



Building a Clinically Relevant and Technically Robust Synthetic Histopathology Dataset for Breast and Gastric Cancer

So Hyeon Lee¹ · Young Seop Lee² · Young Jae Kim³ · Jisup Kim⁴ · Kwang Gi Kim^{1,2}

Received: 2 February 2026 / Accepted: 5 June 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Data scarcity, inter-institutional stain variability, and privacy constraints are major challenges impeding the development of generalizable artificial intelligence models in digital pathology. Although recent generative adversarial network (GAN)-based synthesis approaches show promise, they struggle to preserve fine-grained nuclear morphology or maintain class-specific histological diversity. Moreover, existing diffusion-based studies have not sufficiently addressed class-conditional synthesis across heterogeneous, multi-institutional pathology data. We introduce the Efficient Pathology Diffusion Pipeline (EPDP), a class-conditional diffusion framework that enables the generation of clinically relevant, subtype-specific synthetic histopathology images for training and validating diagnostic AI models. By providing high-fidelity, subtype-specific synthetic datasets, EPDP lowers the barriers to developing robust AI diagnostic tools and supports the establishment of standardized evaluation frameworks in digital pathology. To achieve this, EPDP integrates a customized denoising U-Net that employs nuclear details, learnable class embeddings, and CycleGAN-based stain normalization with reference-guided alignment. This design explicitly targets the preservation of fine-grained nuclear morphology, class-specific histological diversity, and stain-invariant visual consistency across institutions. Multi-institutional hematoxylin and eosin whole-slide images are curated via a two-round pathology review to build subtype-labeled training and evaluation sets. Image fidelity is assessed using FID, 1-LPIPS, and SSIM, and compared against state-of-the-art models including PathDiff, PathLDM, and DiffInfinite. Diagnostic utility was evaluated using cross-domain classification (EfficientNetV2-L), and perceptual realism was assessed using a visual Turing test (VTT). The real–synthetic FID was lower than that of real–real FID by 11.0% (breast) and 8.6% (gastric). The subtype 1-LPIPS gaps were $\leq 12\%$, and the SSIM gaps were $\leq 0.3\%$. Classifiers trained only on synthetic patches matched real-trained baselines within an $\sim 1\%$ F1 score on real-image validation. Pathologists performed at near-chance levels in the VTT, with accuracies ranging from 50% to 56%.

Keywords Diffusion model · Digital pathology · Class-conditional generation · Histopathological image synthesis · Gastric cancer · Breast cancer

✉ Kwang Gi Kim
kimkg@gachon.ac.kr

✉ Jisup Kim
jspath@gilhospital.com

¹ Department of Biomedical Engineering, Gachon University, Seongnam-si, Gyeonggi-do, Republic of Korea

² Department of Biomedical Engineering, Gil Medical Center, Gachon University College of Medicine, Incheon, Republic of Korea

³ Gachon Biomedical & Convergence Institute, Gachon University Gil Medical Center, Incheon, Republic of Korea

⁴ Department of Pathology, Gil Medical Center, Gachon University College of Medicine, Incheon, Republic of Korea

Introduction

Gastric and breast cancers are representative malignancies with high global prevalence and mortality rates, making them primary targets for evaluating diagnostic AI models. In 2020, ~ 1.1 million new gastric cancer cases were reported globally. The highest incidence rates were observed in Eastern Asia and Central/Eastern Europe, with higher rates noted among Asian and Hispanic populations in the United States. Gastric cancer is a highly lethal disease with a five-year survival rate of $\sim 20\%$, and about 0.8 million cancer-related deaths occurred worldwide [1].

Breast cancer accounted for ~ 2.3 million cases in 2022, which makes it the most common cancer among women,

with particularly high incidence rates in Europe and North America, and ~0.7 million deaths worldwide [2]. Despite this heavy burden, artificial intelligence (AI) research utilizing pathological images of gastric and breast cancers faces significant limitations. For example, histological diversity and heterogeneity make constructing large-scale pathology datasets highly time- and cost-intensive, while ethical and legal restrictions further hinder data sharing [3, 4]. Publicly available pathology datasets remain limited in scale and suffer from differences in staining and scanning across institutions, restricting the generalizability of AI models [4, 5]. Collectively, these factors constitute major barriers to the clinical translation and scalability of AI models based on the pathological images of gastric and breast cancers.

Synthetic data have recently emerged as a promising alternative to complement real clinical data because deep-learning-based generative models demonstrate the potential to preserve the structural characteristics and cytological patterns of pathology images while producing diverse variations [6]. Although earlier studies reported GAN-based approaches to synthesize histopathological images, including nuclei segmentation tasks in gastric and breast cancers [7, 8], there are several challenges such as ambiguous mappings that lead to difficulties in representing fine structures and structural distortions that limit their clinical applicability [9]. In contrast, diffusion-based generative methods have gained considerable attention as a promising strategy to overcome these limitations because of their stable training and ability to generate high-quality images; they are increasingly being applied to medical image synthesis, including histopathology [10]. In practice, a morphology-focused diffusion probabilistic model with color normalization outperforms ProGAN, which achieves better results in terms of the Fréchet inception distance (FID) and inception score compared with that of the diffusion model [11]. These results highlight the superiority of diffusion-based approaches over GANs in histopathological image synthesis, particularly in terms of preserving structural details and visual fidelity [11]. However, diffusion-based studies remain in nascent stages, and the class-conditional synthesis of gastric and breast cancer pathological images have not been reported.

Although CycleGAN-based class-conditional methods have been reported in other cancers [12, 13], diffusion-based approaches have not been explored in this context. Pozzi et al. proposed a diffusion-based synthesis pipeline for nonneoplastic tissue samples (brain, kidney, lung, pancreas, and uterus); however, their approach did not extend to class-specific tumor conditions [14]. Chen et al. proposed a conditional diffusion model for breast cancer MRI images to predict tumor subtypes; however, its application remained confined to radiological imaging rather than the histopathological slide images [15].

Therefore, in this paper, we propose the Efficient Pathology Diffusion Pipeline (EPDP), a class-conditional diffusion-based framework for synthesizing gastric and breast cancer histopathology images. These specific cancer types were selected not only for their clinical significance but also for their high histological diversity and morphological heterogeneity across diagnostic subtypes, which provide an ideal environment for validating EPDP's ability to preserve fine-grained nuclear and cellular characteristics. Unlike prior studies focused on single-center data, EPDP leverages multi-institutional datasets and incorporates a reference-guided CycleGAN-based stain normalization to effectively mitigate the domain shift caused by heterogeneous laboratory protocols. We evaluate the quality of the synthetic images using qualitative and quantitative metrics and further examine their effect on the downstream classification performance to demonstrate the utility of the proposed approach. By enabling clinically relevant synthetic pathology data across institutions, EPDP directly addresses the challenges of data privacy and inter-institutional variability by providing a framework for generating clinically plausible and standardized synthetic datasets.

Methods

We propose a comprehensive framework for synthesizing and evaluating high-resolution patch-level histopathological images of gastric and breast cancers using a stable diffusion-based generative model [16]. The methodological pipeline includes four key stages:

- **Real data:** Construction and preprocessing of large-scale, pathologist-curated datasets.
- **Synthetic data:** Generation of synthetic patches via a denoising U-Net integrated within a diffusion model architecture [17, 18].
- **Color normalization:** Color normalization of synthetic patches based on unpaired image translation and reference selection using CycleGAN.
- **Evaluation:** Quantitative and qualitative assessment of synthetic data utility through real and synthetic data classification and expert evaluation.

Initially, we collected a large dataset of whole-slide images (WSIs) of hematoxylin and eosin (H&E)-stained 20x slides from four tertiary medical institutions. Board-certified pathologists reviewed and selected diagnostically relevant regions to extract high-quality image patches. These curated patches served as the foundation for training a histologic-type-specific stable diffusion model, which incorporated a customized residual block within a U-Net backbone for

enhancing the modeling of nuclear details and tissue architecture. We employed a CycleGAN-based color-normalization strategy to mitigate intersample color variability and promote visual consistency between generated images [19, 20]. This process involved automatic reference-image selection and the construction of pseudo-paired datasets to guide color translation.

For downstream evaluation, we trained two Efficient-NetV2-L classifiers [21] on real and synthetic patch datasets for evaluating their performance on real and synthetic data configurations. We evaluated how well synthetic data supported diagnostic tasks relative to real pathology images by assessing the classifier performance on diagnostically curated validation patches of real data. Further, we conducted a perceptual evaluation using a visual Turing test (VTT) that involved two independent board-certified pathologists recruited externally through the National Information Society Agency (NIA), without prior exposure to the training process or construction of the dataset, and whose identities were not disclosed to the research team [22].

Dataset Description and Preprocessing

The dataset used in this study comprised 5,729 WSIs stained with H&E (1,692 and 4,037 WSIs of gastric and breast cancer, respectively). These images were sourced from four tertiary medical institutions with institutional review board approval to ensure ethical compliance: Gachon University Gil Medical Center (GBIRB2024-121), Seoul St. Mary's Hospital of the Catholic University of Korea (20230927-C-292), Korea University Anam Hospital (2024AN0275), and Ajou University Hospital (AJOU-IRB-DB-2024-247). All pathology slides were digitized using institution-specific whole-slide scanners, including Panoramic 250 Flash III (3DHISTECH, Hungary), Hamamatsu NanoZoomer S360 (Hamamatsu Photonics, Japan), and Leica Aperio GT450 and AT2 (Leica Biosystems, USA).

Patch-level images were extracted from WSIs and classified according to histopathological diagnoses provided by board-certified pathologists at each institution. For breast cancer, the patches were classified into five categories [23]: BRNT (normal breast tissue), BRLC (lobular carcinoma in situ), BRDC (ductal carcinoma in situ), BRIL (invasive lobular carcinoma), and BRID (invasive ductal carcinoma). Gastric cancer cases were classified into four subtypes [24]: STNT (normal gastric tissue), STIN (intestinal-type adenocarcinoma), STDI (diffuse-type adenocarcinoma), and STMX (mixed-type adenocarcinoma).

Each 40× WSI was downsampled to 20× magnification and cropped into non-overlapping tiles of 1024×1024 pixels to construct the patch dataset. Only patches exhibiting diagnostically relevant features such as distinguishable nuclear morphology and tissue architecture were retained to

ensure data quality and relevance [25, 26]. The quantitative distributions of WSIs, raw patch images, and refined patch images per category are summarized in Table 1.

The patch dataset was built in a stepwise manner: raw patches were extracted from WSIs, and they were subsequently refined into a high-quality subset (refined patches) through two rounds of expert review. The first round was conducted by senior pathologists with more than ten years of experience and subspecialty in breast or gastrointestinal pathology from each participating institution. In this round, pathologists independently selected representative 20× magnification WSIs from their institutional cases, followed by extracting key diagnostic patches. For gastric cancer, patches were selected to illustrate Laurén subtypes, including cohesive glandular structures in the intestinal type, discohesive growth with poorly cohesive and signet-ring cells in the diffuse type, and overlapping features in the mixed type with normal gastric mucosa [27]. For breast cancer, patches were selected to reflect key diagnostic categories, including the distinction between ductal and lobular morphology (cohesive glandular structures versus discohesive single-file growth) and the separation of invasive lesions from in situ lesions, based on the loss of the myoepithelial cell layer and stromal invasion versus confinement within ducts or lobules, along with normal breast tissue [28, 29]. Following the initial filtering, the selected patches underwent a second-round consensus review by the full panel of participating pathologists to ensure inter-institutional consistency and standardization. This review was not blinded to class labels for preserving the histological-context alignment.

The number of refined patches per class was determined based on an expert consensus rather than a fixed quota. A minimum of 1,000 representative patches per subtype were selected whenever possible. All retained patches were required to have less than 25% of their area occupied by

Table 1 WSIs and patch-level image statistics used for synthetic data generation

Cancer type	Subtype	WSI	Whole patch image	Refined patch image
Breast	BRNT	1,374	147,583	3,594
	BRLC	217	23,267	1,150
	BRDC	139	24,963	2,305
	BRIL	114	33,848	1,062
	BRID	2,193	739,552	2,518
Total		4,037	969,213	10,629
Gastric	STNT	218	5,449	1,278
	STIN	502	230,408	1,346
	STDI	879	378,198	1,253
	STMX	93	31,010	1,464
Total		1,692	645,065	5,341

diagnostically irrelevant regions (e.g., background or artifacts), which helped ensure high pathological fidelity. This rigorous curation process was essential to construct a reliable training dataset for developing generative models and subsequently evaluating diagnostic validity.

Design of the Stable Diffusion Model for Histopathological Image Generation

A stable diffusion-based generative model [16, 30] was developed to synthesize patch-level histopathological images for breast and gastric cancers. Separate models for each organ were trained to capture disease-specific morphological patterns. Figure 1 shows the overall architecture of EPDP, which incorporates class-conditional embeddings, a

customized U-Net denoising, and sampling based on a diffusion probability model (DDPM).

Unlike latent diffusion models that employ variational autoencoders (VAEs) for feature compression, EPDP is designed as a pixel-space diffusion framework. The model performs the denoising process directly on the image pixels, utilizing a customized U-Net backbone to capture spatial features without latent-space dimensionality reduction.

The model operates under a class-conditional diffusion framework. Cancer-type labels (e.g., invasive ductal carcinoma or diffuse-type adenocarcinoma) are embedded in a 256-dimensional latent space through a learnable embedding layer. These embeddings serve as conditional guidance throughout the denoising process. The generative backbone is a denoising U-Net [31–33], which receives a noisy image

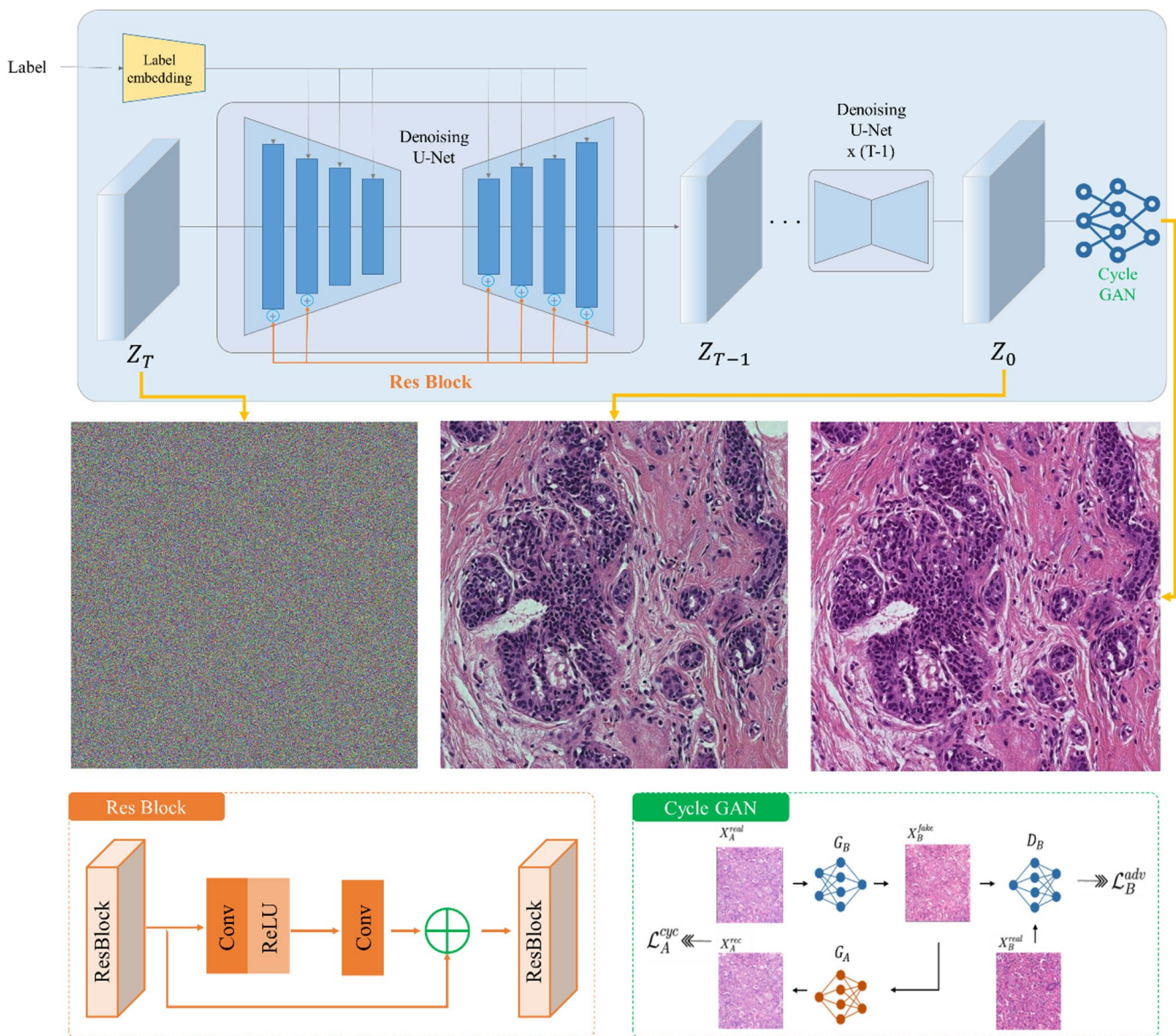


Fig. 1 Overall architecture of EPDP for class-conditional histopathological patch synthesis

and class embedding at each time step of the diffusion process. To preserve fine-grained pathological characteristics in a high-resolution 1024×1024 setting, strategically placed residual blocks are inserted into the upper two downsampling layers of the encoder and all decoder layers. This integration ensures that high-frequency features, such as nuclear chromatin texture, remain intact by mitigating the information bottleneck during the denoising process.

Image generation is conducted using DDPM, which iteratively transforms Gaussian noise into realistic image samples conditioned on the embedded class label. This iterative refinement enables the model to produce diverse, high-resolution, and diagnostically significant histopathologic patch images.

The diffusion process is configured with 1,000 time steps and a batch size of one, utilizing the Adam optimizer with a fixed learning rate to maintain numerical stability. Model training is performed using an NVIDIA A100-SXM4-80GB GPU. Full implementation details, including the U-Net architecture and training specifications, are available in the repository linked in the ‘Data availability’ section. The denoising U-Net has the following configuration: parameter count (EA) of 1,312,768, input size (MB) of 12.58, forward/backward pass size (MB) of 42,736.39, and parameter size (MB) of 715.22. The estimated total memory footprint of the model is $\sim 43,464.19$ MB.

Color Normalization with CycleGAN

In this study, a CycleGAN-based color normalization technique was used to address the absence of paired patch data. Ten reference images were selected under the supervision of pathology experts for standardization in color normalization, and dataset A was generated by extracting images with a histogram correlation greater than 0.9. Dataset B was selected randomly to match the size of dataset A for ensuring comparability. The total number of images used in this analysis is listed in Table 2.

Table 2 Dataset configuration information obtained for patch image color normalization

Cancer type	Subtype	Dataset A	Dataset B	Dataset $\hat{\beta}$
Breast	BRNT	426	351	351
	BRLC	0	100	100
	BRDC	0	426	426
	BRIL	386	572	572
	BRID	2,564	594	594
Total		3,376	2,043	2,043
Gastric	STNT	390	617	617
	STIN	0	632	632
	STDI	0	656	656
	STMX	712	530	530
Total		1,102	2,435	2,435

In this study, we used a CycleGAN-based color normalization technique to enhance the quantitative accuracy of medical image processing by leveraging the advantages of paired images over unpaired images. Previous research showed that paired image datasets achieved higher accuracy in image standardization tasks, especially in medical imaging, where color fidelity is crucial for diagnostic consistency [34, 35].

We developed an unpaired CycleGAN model using datasets A and B, which enabled the model to adaptively learn transformations without direct correspondence between images. This preliminary model was applied to dataset B, selectively excluding non-informative regions to create dataset B^* , which was a synthesized dataset with color distribution and diagnostic regions similar to those of dataset A. We used the RGB standard deviation and extracted white regions with a standard deviation below 5 to identify non-informative regions [36, 37], generating a mask of these areas. This mask was applied to align the non-informative regions consistently between datasets B and B^* . The resulting dataset B^* included synthesized images in which diagnostically irrelevant areas from dataset B were replaced to match the distributional characteristics of dataset A.

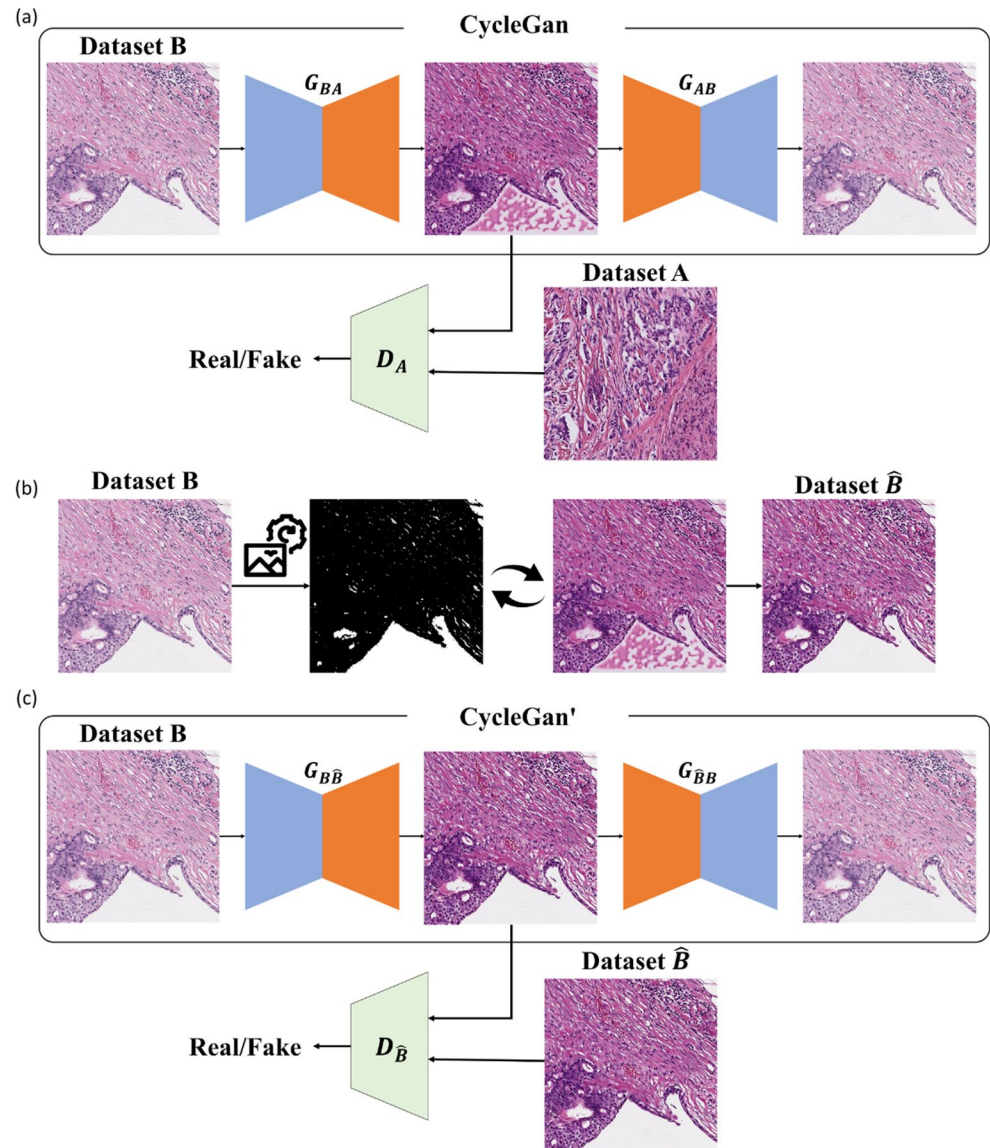
The same board-certified pathologists who validated the refined patches in Table 1 also conducted a review of processed images to ensure the pathological integrity of dataset B^* . This expert verification step was performed to confirm that the color-normalized and masked patches retained diagnostically meaningful structures and satisfied the quality standards required for downstream applications. We trained the final CycleGAN model for achieving enhanced color normalization using these paired datasets (datasets B and B^*). This approach facilitated robust alignment between images, enabling standardized color representation and improving diagnostic accuracy. This process is illustrated in Fig. 2.

Quantitative Evaluation of Image Similarity

We employed three commonly used quantitative metrics to objectively assess the quality of synthetic pathology images: FID [38], learned perceptual image patch similarity (LPIPS) [39], and structural similarity index measure (SSIM) [40]. All metrics were calculated across the entire dataset, and the final results were reported as average values.

FID was used to measure distributional similarity between the real and synthetic images [38]. We randomly split 50 real images into two groups and computed intra-real FID to establish a baseline. Subsequently, we calculated the FID between 50 real and 50 synthetic images for evaluating the divergence between the real and generated distributions. Lower FID scores indicated greater similarity.

Fig. 2 Color normalization process: (a) Unpaired image color normalization pipeline using CycleGAN. (b) Image processing workflow for non-informative region removal in pathology images. (c) Paired image color normalization pipeline using CycleGAN



In our evaluation, the FID was calculated at the organ level (i.e., breast and gastric) without further subdivision into specific histological subtypes. This approach was selected to evaluate the overall distributional similarity of images generated per cancer type instead of per subtype because of the limited number of images available in certain subcategories and focus on global fidelity in each cancer group [41].

Previous studies indicated that widely used metrics, such as FID designed for natural images, may be misleading when applied to medical or histopathology images because of domain-specific factors such as complex cellular structures, stain variability, and large background/tissue heterogeneity [42–44]. Therefore, we considered the relative gap between the real-to-real and real-to-synthetic distributions under identical evaluation conditions as a more appropriate measure of image fidelity

in our 1024×1024 patch setting, instead of using absolute FID thresholds.

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (1)$$

Perceptual similarity was assessed using the LPIPS metric, and the results were expressed as 1-LPIPS to facilitate intuitive comparison (higher values indicate greater similarity) [39]. For each cancer subtype, we computed 1-LPIPS scores between pairs of real–real images and between pairs of real–synthetic images. All pairwise values were averaged to obtain the final metric.

$$1\text{-LPIPS}(x, y) = 1 - \sum_l \frac{1}{H_l W_l} \sum_{h, w} w_l \odot (f_l^x(h, w) - f_l^y(h, w))^2 \quad (2)$$

Structural similarity was evaluated using SSIM [40]. For each subtype, we calculated the SSIM between the real–real

and real–synthetic image pairs. The average SSIM across all image pairs was used to represent the overall structural fidelity.

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3)$$

This set of quantitative evaluations enabled us to comprehensively assess the distributional consistency (FID), perceptual similarity (1-LPIPS), and structural coherence (SSIM) of the synthetic images, providing a robust analysis of their diagnostic utility.

Comparative Benchmarking Against Existing Generative Frameworks

EPDP was benchmarked against representative recent generative frameworks in digital pathology, including PathDiff, PathLDM, and DiffInfinite, under a controlled head-to-head setting. All baseline models were retrained using the same organ-specific subtype configuration, dataset split protocol, and number of generated images as those used for EPDP.

Quantitative comparisons were performed on the same real reference set using an identical evaluation pipeline for FID, 1-LPIPS, and SSIM. Qualitative comparisons were additionally performed using subtype-matched cellular-level image examples to assess the preservation of nuclear morphology, stromal architecture, and overall histologic fidelity.

Assessing Synthetic Data Utility through Real and Synthetic Dataset Interchange

We performed subtype-level multiclass classification experiments on breast and gastric cancers to evaluate the utility of synthetic data. Instead of simplifying the problem into a tumor versus non-tumor binary task, each histopathological subtype was treated as an independent class to more accurately assess if synthetic images preserved clinically meaningful morphological distinctions. For each cancer type, we trained two classification models, one using real patch images, and the other using EPDP-generated synthetic images while keeping all architectural and training configurations identical. EfficientNetV2-L, a high-performance convolutional neural network widely used for fine-grained medical image classification, was employed across all experiments to ensure a consistent evaluation of the subtype-level predictive performance [21].

For each cancer type, two EfficientNetV2-L classifiers were trained independently to compare the utility of real and synthetic data. The first model (EffNet-Real) was exclusively trained on real patch images, whereas the second

model (EffNet-Syn) was trained on EPDP-generated synthetic patch images corresponding to the same subtype categories. Both models were trained using datasets with identical class structures and preprocessing pipelines to maintain strict equivalence under experimental conditions. To ensure rigorous evaluation and prevent data leakage, the dataset was partitioned into training and validation sets at the WSI (slide) level prior to patch extraction. Specifically, all patches from a single WSI were assigned exclusively to either the training or the validation set, ensuring that no morphological information from the same slide appeared in both distributions. All evaluations were conducted on a shared validation set composed solely of real images, which enabled a controlled assessment of how well synthetic data can support subtype-level classification. Model performance was quantified using F1 scores and per-class metrics for enabling direct and interpretable comparison between real-trained and synthetic-trained classifiers.

This cross-evaluation approach enabled us to assess the generalizability and diagnostic relevance of the synthetic dataset [45]. We aimed to determine how closely synthetic images approximate real-world pathology and if they can effectively support downstream diagnostic tasks by quantifying the performance difference between the trained models EffNet-Real and EffNet-Syn.

Table 3 summarizes the composition of the training and validation datasets used for evaluating synthetic data utility. EffNet-Real was trained using real patch images, whereas EffNet-Syn was trained exclusively on synthetic patches generated using the EPDP model [22, 46, 47]. For both breast and gastric cancer experiments, a subset of real patch images for each subtype was reserved to construct a shared validation set, ensuring that both models were evaluated on identical real-image distributions. This standardized validation protocol enabled fair and unbiased comparisons between the models trained on real and synthetic data.

Table 3 Composition of training and validation datasets for evaluating synthetic data utility

Cancer type	Subtype	Real patch image	Synthetic patch image	Real patch images for validation
Breast	BRNT	2,883	6,504	711
	BRLC	927	3,033	223
	BRDC	1,851	4,218	454
	BRIL	830	1,652	232
	BRID	2,013	5,909	505
Total		8,504	21,316	2,125
Gastric	STNT	1,005	2,637	273
	STIN	1,096	2,538	250
	STDI	995	2,938	258
	STMX	1,176	2,747	288
Total		4,272	10,860	1,069

In addition to this real-versus-synthetic comparison, we further assessed the augmentation effect of EPDP-generated synthetic images through real–synthetic mixing experiments. For each cancer type, the total number of training images was fixed at $N=1,000$, and the real-to-synthetic ratio was varied across five settings: real only, real 75% + synthetic 25%, real 50% + synthetic 50%, real 25% + synthetic 75%, and synthetic only. To avoid class imbalance, training samples were uniformly distributed across diagnostic subtypes within each cancer type. Specifically, breast cancer included 200 images per subtype across five subtypes, and gastric cancer included 250 images per subtype across four subtypes. All models were trained using the same EfficientNetV2-L architecture and evaluated on the same real-image validation sets used in Table 3.

Expert-Based VTT for Realism Evaluation

We assessed the perceptual realism of the generated synthetic pathology images by conducting a VTT involving two independent board-certified pathologists recruited externally through the NIA, each with more than ten years of diagnostic experience [46]. These experts had no prior involvement in the dataset curation or model training stages, which helps ensure an unbiased evaluation.

A total of 100 patch-level images were prepared for the test, comprising 50 real images and 50 synthetic images sampled from the diffusion model after CycleGAN-based color normalization. The images were presented one at a time in randomized order, and the reviewers were asked to independently classify each image as 'real' or 'fake' (synthetic) based solely on visual inspection, without additional clinical or contextual information. Figure 3 shows the graphical user interface used for the VTT, which presented a single patch image with binary real/fake response options.

The primary evaluation metric was classification accuracy, which was defined as the proportion of correctly identified real and synthetic images. In typical VTTs, human observers can distinguish real images from synthetic images with high accuracy (often more than 80%) when visual differences are clear. In contrast, an accuracy score within the range of 40%–60% is interpreted as an indication that the observer cannot reliably differentiate between two classes [22, 46, 47]. Therefore, if the precision of the pathologists falls within this range, it indicates that synthetic images exhibit a high degree of visual similarity to real pathology images and are effectively indistinguishable.

The VTT results provide an independent, expert-based validation of the visual fidelity of generated images and complement quantitative classification performance results

described in the previous section. Collectively, these evaluations offer a holistic assessment of both the functional and perceptual qualities of synthetic pathology data.

For statistical analysis of the VTT results, inter-rater agreement between the two pathologists was assessed using Cohen's kappa coefficient. The calculation was based on the binary real/fake decisions assigned by the two pathologists for each image, rather than on the agreement between each pathologist and the ground-truth label. Cohen's kappa was defined as:

$$\kappa = (Po - Pe) / (1 - Pe),$$

where Po represents the observed proportion of agreement between the two pathologists, and Pe represents the expected agreement by chance based on the marginal response distributions of the two raters. In addition, the ability of each pathologist to distinguish real from synthetic images was evaluated using an exact two-sided binomial test. For each pathologist and cancer type, the number of correctly classified images was modeled as $X \sim \text{Binomial}(n, 0.5)$, where n is the total number of VTT images and 0.5 represents chance-level performance in a binary real/fake classification task. A two-sided p-value was used to determine whether the observed accuracy differed significantly from random guessing.

Results

We performed comprehensive quantitative and qualitative evaluations to validate the quality and diagnostic relevance of synthetic pathology images generated by the EPDP model. We evaluated the visual similarity between the synthetic and real images using established image quality metrics (FID, 1-LPIPS, and SSIM). Subsequently, we examined the practical utility of synthetic data in downstream classification tasks by training and testing a convolutional neural network (EfficientNetV2-L) on real (EffNet-Real) and synthetic (EffNet-Syn) datasets. Finally, we performed VTT with board-certified pathologists to assess the perceptual realism of generated images.

Quantitative Evaluation of Synthetic Image Quality

We performed a quantitative evaluation using three standard image quality metrics to assess the similarity between synthetic and real pathology images generated by the EPDP model: FID for distributional similarity, 1-LPIPS for perceptual similarity, and SSIM for structural consistency. These metrics were calculated for the cancer subtypes in the breast and gastric datasets, and the results

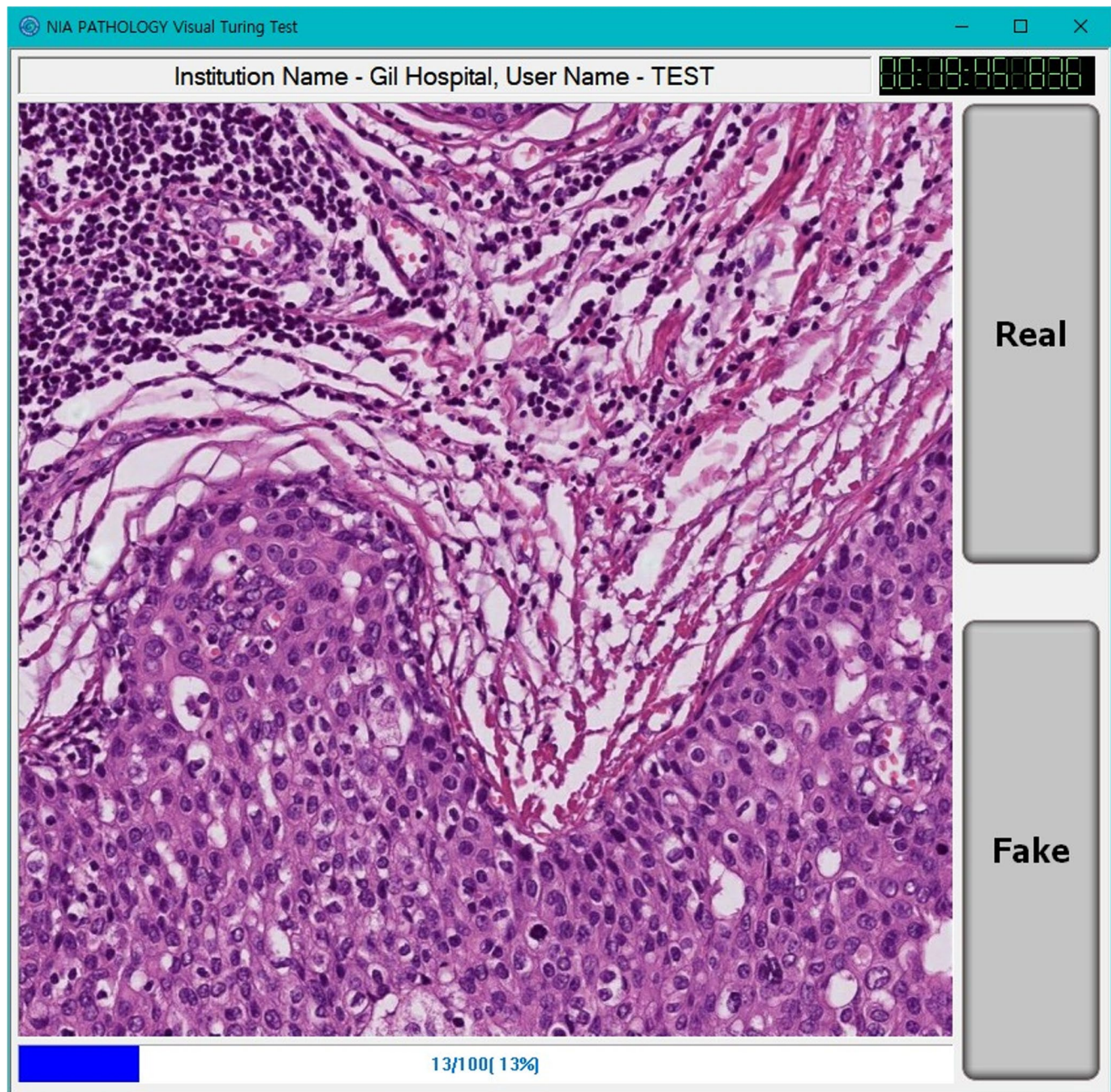


Fig. 3 Visual Turing test program graphical user interface

are summarized in Table 4. Furthermore, a comparative benchmark analysis with other generative frameworks is presented in Table 6.

We calculated the relative difference (Rel. Diff.) for each metric to quantify the degree of similarity between real–real and real–synthetic image pairs as follows:

$$\text{Rel. Diff. (\%)} = \left| \frac{S_{\text{real-syn}} - S_{\text{real-real}}}{S_{\text{real-real}}} \right| \times 100 \quad (4)$$

where $S_{\text{real-syn}}$ and $S_{\text{real-real}}$ represent metric values for pairs of real and real images, respectively. This provided a normalized measure of the deviation between the synthetic and real image similarity scores. The metric values for diagnostic subtype categories in breast and gastric cancers indicated a high level of similarity between the synthetic and real images with differences ranging from 0% to 12% depending on the metric and subtype.

For FID, real–synthetic comparisons yielded lower scores of 11.0% in breast cancer and 8.6% in gastric cancer

Table 4 Quantitative similarity metrics (mean $\pm$ standard deviation) between real and synthetic pathology images

Evaluation metrics	Cancer type	Subtype	Real vs. Real	Real vs. synthetic	Relative difference (%)
FID	Breast	Whole	107.936 (± 4.699)	96.050 (± 2.688)	11.0
	Gastric	Whole	104.790 (± 5.939)	95.790 (± 3.424)	8.6
1-LPIPS	Breast	BRNT	0.533 (± 0.067)	0.512 (± 0.063)	3.9
		BRLC	0.545 (± 0.054)	0.519 (± 0.055)	4.8
		BRDC	0.575 (± 0.064)	0.529 (± 0.055)	8.0
		BRIL	0.471 (± 0.070)	0.451 (± 0.063)	4.2
		BRID	0.487 (± 0.062)	0.476 (± 0.055)	2.3
		STNT	0.536 (± 0.067)	0.507 (± 0.063)	5.4
	Gastric	STIN	0.552 (± 0.061)	0.485 (± 0.060)	12.1
		STDI	0.479 (± 0.074)	0.500 (± 0.061)	4.4
		STMX	0.520 (± 0.058)	0.494 (± 0.053)	5.0
		SSIM	Breast	BRNT	0.979 (± 0.007)
BRLC	0.981 (± 0.006)			0.982 (± 0.006)	0.1
BRDC	0.980 (± 0.008)			0.983 (± 0.006)	0.3
BRIL	0.983 (± 0.005)			0.983 (± 0.005)	0.0
BRID	0.982 (± 0.006)			0.983 (± 0.006)	0.1
STNT	0.981 (± 0.009)			0.981 (± 0.007)	0.0
Gastric	STIN		0.981 (± 0.006)	0.981 (± 0.005)	0.0
	STDI		0.982 (± 0.006)	0.980 (± 0.006)	0.2
	STMX		0.983 (± 0.006)	0.982 (± 0.006)	0.1

compared with their real counterparts, suggesting that the generative process might have reduced staining variability or intraclass noise. Although this apparent improvement indicates strong distributional concordance, it can reflect smoothing effects inherent to diffusion models rather than a true increase in pathological diversity.

In terms of 1-LPIPS, differences between the real and synthetic image pairs ranged from 2.3% to 8.0% across breast cancer subtypes, with BRDC showing the largest difference (8.0%). Among the gastric cancer subtypes, STIN exhibited the highest difference (12.1%), whereas STDI showed the smallest increase (4.4%), followed by STNT (5.4%) and STMX (5.0%) in the real–synthetic versus real–real comparison.

The SSIM analysis revealed that differences between the real and real synthetic comparisons were consistently small in all subtypes, remaining within 0.3%. This suggests that synthetic images retained essential structural patterns, which makes them visually and anatomically consistent with real samples. Furthermore, a controlled head-to-head benchmark analysis against representative generative frameworks, including PathDiff [50], PathLDM [49], and DiffInfinite [48], was conducted under the same experimental setting used for EPDP. As summarized in Table 6, quantitative comparisons were performed using matched subtype-specific samples and an identical evaluation pipeline to assess the relative fidelity of the generated pathology images. Corresponding qualitative comparisons of morphological realism at the cellular level are presented in Fig. 6.

Classification Performance Using Synthetic Data

We evaluated classification performance using models trained on real pathology images (EffNet-Real) and synthetic images generated by the EPDP model (EffNet-Syn). Assessments were conducted independently for breast and gastric cancers, and the performance was quantified using precision, recall, specificity, and F1 score with corresponding 95% confidence intervals, as summarized in Tables 5 and 6.

In breast cancer classification, EffNet-Syn achieved a macro-averaged F1 score of 0.856, which was slightly higher than that obtained by EffNet-Real (0.843). The weighted F1 score showed comparable performance between training sources, with 0.887 for synthetic-trained models, and 0.880 for real-trained models. Among five breast cancer subtypes, BRIL demonstrated the highest performance improvement when trained using synthetic images with an F1 increase of 0.057. In gastric cancer classification, EffNet-Syn achieved a macro-averaged F1 score of 0.834 compared to 0.822 for EffNet-Real, maintaining a similar level of weighted F1 performance (0.832 vs. 0.821). Performance differences across the four gastric cancer subtypes remained within the confidence ranges of the trained model Table 7.

To further evaluate the effect of real–synthetic training composition, we compared classification performance across five training settings while fixing the total training size at $N=1,000$. As summarized in Table 8, performance gradually increased as the proportion of synthetic images

Table 5 Subtype classification performance with real vs. synthetic training data ($\pm 95\%$ confidence interval)

Cancer type	Subtype	Training data	Precision	Recall	Specificity	F1 Score	F1 Score Diff. (Syn – Real)
Breast	BRNT	Real	0.987 (± 0.006)	0.944 (± 0.015)	0.994 (± 0.003)	0.965 (± 0.014)	-0.002
		Synthetic	0.948 (± 0.014)	0.979 (± 0.008)	0.973 (± 0.007)	0.963 (± 0.014)	
	BRID	Real	0.901 (± 0.024)	0.844 (± 0.029)	0.971 (± 0.007)	0.871 (± 0.029)	0
		Synthetic	0.791 (± 0.030)	0.968 (± 0.012)	0.920 (± 0.012)	0.871 (± 0.029)	
	BRIL	Real	0.542 (± 0.048)	0.944 (± 0.023)	0.902 (± 0.013)	0.689 (± 0.060)	0.057
		Synthetic	0.860 (± 0.043)	0.659 (± 0.058)	0.987 (± 0.004)	0.746 (± 0.056)	
	BRLC	Real	0.957 (± 0.022)	0.700 (± 0.056)	0.996 (± 0.002)	0.808 (± 0.052)	0.031
		Synthetic	0.939 (± 0.027)	0.758 (± 0.052)	0.994 (± 0.003)	0.839 (± 0.048)	
	BRDC	Real	0.933 (± 0.020)	0.833 (± 0.032)	0.984 (± 0.005)	0.880 (± 0.030)	0.020
		Synthetic	0.942 (± 0.019)	0.861 (± 0.029)	0.986 (± 0.005)	0.900 (± 0.028)	
	Macro average	Real	0.864	0.854	0.968	0.843	0.013
		Synthetic	0.868	0.857	0.966	0.856	
	Weighted average	Real	0.961	0.843	0.990	0.880	0.007
		Synthetic	0.954	0.863	0.988	0.887	
Gastric	STNT	Real	0.935 (± 0.020)	0.736 (± 0.056)	0.985 (± 0.006)	0.821 (± 0.047)	0.034
		Synthetic	0.943 (± 0.018)	0.771 (± 0.052)	0.987 (± 0.005)	0.855 (± 0.043)	
	STIN	Real	0.821 (± 0.030)	0.859 (± 0.029)	0.911 (± 0.011)	0.839 (± 0.029)	0.013
		Synthetic	0.837 (± 0.027)	0.870 (± 0.027)	0.915 (± 0.010)	0.852 (± 0.028)	
	STDI	Real	0.833 (± 0.030)	0.803 (± 0.033)	0.928 (± 0.009)	0.818 (± 0.031)	-0.008
		Synthetic	0.837 (± 0.028)	0.785 (± 0.035)	0.930 (± 0.009)	0.810 (± 0.032)	
	STMX	Real	0.851 (± 0.027)	0.774 (± 0.038)	0.937 (± 0.009)	0.810 (± 0.034)	0.006
		Synthetic	0.845 (± 0.028)	0.793 (± 0.036)	0.935 (± 0.009)	0.816 (± 0.033)	
	Macro average	Real	0.860	0.793	0.940	0.822	0.012
		Synthetic	0.866	0.805	0.942	0.834	
	Weighted average	Real	0.865	0.793	0.939	0.821	0.011
		Synthetic	0.870	0.803	0.940	0.832	

Table 6 Performance benchmarks of various generative models at 1024 $\times$ 1024 resolution based on FID, SSIM, and 1-LPIPS metrics

Model	Cancer type	FID $\downarrow$	SSIM $\uparrow$	1-LPIPS $\uparrow$
Real(Baseline)	Breast	107.936 (± 4.699)	0.981 (± 0.006)	0.522 (± 0.063)
	Gastric	104.790 (± 5.939)	0.982 (± 0.007)	0.522 (± 0.065)
PathDiff [50]	Breast	124.32 (± 5.12)	0.942 (± 0.012)	0.382 (± 0.045)
	Gastric	119.54 (± 6.01)	0.938 (± 0.015)	0.364 (± 0.052)
PathLDM [49]	Breast	121.85 (± 4.87)	0.945 (± 0.011)	0.395 (± 0.042)
	Gastric	115.20 (± 5.44)	0.941 (± 0.014)	0.378 (± 0.048)
DiffInfinite [48]	Breast	117.67 (± 4.23)	0.958 (± 0.010)	0.422 (± 0.038)
	Gastric	113.45 (± 5.89)	0.952 (± 0.012)	0.465 (± 0.041)
EPDP (Ours)	Breast	96.050 (± 2.688)	0.982 (± 0.006)	0.497 (± 0.058)
	Gastric	95.790 (± 3.424)	0.981 (± 0.006)	0.496 (± 0.059)

increased. In breast cancer classification, the macro F1 score increased from 0.781 in the real-only setting to 0.851 in the synthetic-only setting, and the weighted F1 score

Table 7 Accuracy of the VTT for perceptual realism assessment

Cancer type	Pathologist	VTT accuracy (%)
Breast	Pathologist #1	56
	Pathologist #2	50
Gastric	Pathologist #1	54
	Pathologist #2	51

increased from 0.819 to 0.882. In gastric cancer classification, the macro F1 score increased from 0.759 to 0.833, and the weighted F1 score increased from 0.757 to 0.830.

Across both cancer types, synthetic-trained models demonstrated classification performance equivalent to or slightly higher than that of real-trained models.

VTT for Perceptual Realism

VTT was performed by two board-certified pathologists to quantitatively assess the perceptual realism of the synthetic pathology images generated by the EPDP model. Each expert evaluated a randomized set of real and synthetic images and labeled them as either “real” or “synthetic” based solely on visual inspection. No additional clinical or contextual information was provided during the assessment.

Table 8 Downstream classification performance under different real–synthetic training data compositions. ($N=1,000$)

Cancer type	Training composition	Precision	Recall	Specificity	Macro F1	Weighted F1
Breast	Real only	0.786	0.779	0.943	0.781	0.819
	Real 75% + Synthetic 25%	0.819	0.811	0.952	0.814	0.848
	Real 50% + Synthetic 50%	0.836	0.828	0.957	0.832	0.864
	Real 25% + Synthetic 75%	0.847	0.841	0.961	0.844	0.875
	Synthetic only	0.854	0.848	0.963	0.851	0.882
Gastric	Real only	0.764	0.755	0.919	0.759	0.757
	Real 75% + Synthetic 25%	0.798	0.790	0.931	0.794	0.791
	Real 50% + Synthetic 50%	0.819	0.810	0.938	0.814	0.811
	Real 25% + Synthetic 75%	0.830	0.822	0.942	0.826	0.823
	Synthetic only	0.836	0.829	0.944	0.833	0.830

Table 7 indicates that the classification accuracy of breast cancer images are 56% and 50% for pathologists # 1 and # 2, respectively. For gastric cancer images, the corresponding precisions were 54% and 51%, respectively. All values remained within $\pm 6\%$ of random chance (50%), indicating that experts were unable to reliably distinguish real images from synthetic images. A qualitative review suggested that the minority cases where pathologists correctly identified synthetic images were primarily due to subtle repetitive artifacts in the background stroma, rather than distortions in nuclear morphology.

To quantify inter-rater agreement in the VTT, Cohen's kappa coefficient was calculated based on the binary real/fake decisions of the two pathologists. The resulting kappa values were 0.12 for breast cancer and 0.15 for gastric cancer, indicating only slight agreement between the two experts. In addition, exact two-sided binomial tests showed that the pathologists' classification accuracies did not differ significantly from chance-level performance ($p > 0.05$ for all assessments).

Representative Visual Samples of Synthetic Pathology Images

Figures 4, 5, and 6 provide a comparative visual assessment of real and synthetic images generated by various models. Figures 4 and 5 display subtype-specific patches for breast and gastric cancers, respectively, highlighting EPDP's superior color fidelity and structural integrity. Figure 6 presents a high-magnification comparison at the cellular level.

Discussion

This study demonstrated the feasibility of using EPDP to generate diagnostically meaningful histopathological image patches for breast and gastric cancers. This approach integrated expert-annotated data, stain normalization using CycleGAN, and diffusion-based denoising U-Net to synthesize high-resolution pathology images that preserve subclass-specific morphological characteristics.

Comprehensive fidelity evaluations confirmed that EPDP-generated images were visually and structurally consistent with real pathology slides. Across the breast and gastric cancer datasets, the relative differences in FID [22, 46, 47], LPIPS [36], and SSIM [37] remained within an acceptable range of $\sim 10\%$, indicating a strong alignment in image distribution and tissue organization. Further, board-certified pathologists could not reliably distinguish synthetic images from real images in a VTT (50%–56% accuracy), supporting the high perceptual realism of generated images [22, 46, 47]. These results confirmed that EPDP effectively captured key histological features such as nuclear and cytologic morphology and tumor and stromal architecture, including the tumor–stromal interface, which made the synthesized data suitable for clinical image interpretation [22, 46, 47].

The downstream performance of synthetic data validated its practical utility in digital pathology. Classification models trained solely on synthetic images exhibited macro-averaged F1 scores within 1%–1.3% of those trained on real images across both cancer types. This performance parity indicated that EPDP-generated data preserved diagnostic cues required for subtype differentiation, including structural heterogeneity and cellular atypia. Such capabilities are particularly relevant in clinical practice, where expert annotation and data imbalances remain major bottlenecks to AI development [22, 46, 47]. By bridging these gaps with high-fidelity, subtype-specific images, EPDP demonstrates its potential to augment existing datasets by capturing key histological features across diverse subtypes, providing a supplementary resource for robust AI development.

Previous studies utilized GAN-based generation for histopathology focused on limited cancer types and often constrained by patch-scale fidelity and color instability [47]. While VAE-based latent diffusion models, such as PathDiff [50] and PathLDM [49], offer computational efficiency, the reconstruction process from latent representations can lead to the loss of high-frequency pathological details. In contrast, EPDP's pixel-space approach prioritizes the preservation of critical diagnostic features, including nuclear chromatin texture and the tumor–stromal interface. This

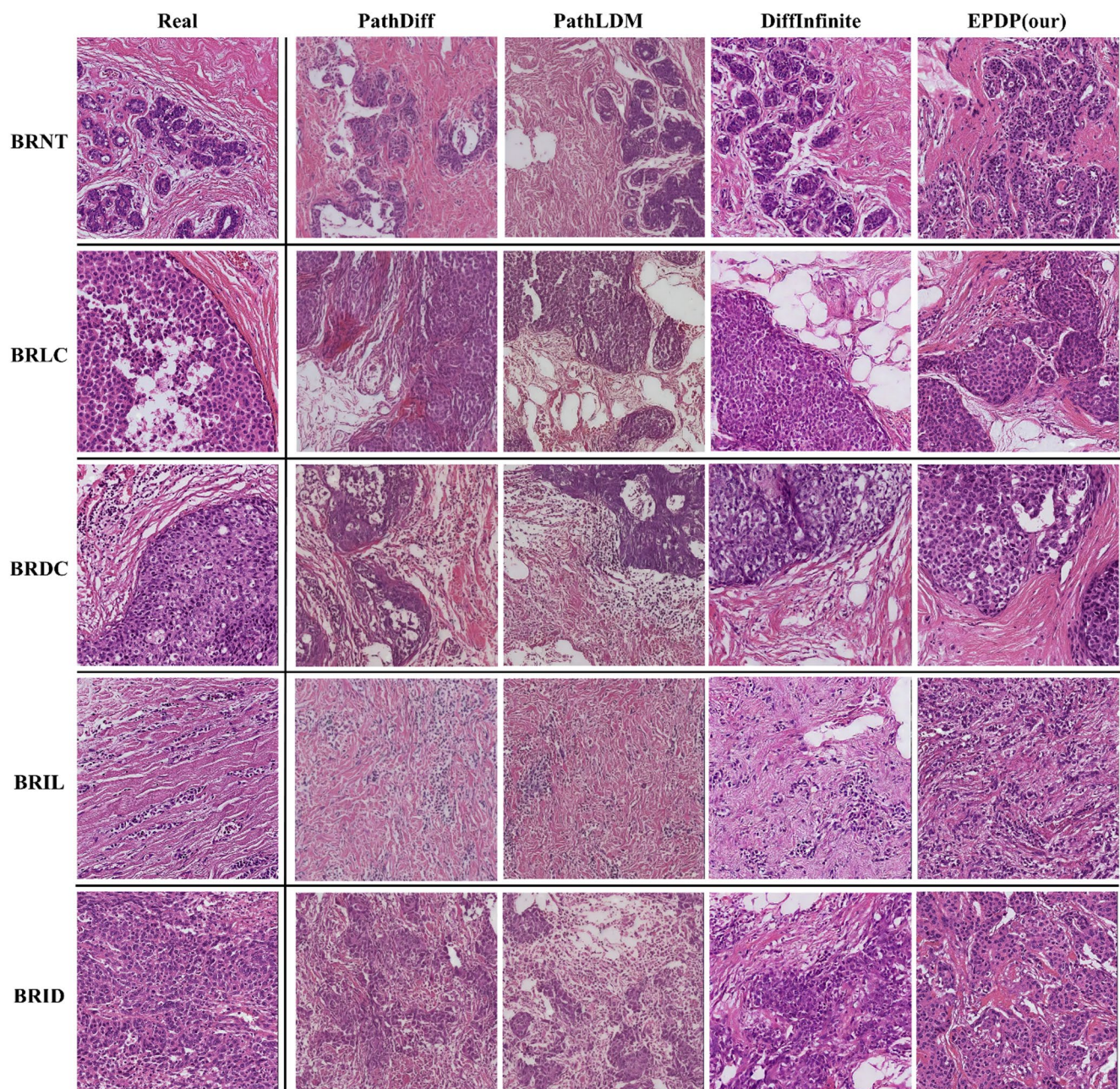


Fig. 4 Comparative visual analysis of real and synthetic breast pathology images across diagnostic subtypes

advantage was qualitatively confirmed through comparisons with PathDiff [50], PathLDM [49], and DiffInfinite [48], where EPDP maintained superior clarity in nuclear morphology. Specifically, by avoiding the information bottleneck inherent in VAE-based decoders, EPDP ensures that fine-grained cytological patterns remain intact and clinically relevant. In addition, most studies evaluated synthetic images using only one or two distributional metrics such as FID or inception score, without validating clinical utility through expert review or downstream tasks [6, 9, 42, 44]. In contrast, our class-conditional EPDP model demonstrated

robust cytologic and nuclear morphology, tumor and stromal architecture, and high fidelity in preserving diagnostic semantic information, as supported by a comprehensive multidimensional evaluation, including VTTs and diagnostic classification performance. These differences highlighted EPDP as a clinically relevant advancement over previous pathological synthesis methods.

EPDP integrates CycleGAN-based stain normalization, class-conditional embedding, and downstream diagnostic classification within a unified pipeline. The main strengths of EPDP arise from the denoising and iterative refinement

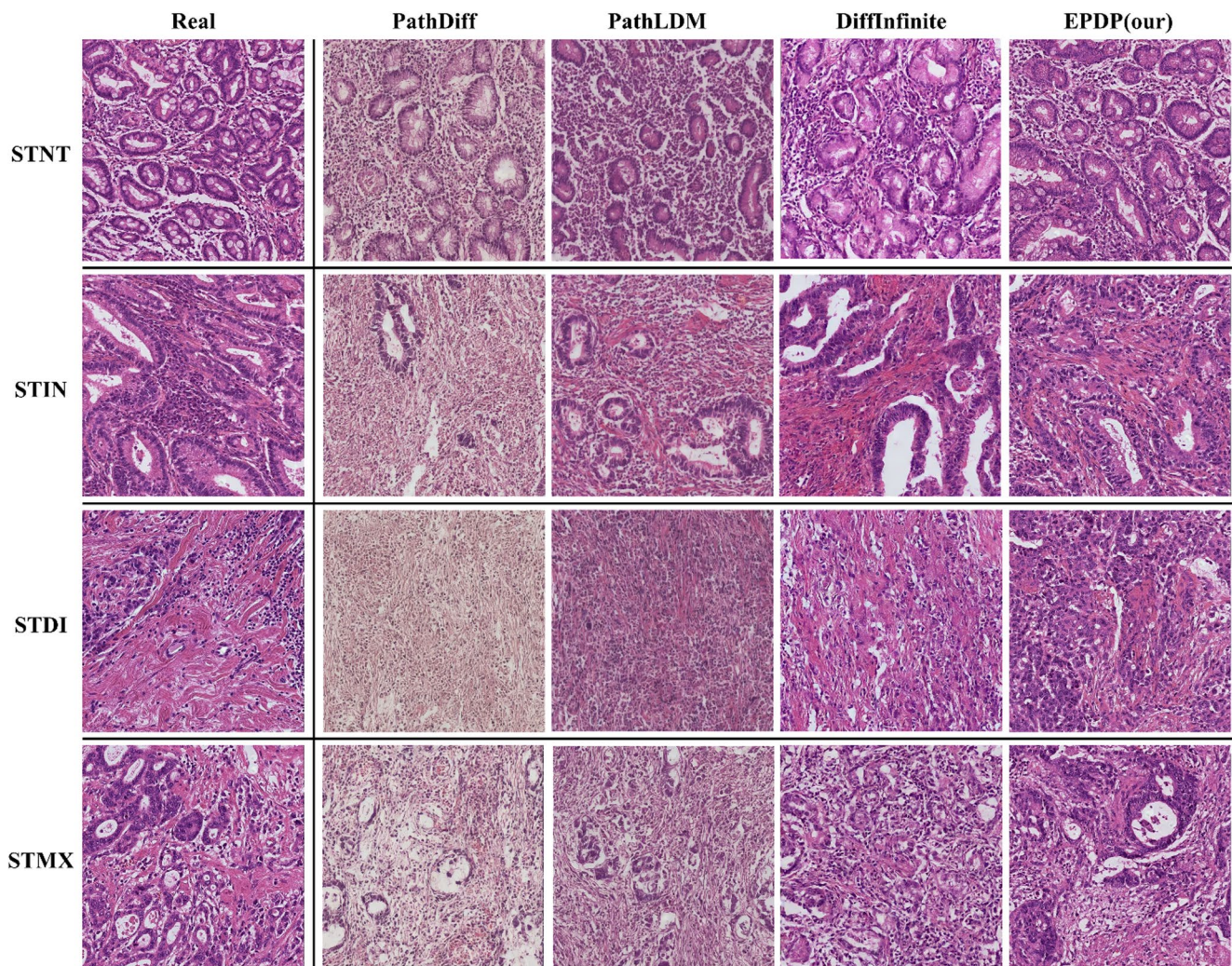


Fig. 5 Comparative visual analysis of real and synthetic gastric pathology images across diagnostic subtypes

mechanisms of diffusion models combined with explicit class-conditional embedding, which together support the generation of subtype-specific pathology images with preserved morphological detail. Within this framework, stain normalization was incorporated as a practical strategy to mitigate inter-institutional stain variability and improve visual consistency across multi-institutional pathology data. By reducing domain shift caused by heterogeneous staining protocols, this component further supports more robust performance of AI models trained on the synthetic datasets in real-world digital pathology workflows.

As illustrated in Fig. 6, EPDP maintained clearer nuclear morphology with a clear chromatin texture and distinct tumor–stromal interface compared to PathDiff [50], PathLDM [49], and DiffInfinite [48]. This superior performance stems from the pixel-space diffusion mechanism, which effectively bypasses the information bottleneck associated with latent-space reconstruction. Unlike previous studies that evaluated FID or inception scores alone, our

study applied a multifaceted evaluation protocol combining distributional, structural, perceptual, and downstream clinical assessments [20, 41]. This comprehensive validation indicated that EPDP-generated images were immediately applicable to real diagnostic workflows and research environments.

Despite the promising results reported in this study, there were several limitations. Although multi-institutional data from four tertiary hospitals were used, the reliance on a single national network and inclusion of only two cancer types may restrict the generalizability of our findings. Subtype imbalances persisted in certain categories, highlighting the need for targeted synthesis strategies that better represent rare or diagnostically challenging patterns.

The slight inter-rater agreement observed in the VTT (Cohen's kappa 0.12–0.15), together with near-chance classification accuracy, suggests that the two pathologists did not identify consistent visual cues that reliably distinguished synthetic images from real images. This finding supports the

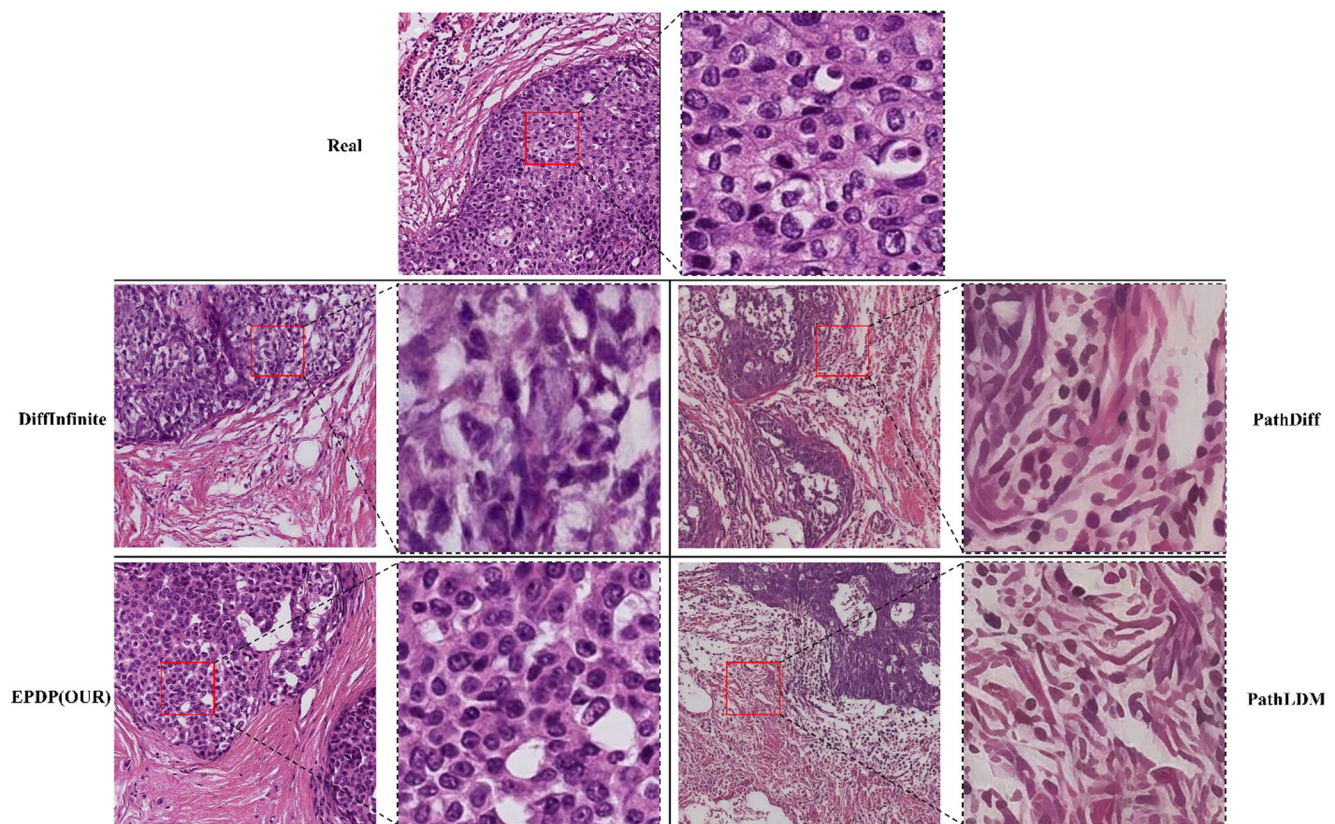


Fig. 6 Cellular level morphological comparison of real and synthetic breast pathology images generated by various generative frameworks

perceptual similarity of EPDP-generated images, although the limited number of expert reviewers indicates that broader multi-institutional reader studies are warranted for further validation. However, VTT involved a limited number of reviewers, suggesting that broader multi-institutional evaluation would strengthen perceptual validation. Future studies will extend EPDP to a more diverse set of cancers and institutions across multiple countries and expand from patch-level to whole-slide image generation.

While the current study demonstrated the effectiveness of EPDP using a large-scale curated dataset, the relationship between training data size and generative performance remains an important area for further exploration. Preliminary observations during the development phase suggested that the clinical validity of synthetic images was more sensitive to the quality of the curated patches (as summarized in Table 1) than to the raw volume of input data. Future investigations will focus on conducting multi-scale experiments to identify the minimum data requirements necessary to achieve diagnostic-grade synthesis, which will be particularly valuable for clinical applications where real-world data collection is logistically constrained.

The EPDP provides a scalable and effective framework to construct high-fidelity synthetic pathology datasets with potential diagnostic relevance. These findings support the use

of synthetic data as a complementary or alternative resource to real clinical data, reinforcing equitable AI development and facilitating research reproducibility. As digital pathology adoption accelerates, EPDP-generated synthetic images facilitate training, benchmarking, and hypothesis-driven studies while also supporting the development of more robust and generalizable AI solutions in clinical practice.

Author Contributions So Hyeon Lee contributed to the conceptualization of the study, manuscript review and editing, and visualization. Young Seop Lee developed the methodology and prepared the original draft of the manuscript. Young Jae Kim contributed to the manuscript review. Jisup Kim and Kwang Gi Kim supervised the study, managed project administration, and provided resources as corresponding authors. All authors reviewed and approved the final manuscript.

Funding This research was supported by the National Institute of Health(NIH) research project (project No.R25TA00471930-00),and by Gachon University Gil Medical Center (FRD2023-18).

Data Availability The synthetic histopathology datasets generated in this study are publicly available through the AI Hub platform (<https://aihub.or.kr/>). Additionally, the complete source code and implementation details for EPDP are available in our public GitHub repository: <https://github.com/Leeyoungsup/EPDP-Pathology>. The original histopathological image data used in this study are not publicly available due to legal and ethical restrictions. However, portions of the data may be available upon reasonable request after discussion with the corresponding author and subject to relevant institutional approval.

Declarations

Ethical Approval This study was approved by the Institutional Review Boards (IRBs) of all participating institutions: Gachon University Gil Medical Center (GBIRB2024-121), Seoul St. Mary's Hospital of the Catholic University of Korea (20230927-C-292), Ajou University Hospital (AJOU-IRB-DB-2024-247), and Korea University Anam Hospital (2024AN0275). The requirement for informed consent was waived by the respective IRBs because of the retrospective nature of this study. All experimental protocols were conducted in accordance with the relevant guidelines and regulations, and they strictly adhered to the ethical principles outlined in the Declaration of Helsinki.

Informed Consent The requirement for informed consent was waived because of the retrospective nature of the study design.

Additional Information Correspondence and requests for materials should be addressed to Kwang Gi Kim (kimkg@gachon.ac.kr) or Jisup Kim (jspath@gilhospital.com).

Competing interests The authors declare no competing interests.

References

- Li, Y. *et al.* Clinicopathological characteristics and risk factors of young-onset gastric carcinoma: a systematic review and meta-analysis. *Clinical and translational gastroenterology* **15**, e1 (2024).
- Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **74**, 229–263 (2024).
- Choi, S. & Kim, S. Artificial intelligence in the pathology of gastric cancer. *Journal of Gastric Cancer* **23**, 410 (2023).
- Tafavvoghi, M., Bongo, L. A., Shvetsov, N., Busund, L.-T. R. & Møllersen, K. Publicly available datasets of breast histopathology H&E whole-slide images: a scoping review. *Journal of Pathology Informatics* **15**, 100363 (2024).
- Song, Z. *et al.* Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nature communications* **11**, 4294 (2020).
- Alajaji, S. A. *et al.* Generative adversarial networks in digital histopathology: current applications, limitations, ethical considerations, and future directions. *Modern Pathology* **37**, 100369 (2024).
- Ding, K. *et al.* A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Scientific Data* **10**, 231 (2023).
- Mahmood, F. *et al.* Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE transactions on medical imaging* **39**, 3257–3267 (2019).
- Tschuchnig, M. E., Oostingh, G. J. & Gadermayr, M. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns* **1** (2020).
- Kazerouni, A. *et al.* Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis* **88**, 102846 (2023).
- Moghadam, P. A., Van Dalen, S., Martin, K. C., Lennerz, J., Yip, S., Farahani, H., & Bashashati, A. (2023). A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2000–2009).
- Falakhkhakh, K. *et al.* Deepfake histologic images for enhancing digital pathology. *Laboratory Investigation* **103**, 100006 (2023).
- Dolezal, J. M. *et al.* Deep learning generates synthetic cancer histology for explainability and education. *NPJ precision oncology* **7**, 49 (2023).
- Pozzi, M. *et al.* Generating and evaluating synthetic data in digital pathology through diffusion models. *Scientific Reports* **14**, 28435 (2024).
- Chen, L. *et al.* Conditional probabilistic diffusion model driven synthetic radiogenomic applications in breast cancer. *PLOS Computational Biology* **20**, e1012490 (2024).
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020).
- Saharia, C. *et al.* Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence* **45**, 4713–4726 (2022).
- Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. in *Proceedings of the IEEE international conference on computer vision*. 2223–2232.
- Shaban, M. T., Baur, C., Navab, N. & Albarqouni, S. in *2019 IEEE 16th international symposium on biomedical imaging (Isbi 2019)*. 953–956 (IEEE).
- Tan, M. & Le, Q. in *International conference on machine learning*. 10096–10106 (PMLR).
- Karras, T., Laine, S. & Aila, T. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- Tahir, M. *et al.* A Comprehensive AI-Based Approach in Classifying Breast Lesions: Focusing on Improving Pathologists' Accuracy and Efficiency. *Clinical breast cancer* (2025).
- Veldhuizen, G. P. *et al.* Deep learning-based subtyping of gastric cancer histology predicts clinical outcome: a multi-institutional retrospective study. *Gastric Cancer* **26**, 708–720 (2023).
- Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).
- Coudray, N. *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature medicine* **24**, 1559–1567 (2018).
- Mariette, C. *et al.* Consensus on the pathological definition and classification of poorly cohesive gastric carcinoma. *Gastric Cancer* **22**, 1–9 (2019).
- Siziopikou, K. P. Ductal carcinoma in situ of the breast: current concepts and future directions. *Arch Pathol Lab Med* **137**, 462–466 (2013).
- Contreras, A. & Sattar, H. Lobular neoplasia of the breast: an update. *Arch Pathol Lab Med* **133**, 1116–1120 (2009).
- Ding, Z., Zhang, M., Wu, J. & Tu, Z. in *The twelfth international conference on learning representations*.
- Zhang, Z. *et al.* in *2023 6th International Conference on Mechatronics, Robotics and Automation (ICMRA)*(185–190 (IEEE).
- Zhang, Y. *et al.* A unified conditional framework for diffusion-based image restoration. *Advances in Neural Information Processing Systems* **36**, 49703–49714 (2023).
- Gao, Q., Li, Z., Zhang, J., Zhang, Y. & Shan, H. CoreDiff: Contextual error-modulated generalized diffusion model for low-dose CT denoising and generalization. *IEEE Transactions on Medical Imaging* **43**, 745–759 (2023).
- Harms, J. *et al.* Paired cycle-GAN-based image correction for quantitative cone-beam computed tomography. *Medical physics* **46**, 3998–4009 (2019).
- Hiasa, Y. *et al.* in *International workshop on simulation and synthesis in medical imaging*. 31–41 (Springer).
- Tellez, D. *et al.* Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging* **37**, 2126–2136 (2018).

37. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics* **7**, 29 (2016).
38. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017).
39. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
40. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).
41. Wang, Y. *et al.* in *Proceedings of the European conference on computer vision (ECCV)*. 218–234.
42. Abdusalomov, A. B., Nasimov, R., Nasimova, N., Muminov, B. & Whangbo, T. K. Evaluating synthetic medical images using artificial intelligence with the GAN algorithm. *Sensors* **23**, 3440 (2023).
43. Cagas, W. *et al.* in *Proceedings of the Asian Conference on Computer Vision*. 207–217.
44. Skandarani, Y., Jodoin, P.-M. & Lalande, A. Gans for medical image synthesis: An empirical study. *Journal of Imaging* **9**, 69 (2023).
45. Bowles, C. *et al.* Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863* (2018).
46. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* **1**, 3 (2022).
47. Rademacher, P., Lengyel, J., Cutrell, E. & Whitted, T. in *Eurographics Workshop on Rendering Techniques*. 235–247 (1996).
48. Aversa, Marco, *et al.* “Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology.” *Advances in Neural Information Processing Systems* **36** (2023).
49. Yellapragada, Srikar, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. “Pathldm: Text conditioned latent diffusion model for histopathology.” In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5182–5191. (2024).
50. Bhosale, Mahesh, Abdul Wasi, Yuanhao Zhai, Yunjie Tian, Samuel Border, Nan Xi, Pinaki Sarder, Junsong Yuan, David Doermann, and Xuan Gong. “Pathdiff: Histopathology image synthesis with unpaired text and mask conditions.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22415–22424. (2025).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.