator is silently incurring.

No conditional-GAN baseline. The reported comparison is between two diffusion-based configurations of the same generator. Pipeline-1 (no rejection) functions as an ablation that isolates the contribution of the classifier-in-the-loop component within the diffusion pipeline; it does not isolate the contribution of diffusion versus alternative generative paradigms[27]. A controlled comparison against a conditional GAN of comparable parameter count and training budget remains future work and would speak directly to the broader question of whether diffusion is the right backbone for histopathology tile generation.

No expert pathologist review. Both pipelines are evaluated by deep classifiers, not by human experts. Classifier accuracy is a useful surrogate for prompt-label adherence and image realism[28], but it does not substitute for histological plausibility judgements that only a trained pathologist can make. We have identified the high-disagreement and low-confidence cases (the reference-seed files disagreement_cases.csv and lowest_confidence_cases.csv, available on request) as a natural priority set for an expert review study; this remains future work.

Limited number of seeds. Multi-seed reporting in this work uses three paired seeds (1337, 42, 2024). This is sufficient for the primary statistical test (paired Wilcoxon on n=27 class × seed observations, $p = 3.5 \times 10^{-3}$) but is underpowered for tests that pair only by seed (n=3); we report these as supporting evidence with explicit caveats. A higher-seed study is straightforward to run and is included in the future work plan.

Single dataset, single tissue type. The evaluation is on PathMNIST (colorectal tissue, derived from the NCT-CRC-HE-100K collection[22]). We do not test out-of-domain generalisation, and the broader question of whether classifier-in-the-loop synthetic data improves downstream clinical-task performance — the standard utility test for medical synthetic data[29] — is left for future work. The classifier-in-the-loop framing is dataset-agnostic, but the size and direction of the per-class improvement are not, and the failure mode (stroma→smooth muscle) is a colorectal-tissue artefact.

Tile-level evaluation only. All evaluation is at the 224×224 tile scale defined by PathMNIST. Whole-slide-level evaluation, in which generated tiles are composited and assessed for spatial coherence, is outside the scope of this study.

Future Work

The five lines of follow-up work suggested by this study are:

1. Conditional-GAN baseline of comparable parameter and training budget, evaluated under the same multi-seed protocol, to disentangle the diffusion-backbone contribution from the classifier-in-the-loop contribution.

2. Expert pathologist review of the high-disagreement and

low-confidence cases identified by this study, with a particular focus on the stroma→smooth muscle confusion pair.

3. Larger seed counts and bootstrap-confidence-interval reporting to tighten the inferential claims on overall metrics, plus a Pipeline-3 role-reversal ablation (ResNet-18 as screening classifier, DenseNet-121 as independent evaluator) to isolate the contribution of the specific classifier-screen choice from the existence of any screen at all.

4. Targeted stroma sub-conditioning. The current generator places too much mass on smooth muscle when prompted for stroma; sub-class textual conditioning or stroma-specific LoRA adapters are natural next experiments.

5. Scale-up and alternative guidance. Two complementary directions extend the approach beyond the present scope: (i) replacing the rejection-sampling gate with classifier-guided or energy-guided sampling that incorporates the classifier signal directly into the denoising trajectory rather than as a post-hoc accept/reject decision[5,30], which could reduce the average number of regeneration attempts; and (ii) scaling from patch-level synthesis on PathMNIST to whole-slide synthesis using larger-scale histopathology corpora such as TCGA[31], which would expose the framework to inter-patient and inter-site variability absent from the present evaluation.

Summary and Outlook

This work set out to determine whether classifier-in-the-loop rejection sampling can produce histopathology images that are both visually realistic and semantically aligned with their target class, and to do so under an evaluation that controls for the most common forms of methodological circularity in this setting. Three findings answer that question. First, the rejection-sampling gate raises overall accuracy on the independent ResNet-18 evaluator from 0.913 ± 0.009 to 0.939 ± 0.003, with the gain concentrated on the single hardest class (cancer-associated stroma, +14.7 percentage points). Second, the improvement is statistically supported under a paired test pooled across class and seed (Wilcoxon, n=27, $p = 3.5 \times 10^{-3}$) and stable across three paired random seeds. Third, the across-seed standard deviation on stroma falls by an order of magnitude under classifier-in-the-loop sampling, indicating that the loop stabilises outcomes rather than merely raising means.

These results connect directly to the objectives stated in the introduction: the framework produces synthetic images that the independent evaluator accepts at the target class with high reliability, and the per-class analysis localises the remaining failure mode (stroma→smooth muscle) to a specific morphological ambiguity that can be addressed in future work. The framing as rejection sampling for medically-conditioned diffusion, with an architecturally distinct held-out evaluator, generalises beyond this benchmark; the specific numerical gains do not, and we explicitly do not claim clinical readiness. The contribution is a methodological recipe — LoRA-fine-tuned latent diffusion, Bayesian hyperparameter selection, a screening classifier with a rejection gate, and an independent cross-architecture evaluator — that can be transferred to other histopathology and medical-imaging tasks with appropriate domain-specific tuning. We hope the explicit instrumentation of the rejection loop, the disagreement and low-confidence case lists, and the multi-seed protocol can serve as a template for rigorous evaluation of synthetic medical images more broadly.

Acknowledgments

The per-seed, per-class results underlying every table and figure in this paper are available from the authors on request and will be deposited in a public repository upon publication. These comprise the full three-seed long-format results file (raw_long.csv), which lists, for each of the three random seeds (1337, 42, and 2024), the nine per-class recall values under each pipeline together with the overall-accuracy, inter-model-agreement, acceptance-rate, and mean-attempts summaries; and, for the reference seed (1337), the two classifier-level case lists discussed above — the classifier-disagreement cases (disagreement_cases.csv) and the lowest-confidence accepted cases (lowest_confidence_cases.csv) — each reporting the prompt, the true label, both classifiers' predicted labels and full nine-class probability vectors, and the inter-model agreement flag. The diffusion fine-tuning, image-generation, screening, and evaluation code will be made available upon publication.

References

1 S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, A. Mukhopadhyay. GANs for medical image analysis: a review. Artificial Intelligence in Medicine. Vol. 109, pg. 101938, 2020, https://doi.org/10.1016/j.artmed.2020.101938.

2 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. Advances in Neural Information Processing Systems. Vol. 27, pg. 2672–2680, 2014.

3 X. Yi, E. Walia, P. Babyn. Generative adversarial network in medical imaging: a review. Medical Image Analysis. Vol. 58, pg. 101552, 2019, https://doi.org/10.1016/j.media.2019.101552.

4 M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, H. Greenspan. GAN-based synthetic medical image augmentation for improved liver lesion classification. Neurocomputing. Vol. 321, pg. 321–331, 2018, https://doi.org/10.1016/j.neucom.2018.09.013.

5 P. Dhariwal, A. Nichol. Diffusion models beat GANs on image synthe-

sis. Advances in Neural Information Processing Systems. Vol. 34, pg. 8780–8794, 2021.

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

7 C. Bluethgen, P. Chambon, J. B. Delbrouck, R. van der Sluijs, M. Połacin, S. Zickler, Z. Zhou, A. Chaudhari. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering. Vol. 8, pg. 1–13, 2024, https://doi.org/10.1038/s41551-023-01039-4.

8 G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, J. N. Kather, D. Truhn. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image generation. Scientific Reports. Vol. 13, pg. 12098, 2023, https://doi.org/10.1038/s41598-023-39278-0.

9 P. A. Moghadam, S. Van Dalen, K. C. Martin, J. Lennerz, S. Yip, H. Farahani, A. Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pg. 2000–2009, 2023, https://doi.org/10.1109/WACV56688.2023.00204.

10 F. Khader, G. Müller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, D. Truhn. Denoising diffusion probabilistic models for 3D medical image generation. Scientific Reports. Vol. 13, pg. 7303, 2023, https://doi.org/10.1038/s41598-023-34341-2.

11 W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, M. J. Cardoso. Brain imaging generation with latent diffusion models. Deep Generative Models (MICCAI 2022 Workshop), Lecture Notes in Computer Science. Vol. 13609, pg. 117–126, 2022, https://doi.org/10.1007/978-3-031-18576-2_12.

12 J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, Z. Zhou, B. Ni. MedMNIST v2: a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data. Vol. 10, pg. 14, 2023, https://doi.org/10.1038/s41597-022-01721-8.

13 E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, B. Wang, W. Chen. LoRA: low-rank adaptation of large language models. arXiv preprint. arXiv:2106.09685, 2021.

14 K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pg. 770–778, 2016.

15 J. Ho, T. Salimans. Classifier-free diffusion guidance. arXiv preprint. arXiv:2207.12598, 2022.

16 A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. Proceedings of the 39th International Conference on Machine Learning (ICML). PMLR Vol. 162, pg. 16784–16804, 2022.

17 S. Azadi, C. Olsson, T. Darrell, I. Goodfellow, A. Odena. Discriminator rejection sampling. International Conference on Learning Representations (ICLR). 2019, arXiv:1810.06758.

18 R. Turner, J. Hung, E. Frank, Y. Saatchi, J. Yosinski. Metropolis-Hastings generative adversarial networks. Proceedings of the 36th International Conference on Machine Learning (ICML). Vol. 97, pg. 6345–6353, 2019.

19 S. Bond-Taylor, A. Leach, Y. Long, C. G. Willcocks. Deep generative modelling: a comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 44, pg. 7327–7347, 2022, https://doi.org/10.1109/TPAMI.2021.3116668.

20 Y. Xue, J. Zou, X. Tian, X. Liu, P. Kumar, X. Huang. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. Medical Image Analysis. Vol. 67, pg. 101816, 2021, https://doi.org/10.1016/j.media.2020.101816.

21 M. E. Tschuchnig, M. Gadermayr. Anomaly detection in medical imaging — a mini review. Data Science — Analytics and Applications. pg. 33–38, 2022, https://doi.org/10.1007/978-3-658-36295-9_5.

22 J. N. Kather, J. Krisam, D. Charon, M. O. Zörner, T. Marotz, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Medicine. Vol. 16, pg. e1002730, 2019, https://doi.org/10.1371/journal.pmed.1002730.

23 D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, J. van der Laak. Quantifying the effects of data augmentation and stain colour normalisation in convolutional neural networks for computational pathology. Medical Image Analysis. Vol. 58, pg. 101544, 2019, https://doi.org/10.1016/j.media.2019.101544.

24 N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pg. 22500–22510, 2023.

25 I. Loshchilov, F. Hutter. SGDR: stochastic gradient descent with warm restarts. International Conference on Learning Representations. 2017.

26 J. N. Kather, A. T. Pearson, N. Halama, D. Jäger, J. Krause, S. H. Loosen, A. Marx, P. Boor, F. Tacke, U. P. Neumann, H. I. Grabsch, T. Yoshikawa, H. Brenner, J. Chang-Claude, M. Hoffmeister, C. Trautwein, T. Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine. Vol. 25, pg. 1054–1056, 2019, https://doi.org/10.1038/s41591-019-0462-y.

27 T. Karras, S. Laine, T. Aila. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pg. 4401–4410, 2019, https://doi.org/10.1109/CVPR.2019.00453.

28 A. Borji. Pros and cons of GAN evaluation measures: new developments. Computer Vision and Image Understanding. Vol. 215, pg. 103329, 2022, https://doi.org/10.1016/j.cviu.2021.103329.

29 R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, F. Mahmood. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering. Vol. 5, pg. 493–497, 2021, https://doi.org/10.1038/s41551-021-00751-8.

30 C. Lu, H. Chen, J. Chen, H. Su, C. Li, J. Zhu. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. Proceedings of the 40th International Conference on Machine Learning (ICML). PMLR Vol. 202, pg. 22825–22855, 2023.

31 K. Tomczak, P. Czerwińska, M. Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology. Vol. 19, pg. A68–A77, 2015, https://doi.org/10.5114/wo.2014.47136.