IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 February 2026, accepted 23 March 2026, date of publication 30 March 2026, date of current version 2 April 2026.

*Digital Object Identifier 10.1109/ACCESS.2026.3678764*

RESEARCH ARTICLE

# Mitigating Data Scarcity in Cancer Classification With Synthetic Data

**HUSSAIN AHMAD MADNI[1], HAFSA SHUJAT[2], SILVIA ZOTTIN[1], AXEL DE NARDIN[1], AND GIAN LUCA FORESTI[1], (Senior Member, IEEE)**

[1]Department of Mathematics, Computer Science, and Physics, University of Udine, 33100 Udine, Italy
[2]International Islamic University Islamabad, Islamabad 44000, Pakistan

Corresponding author: Hussain Ahmad Madni (hussain.madni@uniud.it)

This work was supported in part by the University of Udine Project on "Piano Stretegico Dipartimentale on Artificial Intelligence" (PSD-AI) Project at the University of Udine under Grant 2022-25, and in part by Friuli Venezia Giulia Region Project "Supporting the Diagnosis of Rare Diseases through Artificial Intelligence" under Project CUP F53C22001770002 and Project CUP F53C22001780002.

**ABSTRACT** Clinical datasets are often limited in availability and subject to strict privacy regulations, posing significant challenges for the development of accurate classification models. Synthetic data generation offers a promising alternative, enabling the augmentation of training datasets while preserving patient privacy. However, generating high-fidelity synthetic images that effectively support model training remains a significant challenge. In this paper, we focus on the classification of colorectal and lung carcinoma as representative tasks in cancer clinical diagnostics. We utilize a stable diffusion model enhanced with Low-Rank Adaptation (LoRA) weights to generate synthetic images from a limited number of real images used for fine-tuning. This method results in a performance improvement of the DeiT-L (Data-efficient Image Transformer–Large) and CLIP (Contrastive Language–Image Pretraining) models for colon and lung datasets, respectively, when trained on excessive synthetic data and a few real samples. Synthetic data closely mirrors the real samples, mitigating the issues of data scarcity and privacy, while enhancing model generalization. These findings support the use of synthetic data as a viable tool in cancer disease classification and demonstrate its potential to strengthen deep learning applications in clinical diagnostics and medical research. Experimental results on colon and lung histopathology datasets demonstrate that augmenting limited real data with diffusion-generated synthetic images consistently improves classification accuracy and generalization across multiple deep learning architectures, particularly in low-data (few-shot) regimes. Notably, the hybrid configuration combining a small number of real samples with synthetic data yields the most practically significant performance gains, highlighting its relevance for real-world clinical settings where annotated data are scarce. Code is available at: https://github.com/h-ahmad/rare_disease_classification

**INDEX TERMS** Colorectal cancer, histopathologic classification, lung cancer, medical image augmentation, synthetic medical data.

## I. INTRODUCTION

Within contemporary diagnostic practice for cancer diseases, pathologists remain indispensable for the microscopic evaluation of high-resolution histopathological images, where they must both distinguish malignant from non-malignant tissues and accurately subtype cancers such as colorectal carcinoma and lung carcinoma. Histological assessment of colorectal cancer is particularly challenging due to the presence of diverse tissue patterns, tumor heterogeneity, and technical artifacts that increase interpretative complexity and workload [1]. Similarly, morphologic differentiation between lung cancer subtypes, especially adenocarcinoma versus squamous cell carcinoma, places considerable demands on

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

pathologists and is known to suffer from inter- and intra-observer variability [2]. Such limitations underline the need for complementary diagnostic tools that can support reproducibility, reduce observer variability, and assist pathologists in routine clinical practice.

Over the past decade, deep learning-based computational techniques have been increasingly applied as adjunct tools capable of supporting automated tissue classification and cancer detection [3]. These approaches have demonstrated considerable potential in improving diagnostic workflows by enabling high-throughput image analysis, detecting subtle morphological patterns, and assisting in subtype classification. However, the performance of machine learning models is fundamentally influenced by the size, diversity, and quality of the training datasets. In the field of histopathology, obtaining large-scale annotated datasets is particularly challenging: annotation requires domain expertise, labeling is time-intensive, and the collection of representative cohorts is constrained by both clinical availability and institutional silos.

Furthermore, medical data are subject to stringent legal and ethical privacy regulations, which impose significant restrictions on data sharing and collaborative research [4], [5]. This results in fragmented datasets, limited accessibility, and potential biases, all of which hinder the development of generalizable and robust classification models. These limitations are particularly pronounced in studies focusing on specific cancer subtypes, where patient cohorts are smaller and morphological diversity is high.

To mitigate these challenges, synthetic data generation has emerged as a promising alternative. By producing realistic artificial samples that mimic the statistical and visual properties of real data, synthetic datasets can reduce direct exposure of identifiable patient records and provide opportunities for data augmentation in data-scarce domains [6], [7]; however, they do not inherently guarantee formal privacy protection without explicit privacy-preserving mechanisms. High-quality synthetic medical images cannot only complement real clinical datasets but also facilitate multi-institutional research by providing a privacy-preserving medium for sharing and benchmarking.

With the advent of text-to-image generative models [8], [9], [10], [11], [12], diffusion models have gained prominence as a more effective approach to generate synthetic data. Their capacity to generate high-resolution images that accurately depict the morphological characteristics of tissue structures makes them particularly suitable for medical applications [13], [14], [15]. Furthermore, diffusion models can produce synthetic data that retain the visual fidelity and unique biological features of the tissue [14]. This capability facilitates the creation of large datasets that mirror the morphological properties of tissues without compromising patient privacy; thus, offering a significant advancement for biomedical research and applications [16].

Despite recent advancements [17], generating high-fidelity synthetic tissue data from a limited number of samples (e.g., 5, 10, 20, or 50) remains a significant challenge. This issue can be addressed through the application of few-shot learning or generative models [18]. These methods hold substantial promise for training models capable of producing extensive synthetic datasets from minimal input. Unlike conventional data augmentation strategies such as color jittering, brightness, or geometric transformations, few-shot generative approaches are more effective in producing synthetic data that not only include novel variations but also finely tuned replicas of original samples.

Following the precise generation of synthetic images from a limited number of real samples, we train multiple convolutional and non-convolutional models exclusively on the synthetic dataset and compare the performance with those trained on the real dataset. The synthetic dataset is designed to be both balanced and representative, thereby demonstrating the efficacy of the data generation process.

Recent studies have explored data augmentation and generative modeling to address data scarcity in medical imaging. Traditional approaches rely on geometric and color-based transformations, which offer limited diversity and fail to capture complex tissue morphology. Generative adversarial networks (GANs) have been applied to histopathological image synthesis; however, they often suffer from training instability and mode collapse, particularly in few-shot settings. Diffusion-based models generate images through an iterative denoising process that progressively refines samples from noise, leading to stable training behavior and improved coverage of the underlying data distribution. Unlike adversarial training, this likelihood-based formulation reduces issues such as mode collapse and enables finer control over generated details, resulting in higher image fidelity and diversity. Nevertheless, most existing diffusion-based approaches require large-scale training data and full model fine-tuning, which is impractical in rare disease scenarios. Few-shot adaptation strategies such as LoRA have emerged as efficient alternatives but remain underexplored for histopathological synthesis. In parallel, transformer-based classifiers and multimodal models such as CLIP have shown strong performance in visual recognition tasks, yet their integration with synthetic medical data generation pipelines has received limited attention. Unlike prior work, our study systematically combines few-shot diffusion-based synthesis with LoRA adaptation and evaluates its impact across both convolutional and transformer-based classifiers, including multimodal models, thereby addressing a critical gap in the literature.

Despite significant progress in few-shot learning, existing methods primarily address either data-centric or model-centric adaptation strategies in isolation. Diffusion-based approaches have been explored for synthetic data generation and augmentation under limited data regimes, whereas parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) focus on adapting pretrained models with reduced computational cost. However, the literature lacks a systematic investigation of their joint effect under few-shot

conditions. Furthermore, prior work typically evaluates these techniques on a single model family, most commonly convolutional or transformer-based architectures, without examining their generalizability across heterogeneous classifiers or multimodal settings. As a result, it remains unclear whether diffusion-based data synthesis and parameter-efficient adaptation provide complementary benefits, how their interaction varies across architectures, and whether such integration improves multimodal classification performance under data scarcity. To address this gap, this work proposes a unified framework that integrates diffusion-based few-shot data synthesis with LoRA adaptation and evaluates their combined impact across convolutional, transformer-based, and multimodal classifiers under a consistent experimental setting. This enables a cross-architecture analysis of data-centric and parameter-efficient strategies that have not been systematically explored in prior methods. To clearly position our contribution relative to existing research directions, Table 1 summarizes the scope of representative study categories in few-shot learning, diffusion-based augmentation, and parameter-efficient fine-tuning. The comparison highlights that prior work typically investigates individual components in isolation, while a unified evaluation of diffusion-based data synthesis and parameter-efficient adaptation across heterogeneous architectures and multimodal models remains underexplored.

As shown in Table 1, existing approaches primarily focus on either data synthesis or model adaptation separately and often evaluate performance within a single architectural paradigm. In contrast, this work provides a unified framework that integrates diffusion-based few-shot synthesis with parameter-efficient adaptation and systematically evaluates their combined impact across convolutional, transformer-based, and multimodal classifiers.

Our main contributions are summarized as follows:

- We utilize a Stable Diffusion model to generate high-fidelity synthetic images, which are subsequently used to train a robust model for the accurate detection and classification of subtypes of colorectal and lung cancer.
- We introduce a synthetic data generation approach that leverages limited real samples, custom text prompts, and LoRA-based fine-tuning to produce data closely aligned with real images.
- We thoroughly evaluate the proposed image generation and classification pipeline to assess its effectiveness and overall performance.
- We evaluate the effectiveness of the proposed approach on two benchmark datasets: colorectal cancer and lung cancer subtypes.

## II. METHODOLOGY

The Methodology section is structured into two subsections: Data Generation, which outlines the procedures for constructing representative datasets, and Classification, which details the approaches employed for model training and evaluation.

### A. DATA GENERATION

In the initial phase of our methodology, we generate high-fidelity synthetic images representing cancer subtypes of colon and lung tissues, with the aim of closely replicating real images in terms of visual appearance and tissue morphology. The primary objective of this step is to generate high-fidelity synthetic samples through a diffusion-based generation process that progressively refines image quality and produces representative data for improving model performance. To achieve this, we employ the Stable Diffusion 2.1 model [12], a generative framework pre-trained on a large corpus of images, capable of producing high-quality visuals from random noise based on textual prompts. The diffusion model is trained by minimizing the denoising loss $\mathcal{L}_{\mathcal{D}}$, which measures the discrepancy between the predicted and true noise at each diffusion step. This loss is used exclusively during the training of the diffusion model to learn the data distribution and does not directly influence the downstream classification stage. While the model demonstrates strong capabilities for synthesizing realistic images, it struggles to generate accurate representations of cancer conditions such as colorectal carcinoma and lung carcinoma. The resulting images often lack the morphological fidelity and diagnostic features required for effective classification tasks.

The relevance of the DataDream method [18] lies in its ability to adapt large-scale diffusion models to domain-specific medical imagery under severe data constraints. Unlike conventional fine-tuning, DataDream leverages LoRA to update only a small subset of parameters, enabling efficient learning from very few histopathological samples. This makes it particularly suitable for rare disease scenarios, where full retraining is infeasible due to limited data availability and computational cost. To enhance synthetic data generation and produce images that closely resemble real ones, we exploit a promising approach introduced in DataDream method [18], which incorporates Low-Rank Adaptation (LoRA), a fine-tuning technique designed to efficiently adapt large pre-trained models. LoRA enables targeted modifications to models such as Stable Diffusion 2.1 [19] by introducing lightweight, yet effective, adjustments. This is achieved through the integration of low-rank matrices into the attention mechanisms of the text encoder and the diffusion U-Net. By doing so, the method enhances the model's capacity to emphasize the most relevant aspects of the input prompt, thereby eliminating the need for full model retraining. This approach modifies the conventional workflow for synthetic image generation by requiring the use of a textual prompt to produce synthetic samples, leveraging LoRA weights that have been fine-tuned on real data. The effectiveness of this technique lies in the combined application of text-to-image synthesis, powered by the Stable Diffusion 2.1 model, and image-to-image translation, enabled through the incorporation of the trained LoRA weights, as illustrated in Fig. 1 where the data $(x, y) \in D$ is provided as input, $x$ representing the input image, and $y$ is a label or class, and $C(y)$ is the caption

**TABLE 1.** Positioning of the proposed method relative to existing research directions.

| Category | Diffusion-based Data Synthesis | Parameter-efficient Adaptation (e.g., LoRA) | Cross-Architecture Evaluation (CNN + Transformer) | Multimodal Evaluation | Joint Analysis of Data + Model Adaptation |
|---|---|---|---|---|---|
| Diffusion-based augmentation | ✓ | × | Limited | Rare | × |
| Parameter-efficient fine-tuning | × | ✓ | Limited | Limited | × |
| General few-shot learning methods | Sometimes | Sometimes | Typically single architecture | Rare | × |
| Proposed method | ✓ | ✓ | ✓ | ✓ | ✓ |

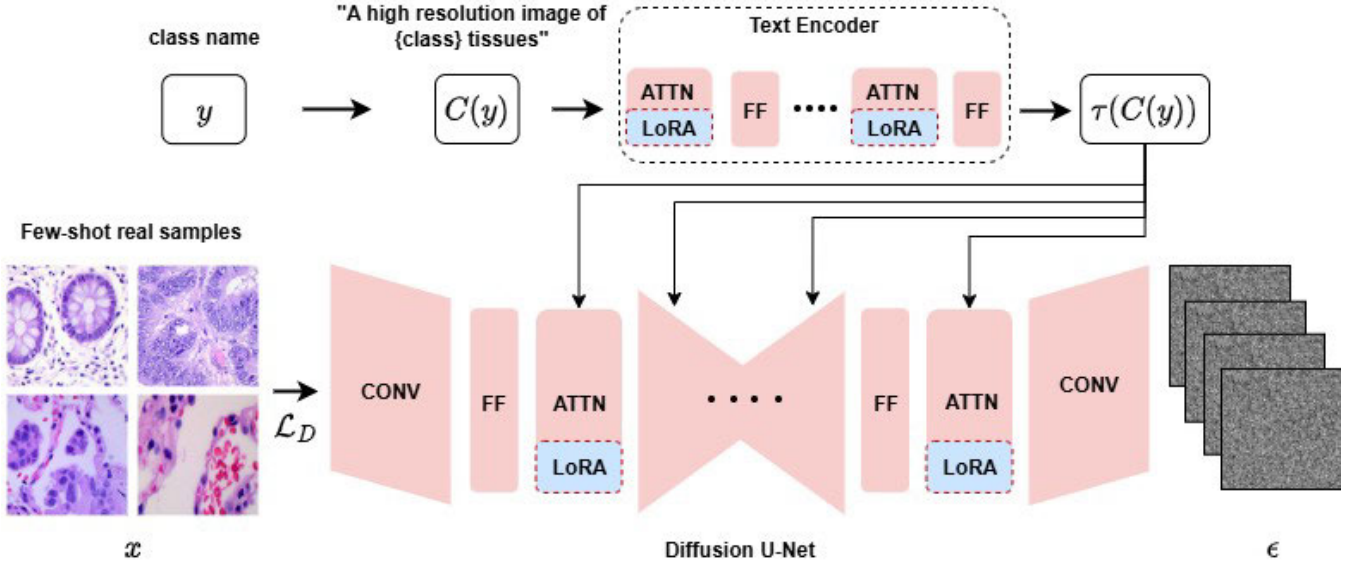



**FIGURE 1.** Overview of the proposed method. We perform fine-tuning of the LoRA parameters associated with the linear projections within the attention modules of both the text encoder and the diffusion U-Net. This optimization is aimed at enhancing image generation fidelity, enabling the model to produce outputs that more closely resemble the few-shot real images. Few-shot real samples are used as input $x$ where class names are used as label $y$, while $\epsilon$ represents intermediate noisy latent representations produced during the forward diffusion process. Diffusion U-Net is comprised of convolutional (CONV), feed-forward (FF), and attention (ATTN) layers.

of the image that describes it. Thus, $\tau(C(y))$ indicates the output transformed by the text encoder. Thus, the Stable Diffusion model captures the conditional data distribution by progressively removing Gaussian noise $\epsilon$ from the latent representation by minimizing the loss $\mathcal{L}_D$.

The image generation process begins with training the LoRA weights using a limited set of real images sampled from the publicly available LC25000 [20] dataset. To evaluate the impact of sample size, four separate experiments are conducted using 5, 10, 20 and 50 real samples per class. Given the variability in image quality across the dataset, it is critical to utilize the highest-quality samples during this phase. Therefore, images exhibiting the most distinct biological features are manually selected for each class to ensure the reliability of the training data.

Since synthetic data quality depends on prompt formulation, we evaluate a small set of candidate prompts and select the one yielding the best validation performance. This process follows standard hyperparameter selection practice, and all candidate and final prompts are explicitly provided to ensure reproducibility.

Furthermore, class-specific prompts are formulated to guide the generation of the corresponding synthetic samples. These prompts are designed to describe not only the class label but also the critical visual characteristics required to produce synthetic images that closely resemble real ones. An iterative trial-and-error strategy is conducted to refine and identify the most effective prompts for each class. The finalized prompts utilized in the proposed method are described below for the colon dataset, comprising adenocarcinoma (colon_acs) and benign colonic tissue (colon_n) class, and the lung dataset, comprising adenocarcinoma (lung_aca), squamous carcinoma (lung_scc), and benign (lung_n) tissue classes.

**"colon_aca":** "Ultra-high-resolution photorealistic histopathology slide of human colon adenocarcinoma tissue, hematoxylin and eosin (H&E) stained, microscopic view at 40x magnification, showing irregular glandular architecture, malignant epithelial cells with hyperchromatic nuclei, prominent nucleoli, high nuclear-to-cytoplasmic ratio, loss of normal crypt structure, desmoplastic stroma, visible mitotic figures, pink eosinophilic cytoplasm, blue-purple nuclear staining, intricate microvascular structures, sharp focus with ultra-detailed cellular morphology, scientifically accurate pathology style, vibrant yet realistic H&E color balance, aspect ratio 1:1 or 4:3, resolution 1024 × 1024, varied tumor grades and gland formation patterns".

**"colon_n":** "Ultra-high-resolution photorealistic histopathology slide of normal benign colonic mucosa, hematoxylin and eosin (H&E) stained, microscopic view at 40x magnification, showing well-organized crypt architecture with uniform tubular glands, evenly spaced goblet cells, nuclei small and basally located, low nuclear-to-cytoplasmic ratio, intact mucosal epithelium, normal lamina propria with sparse inflammatory cells, pink eosinophilic cytoplasm, blue-purple nuclear staining, sharp focus with ultra-detailed cellular morphology, scientifically accurate pathology style, vibrant yet realistic H&E color balance, aspect ratio 1:1 or 4:3, resolution 1024 × 1024".

**"lung_aca":** "Ultra-detailed, high-resolution histopathology slide of lung adenocarcinoma tissue, stained with hematoxylin and eosin (H&E), microscopic view at 40x magnification, vivid cell nuclei and glandular patterns, realistic color tones, crisp fine details, scientific medical imaging style, 'pathology laboratory quality, professional lighting, 8k resolution".

**"lung_n":** "Ultra-detailed and high-resolution histopathology slide of healthy benign lung tissue, stained with hematoxylin and eosin (H&E), microscopic view at 40x magnification, showing clear alveolar structures, thin alveolar walls, uniform cell nuclei, natural pink and purple tones, realistic medical imaging style, pathology lab quality, crisp fine details, 8k resolution".

**"lung_scc":** "Ultra-detailed, high-resolution histopathology slide of lung squamous cell carcinoma tissue, stained with hematoxylin and eosin (H&E), microscopic view at 40x magnification, showing keratin pearls, intercellular bridges, dense irregular sheets of malignant squamous cells, vivid pink and purple tones, realistic medical imaging style, pathology lab quality, crisp fine details, 8k resolution".

Prompt design followed an iterative trial-and-error process aimed at maximizing morphological fidelity of the generated histopathological images. Although multiple prompt variants were explored during development, a systematic prompt sensitivity analysis and reporting of failed prompt configurations were beyond the scope of this work.

To adapt the Stable Diffusion v2.1 model to the histopathological domain under few-shot conditions, we employ LoRA applied to the linear projection layers of the self-attention and cross-attention modules in both the diffusion U-Net and the text encoder. The rank of the LoRA matrices is fixed to $r = 4$, with a corresponding scaling factor $\alpha = 4$, providing a balance between parameter efficiency and representational capacity. During fine-tuning, only the LoRA parameters are optimized, while all original Stable Diffusion weights remain frozen. Optimization is performed using the AdamW optimizer with a learning rate of $1e{-}4$ and a weight decay of $1e{-}4$. LoRA fine-tuning is carried out for a maximum of 1,000 training steps, corresponding to approximately 20–30 epochs depending on the number of few-shot samples available per class. Training is terminated using early stopping, based on convergence of the diffusion loss and stabilization of the visual quality of generated samples, thereby preventing overfitting to the limited real examples.

### B. CLASSIFICATION

Following the generation of synthetic data, a series of experiments is conducted using a range of convolutional and non-convolutional models. Each model is independently trained on both real and synthetic datasets to facilitate a comprehensive evaluation. Training on real data served two primary purposes: first, to determine whether the models could perform effectively despite the limited number of samples per class; and second, to establish a performance baseline for addressing the central research question, whether synthetic data generated from a small number of real images can enhance model performance and support the detection. In contrast, training exclusively on synthetic data facilitates the assessment of whether models could learn and recognize the morphological characteristics embedded in the synthetically generated images.

To assess model performance on both real and synthetic datasets, we use a range of convolutional and transformer-based architectures. Convolutional Neural Networks (CNNs), particularly those with deep layers and residual connections, are extensively utilized in computer vision tasks due to their capacity to extract hierarchical feature representations. In our approach, CNN-based models are fine-tuned by adapting their final fully connected layers to perform binary and multi-class classification. These architectures are trained in a fully supervised manner to learn morphological characteristics from real and synthetic tissue samples.

Transformer-based models are designed to capture relationships throughout the image by using self-attention mechanisms, enabling them to understand global context

rather than relying solely on localized features as in CNN. This capability is particularly useful for the classification of rare diseases in tissues, such as differentiating between benign and malignant tissues, where the relevant features may be subtle and dispersed. We train vision-specific transformer architectures on both real and synthetic datasets to recognize morphological patterns linked to tissue conditions. These models demonstrate strong classification performance, with their ability to model long-range dependencies contributing significantly to effective feature learning during supervised fine-tuning.

CLIP (Contrastive Language–Image Pretraining) is a dual encoder model that simultaneously learns feature representations for both images and text through contrastive learning. We utilize two CLIP variants, ViT-B/16 and ViT-B/32, which are based on transformer architectures for image encoding, while the text encoder also exploits a standard transformer model. ViT-B/16 and ViT-B/32 differ in patch size, where ViT-B/16 processes images using smaller patches, enabling finer-grained spatial feature extraction at the cost of higher computational complexity, while ViT-B/32 uses larger patches, offering improved efficiency but reduced spatial resolution. This distinction allows us to analyze the trade-off between representation granularity and efficiency when applying CLIP to histopathological classification. Unlike conventional classification approaches, CLIP projects both visual inputs and textual prompts into a shared embedding space, enabling direct comparison between the two modalities. During inference, classification is performed by computing the similarity between the test image embedding and the embeddings of predefined, class-specific prompts. Finally, the label corresponding to the most similar prompt is assigned to the image. This design allows for a robust evaluation of how well the visual features align with the semantic content of the prompts, making it particularly suitable for tasks requiring cross-modal consistency. By analyzing model architectures across both real and synthetic data settings, the proposed approach aims to evaluate the generalizability of synthetic data and its effectiveness when incorporated into both conventional and prompt-based classification frameworks.

In addition to generic vision transformer architectures, we also evaluate the Data-efficient Image Transformer (DeiT) models, specifically DeiT-L and DeiT-B. DeiT is a vision transformer architecture designed to achieve strong performance with limited training data through knowledge distillation and efficient training strategies. The DeiT-L variant provides higher model capacity and is included to investigate the effectiveness of large-scale transformer representations in histopathological image classification under data-scarce conditions.

For all experiments, we use the cross-entropy loss function, a widely adopted objective in classification tasks due to its effectiveness in measuring the divergence between predicted class probabilities and ground truth labels. This loss function encourages the model to assign higher probabilities to the correct classes, thus improving classification accuracy. During training, whether on real or synthetic datasets, the cross-entropy loss is computed across the entire training set to guide model optimization. Moreover, the AdamW optimizer is used to update the model parameters via backpropagation, providing improved weight regularization over standard Adam. All models are trained for 50 epochs using a learning rate of $1e-4$.

All convolutional and transformer-based models reported in the experimental tables in the following section are fully implemented and fine-tuned by the authors under identical training and evaluation protocols. These models serve as comparative baselines to assess the effectiveness of the proposed synthetic data generation strategy and are not directly adopted from prior experimental results.

## III. EXPERIMENTS

### A. DATASET

In this paper, we use a publicly available histopathological image dataset LC25000 [20] treated as real data, which comprises digitized samples of lung and colon tissues representing both benign and cancerous subtypes. The colon subset is divided into two categories, colon adenocarcinomas and benign colon tissues, with $5,000$ images per class, resulting in a total of $10,000$ samples. Likewise, the lung subset consists of three categories, lung adenocarcinomas, squamous cell carcinomas, and benign lung tissues, each containing $5,000$ images, leading to a total of $15,000$ samples. All images are standardized to a resolution of $768 \times 768$ pixels, ensuring consistency across both malignant and non-malignant tissue representations for robust comparative evaluation.

Although the LC25000 dataset contains $5,000$ images per class, we intentionally subsampled 500 images per class to simulate realistic data-scarce clinical settings. The subset was selected using random stratified sampling to preserve class balance and avoid sampling bias. From this subset, 60% of the samples were used for training, 20% for validation, and 20% for testing.

As an initial step, each model is trained on the real subsets of the LC25000 data having a total of 1000 samples for the colon dataset and 1500 samples for lung tissues, where each class contains 500 images for both datasets. For this purpose, the images in each class are divided into 60% for training, 20% for validation to monitor potential overfitting and optimize hyperparameters, and 20% for testing, thus ensuring a balanced and systematic evaluation protocol. The models trained on real data are then evaluated to establish baseline performance and to enable a direct comparison with models trained on synthetic data. This experimental setup was designed to rigorously evaluate not only the classification accuracy but also the generalization capability of the proposed approach across real and synthetic domains.

To ensure fair comparison and avoid data leakage, the real dataset is partitioned once into training, validation, and

**TABLE 2.** A summary of data splits of both real and synthetic datasets used in the preliminary experiments.

| Dataset Type | Dataset | Class | Train Samples | Validation Samples | Test Samples | Total |
|---|---|---|---|---|---|---|
| Real (LC25000) | Colon | adenocarcinomas | 350 | 150 | 150 | 650 |
| | | benign | 350 | 150 | 150 | 650 |
| | Lung | adenocarcinomas | 350 | 150 | 150 | 650 |
| | | squamous cell carcinomas | 350 | 150 | 150 | 650 |
| | | benign | 350 | 150 | 150 | 650 |
| Synthetic | Colon | adenocarcinomas | 350 | 150 | 150 (real) | 650 |
| | | benign | 350 | 150 | 150 (real) | 650 |
| | Lung | adenocarcinomas | 350 | 150 | 150 (real) | 650 |
| | | squamous cell carcinomas | 350 | 150 | 150 (real) | 650 |
| | | benign | 350 | 150 | 150 (real) | 650 |

a fixed 20% held-out test set. This same real test set is used for evaluating both models trained on real data and models trained on synthetic data. For the synthetic training configuration, models are trained exclusively on synthetic images generated from the training portion of the real dataset, and no real test images are used during training.

Furthermore, to rigorously assess the classification performance and generalization capability of the proposed approach, we train models on synthetically generated histopathological data and subsequently evaluate their performance on real tissue samples. For each dataset, we construct balanced synthetic datasets by generating 500 samples per class, resulting in $1,000$ images for the colon dataset (comprising colon adenocarcinomas and benign colonic tissues) and $1,500$ images for the lung dataset (including adenocarcinomas, squamous cell carcinomas and benign lung tissues). From the synthetic data, 350 samples per class are allocated for training and 150 for validation, while an additional 150 real samples per class from the colon and lung datasets are reserved exclusively for testing. This experimental design intentionally reflects real-world clinical scenarios, where the availability of annotated medical data is often constrained due to privacy regulations, limited accessibility, and high annotation costs. The restricted number of real examples per class introduces a significant challenge for deep learning models, which typically require large-scale annotated datasets to achieve robust performance. In this context, the proposed method demonstrates its effectiveness by enhancing both generalization and classification accuracy under data-scarce conditions, thereby highlighting its potential applicability in practical medical imaging settings. The summary of data splits of both real and synthetic datasets is given in Table 2.

Prior to model training and evaluation, all images undergo a standardized preprocessing pipeline. Initially, each image is resized to a uniform resolution to meet the input dimensional requirements of the model. Subsequently, the intensity values of the pixels are normalized according to the model requirement to improve the training stability and accelerate the model convergence. For example, different values of mean and standard deviation are used for the ResNet-50 and CLIP model. To enhance the robustness and generalization capability of the model, conventional data augmentation techniques such as random rotations, color jittering, and horizontal flipping are applied exclusively to the training set, thus simulating natural variations encountered in real-world conditions. In contrast, no augmentation is performed on the validation and test sets to maintain the integrity and impartiality of the performance assessment.

### B. IMPLEMENTATION DETAILS

To evaluate the effectiveness of the available real dataset, classification tasks are performed using a range of deep learning architectures, including both CNN and transformer-based models. These models are fine-tuned using the AdamW optimizer, which integrates adaptive moment estimation with decoupled weight decay of $1e-4$ to enhance generalization performance. Each model is trained over 50 epochs with a learning rate of $1e-4$ and a batch size of 64 for all models except DeiT-L (Data-efficient Image Transformer-Large), for which we use a batch size of 32 due to the complex structure, applied consistently across the training, validation, and testing phases.

In the proposed approach, we take advantage of two fine-tuning strategies for the CLIP model, adapted to different data availability scenarios. Under limited real data conditions where only images and their corresponding class labels are available, the image encoder is fine-tuned using classification supervision, while the text encoder remains fixed due to the absence of prompt information. In contrast, when a sufficient synthetic dataset is available, including both labeled images and associated text prompts, we leverage the full capabilities of CLIP by fine-tuning both image and text encoders. To support efficient adaptation in both settings, we integrate LoRA, a parameter-efficient fine-tuning technique that introduces trainable low-rank matrices into the attention layers while keeping the pre-trained weights largely frozen. This approach significantly reduces computational overhead and enables effective model adaptation. In the real-data setting, LC25000 provides only image–label

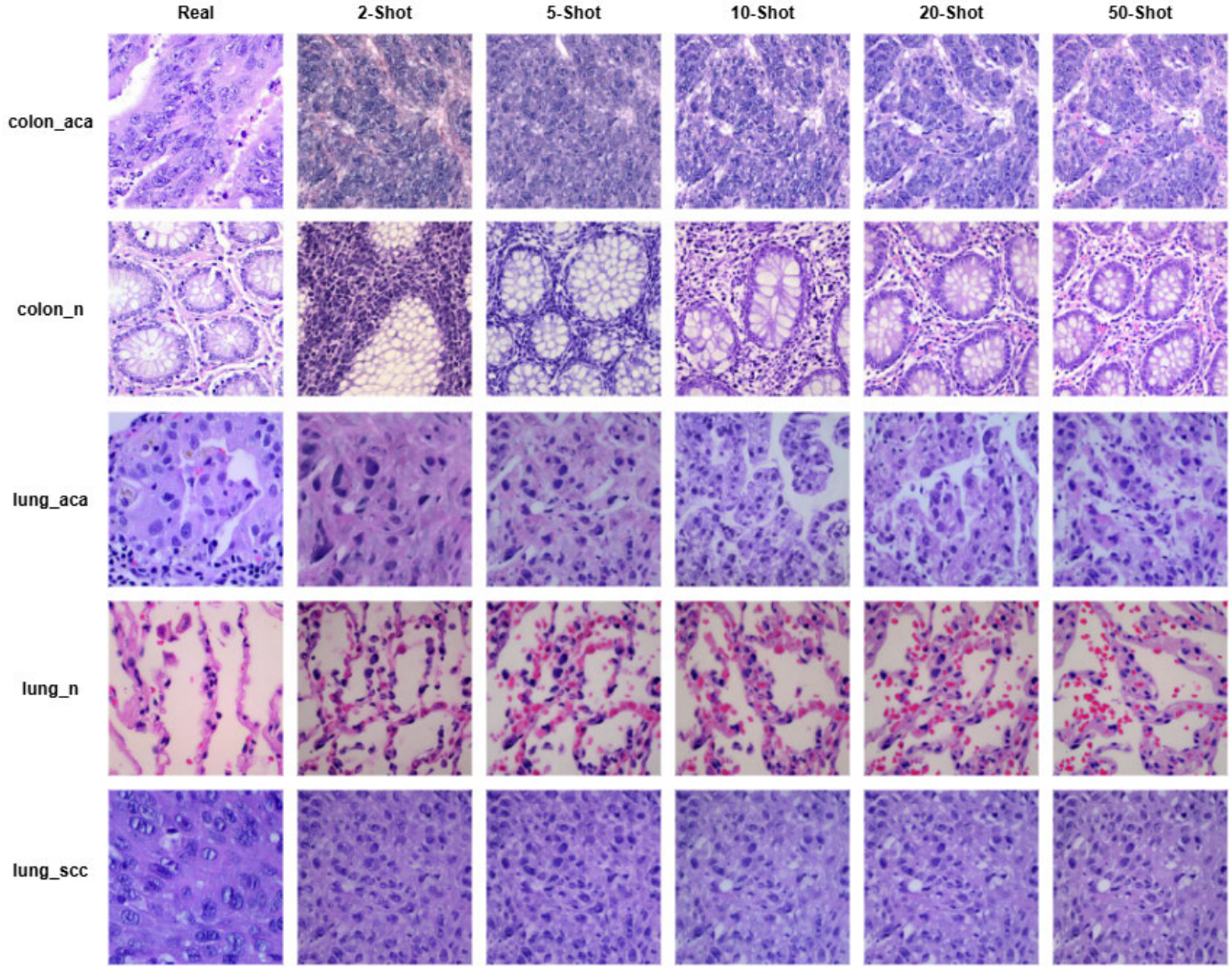



**FIGURE 2.** Comparison of synthetic images produced by stable diffusion v2.1 under varying few-shot conditions. The first column shows the real sample from each class. The top row displays randomly selected real examples from each class of both datasets used for guidance, while the subsequent rows present randomly selected synthetic samples generated for each class based on the respective few-shot inputs. The figure demonstrates that as the number of few-shot samples increases, the generated images exhibit greater realism, coherence, and alignment with class-specific characteristics.

pairs without associated textual descriptions. Consequently, CLIP is fine-tuned using classification supervision on the image encoder only, while the text encoder remains frozen. In contrast, the synthetic-data setting includes class-specific textual prompts as an inherent component of the generation process, allowing full multimodal fine-tuning of both image and text encoders. This distinction reflects realistic clinical data constraints and highlights an additional benefit of synthetic data in enabling prompt-based multimodal learning.

For all synthetic data generation experiments, no additional explicit regularization is applied beyond the intrinsic low-rank constraint imposed by LoRA, which acts as an effective regularizer in few-shot settings. Empirically, this configuration was sufficient to ensure stable training and visually consistent synthetic image generation across all few-shot scenarios.

All diffusion fine-tuning and synthetic image generation experiments are conducted on a workstation equipped with an NVIDIA A100 GPU (40 GB VRAM), 64 GB RAM, and an Intel Xeon CPU. LoRA fine-tuning of Stable Diffusion v2.1 required approximately 25–40 minutes per class for the 50-shot setting, with substantially lower time for fewer shots. Once trained, synthetic image generation required approximately 0.8–1.2 seconds per image at $1024 \times 1024$ resolution, enabling the generation of several thousand images within a few hours. Importantly, this computational cost is incurred only once during dataset preparation and does not impact downstream model training or inference. The use of parameter-efficient LoRA adaptation ensures scalability and makes the proposed approach practical for deployment in clinical research environments with modern GPU infrastructure.

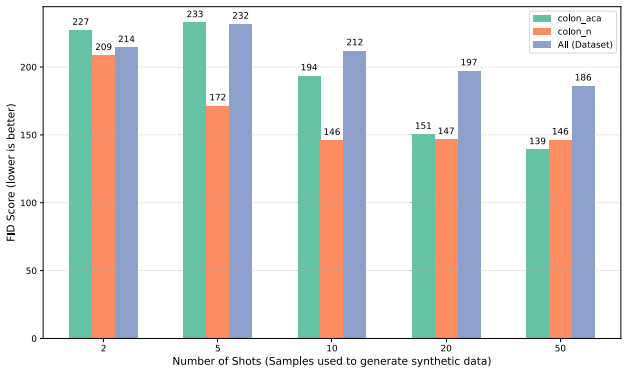



**FIGURE 3.** FID scores for synthetic images of the colon dataset generated under different few-shot settings. The x-axis represents the number of few-shot samples used to guide image generation, while the y-axis indicates the corresponding FID score. Each group of bars shows the FID scores for the colon_aca, colon_n, and all (i.e., both colon_aca and colon_n) classes, illustrating how image quality varies with the number of shots and across different categories.

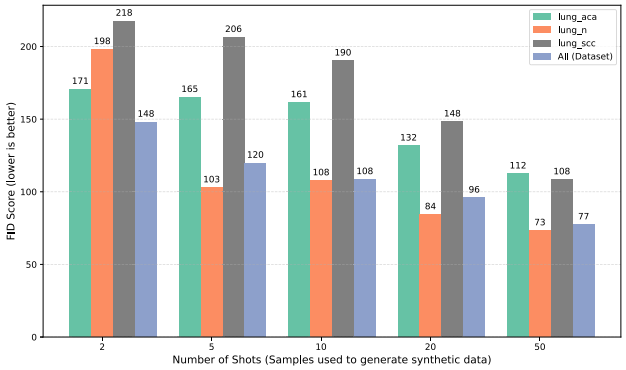



**FIGURE 4.** FID scores for synthetic images of the lung dataset generated under different few-shot settings. The x-axis represents the number of few-shot samples used to guide image generation, while the y-axis indicates the corresponding FID score. Each group of bars shows the FID scores for the lung_aca, lung_n, lung_scc, and all (i.e., including lung_aca, lung_n, and lung_scc) classes, illustrating how image quality varies with the number of shots and across different categories.

### C. EVALUATION METRICS

To assess the quality of the generated synthetic images, we use the Fréchet Inception Distance (FID), a widely used metric that quantifies the similarity between the distributions of real and synthetic images. This is achieved by comparing the statistical features extracted from both image sets using a pre-trained Inception network [21]. A lower FID score signifies greater similarity to real images, indicating higher visual fidelity of the synthetic images. FID is defined as:

$$\mathrm{FID} = \|\mu_r - \mu_s\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_s - 2\left(\Sigma_r \Sigma_s\right)^{1/2}\right), \quad (1)$$

where $\mu$ and $\Sigma$ denote the mean and covariance of feature embeddings of real ($r$) and synthetic ($s$) images.

We employ FID to quantify the similarity between real and synthetic image distributions. While FID is widely adopted in generative modeling, it relies on feature embeddings extracted from Inception v3, which is trained on natural images and may not optimally capture fine-grained histopathological textures. As previously noted in the literature [21], this can introduce bias when evaluating medical images with complex tissue-specific morphology. In this work, FID is therefore interpreted as a relative metric for comparing generative quality across methods and few-shot settings under identical evaluation conditions, rather than as a definitive measure of clinical realism. To mitigate this limitation, we complement FID with downstream classification performance on real images and qualitative visual assessment by domain-relevant criteria (Fig. 2), providing a more holistic evaluation of synthetic data utility.

Mode collapse refers to the failure of a generative model to capture the full diversity of the data distribution, resulting in highly similar or repetitive outputs. Although diffusion models are generally more robust to mode collapse than adversarial approaches, few-shot fine-tuning may still reduce diversity if not carefully constrained. In this work, the risk of mode collapse is mitigated through LoRA-based parameter-efficient fine-tuning, which limits adaptation to low-rank updates while preserving the pre-trained generative capacity of the diffusion model. Furthermore, stochastic sampling during the denoising process and class-specific prompt variation encourage diversity in the generated samples. To assess diversity indirectly, we analyze FID scores across increasing few-shot settings and evaluate downstream classification performance on real test data. The observed decrease in FID and consistent performance gains across architectures suggest that the synthetic datasets retain sufficient intra-class variability and do not collapse to a small number of modes.

While FID provides a quantitative measure of distributional similarity, it does not fully capture fine-grained histopathological structures critical for clinical interpretation. Therefore, we complement FID with a qualitative visual assessment of the generated images based on domain-relevant morphological criteria, including glandular architecture, nuclear morphology, tissue organization, and staining consistency. Representative examples of real and synthetic images under varying few-shot settings are shown in Fig. 2, where increasing visual realism and class-specific features can be observed as the number of guiding samples increases. This qualitative analysis serves as an essential complementary validation of synthetic image fidelity in the medical imaging context. In addition to FID, we compute LPIPS to quantify perceptual diversity among synthetic samples and use PRD analysis to jointly evaluate fidelity and coverage of the generated distribution.

Although synthetic data are commonly regarded as privacy-preserving, this work does not provide formal guarantees against privacy risks such as memorization, membership inference, or model inversion attacks. The generative model is fine-tuned using LoRA on a limited number of samples, which reduces but does not eliminate the possibility of reproducing training examples. Therefore, FID is used strictly as a distribution-level similarity measure

and not as an indicator of privacy or sample uniqueness. A rigorous privacy risk assessment is left for future work.

To evaluate the performance of the classification models, several standard metrics are utilized, each capturing distinct aspects of predictive capability. Overall accuracy, a commonly used metric, is calculated as the ratio of correctly predicted instances to the total number of predictions, providing a general indication of the model's correctness.

In our experiments, precision and recall are used as key performance metrics to enable a more comprehensive assessment of the model. Precision quantifies the proportion of true positive predictions among all instances classified as positive, reflecting the model's ability to limit false positives, which is an important consideration in our approach, where incorrect positive predictions can produce misleading results. In contrast, recall measures the model's ability to correctly detect all actual positive cases, indicating its sensitivity and effectiveness in minimizing false negatives. Together, these metrics provide deeper insight into the trade-offs between false positives and false negatives, offering a more robust evaluation of the model's performance than overall accuracy alone.

In our classification task, we also use the F1-score and the Area Under the Receiver Operating Characteristic Curve (AUROC) to obtain a more nuanced evaluation of the model's performance. The F1-score, defined as the harmonic mean of precision and recall, is particularly valuable in cases where there is an imbalance between classes or where both false positives and false negatives carry significant consequences. It provides a balanced measure that accounts for both types of errors, making it well-suited for assessing model effectiveness in real-world data conditions where accuracy alone may be misleading. The AUROC, on the other hand, evaluates the model's ability to discriminate between the positive and negative classes across all possible classification thresholds. It reflects the trade-off between the true positive rate and the false positive rate, offering a threshold-independent measure of separability. A higher AUROC indicates that the model is capable of distinguishing between classes more effectively. By incorporating both the F1-score and AUROC, we ensure a comprehensive and reliable assessment of model performance that captures both predictive accuracy and class discrimination ability. If $TP$, $TN$, $FP$, and $FN$ are true positive, true negative, false positive, and false negative, respectively, then we define precision, recall, accuracy, and F1-score as follows.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

The formula for the AUC using the trapezoidal rule for a set of points $(x_i, y_i)$ is:

$$\text{AUC} = \sum_i \frac{1}{2} (x_{i+1} - x_i) (y_{i+1} + y_i) \tag{6}$$

### D. EXPERIMENTAL RESULTS

We divide this section into two parts to clearly present our experimental results. The first part focuses on evaluating the quality of the synthetic data generation process, assessing how well the generated data replicate the characteristics of the real dataset. The second part examines the performance of various classification models trained on real and synthetic datasets, highlighting the effectiveness and generalizability of synthetic data. Each part is described in detail in the following subsections.

#### 1) EVALUATION OF SYNTHETIC DATA GENERATION

We first assess the quality of the synthetic images generated using the Stable Diffusion v2.1 model integrated with LoRA attention, guided by few-shot samples, and present the qualitative comparison in Fig. 2, which shows examples of synthetic images for colon and lung tissues under different few-shot settings. The figure illustrates how the number of few-shot examples influences the visual fidelity and variability of the generated images. As the number of guiding samples increases, the synthetic images tend to exhibit improved texture, structure, and class-specific characteristics. Complementing this qualitative assessment, Figs. 3 and 4 present bar graphs of the corresponding FID scores for each class of colon and lung datasets across the same few-shot configurations. For clarity, FID results are reported graphically in Figs. 3 and 4 rather than in tabular form, as these plots more effectively illustrate trends across different few-shot configurations and classes. The FID results show a clear trend in which given classes, and the whole dataset (i.e., all classes of a dataset) relatively achieve comparatively lower scores with increasing few-shot samples, indicating better alignment with the distribution of real data. These findings suggest that incorporating diverse examples from all classes of a given dataset during generation enhances the overall quality and generalizability of the synthetic data.

To further quantify intra-class diversity and distribution coverage of the generated samples, we compute LPIPS and Precision–Recall Distributions (PRD) between synthetic and real image sets. The results are summarized in Table 4. As the number of guiding few-shot samples increases, LPIPS scores consistently rise, indicating greater perceptual variability among generated samples. Simultaneously, PRD precision improves, reflecting higher fidelity to real data, while PRD recall increases, demonstrating enhanced coverage of the real data distribution. These findings confirm that the proposed LoRA-based diffusion adaptation does not suffer from mode collapse and maintains both diversity and realism across few-shot configurations.

**TABLE 3.** Comparison of model performance when trained on real data versus synthetic data, evaluated on real test sets. Left half of the table summarizes the evaluation metrics Accuracy, Precision, Recall, F1-score, and AUROC computed for the real dataset, while the same metrics are calculated for the synthetic dataset on the right half of the table. We use different convolutional models such as ResNet-50, DenseNet-121, EfficientNet-B0, Swin-B, MobileNetV2, ConvNet, ResNet, and also non-convolutional models such as MaxViT, DeiT-L, DeiT-B, and VOLO-d1 to asses their performance on both datasets.

| Dataset | Model | Real train set / Real test set (Real/Real) | | | | | Synthetic train set / Real test set (Real/Syn) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1-score | AUROC | Accuracy | Precision | Recall | F1-score | AUROC |
| Colon | ResNet-50 [22] | 0.9933 | 0.9933 | 0.9933 | 0.9933 | 0.9998 | 0.9000 | 0.9144 | 0.9000 | 0.8991 | 0.9743 |
| | DenseNet-121 [23] | 0.9833 | 1.0 | 0.9667 | 0.9831 | 0.9998 | 0.9167 | 0.9189 | 0.9167 | 0.9166 | 0.9842 |
| | EfficientNet-B0 [24] | 0.9933 | 1.0 | 0.9867 | 0.9933 | 0.9997 | 0.8433 | 0.8489 | 0.8433 | 0.8427 | 0.9424 |
| | Swin-B [25] | 0.9833 | 0.9866 | 0.9800 | 0.9833 | 0.9996 | 0.8367 | 0.8733 | 0.8367 | 0.8326 | 0.9772 |
| | MobileNetV2 [26] | 0.9900 | 0.9868 | 0.9933 | 0.9900 | 0.9995 | 0.8867 | 0.8901 | 0.8867 | 0.8864 | 0.9569 |
| | ConvNet [27] | 0.8267 | 0.7784 | 0.9133 | 0.8405 | 0.8949 | 0.7633 | 0.7090 | 0.8933 | 0.7906 | 0.7874 |
| | ResNet-101 [22] | 0.9933 | 1.0 | 0.9867 | 0.9933 | 0.9999 | 0.8600 | 0.8906 | 0.8600 | 0.8572 | 0.9684 |
| | MaxViT [28] | 0.9933 | 0.9868 | 1.0 | 0.9934 | 0.9997 | 0.9067 | 0.9174 | 0.9067 | 0.9061 | 0.9848 |
| | DeiT-L [29] | 0.9967 | 0.9934 | 1.0 | 0.9967 | 0.9997 | 0.9167 | 0.9250 | 0.9167 | 0.9163 | 0.9885 |
| | DeiT-B [29] | 0.9900 | 0.9868 | 0.9933 | 0.9900 | 0.9994 | 0.8500 | 0.8783 | 0.8500 | 0.8471 | 0.9816 |
| | VOLO-d1 [30] | 0.9967 | 1.0 | 0.9933 | 0.9967 | 1.0 | 0.8267 | 0.8713 | 0.8267 | 0.8213 | 0.9806 |
| | CLIP (16) [31] | 0.9633 | 0.9793 | 0.9467 | 0.9627 | 0.9964 | 0.6933 | 0.7154 | 0.6933 | 0.6853 | 0.7908 |
| | CLIP (32) [31] | 0.6667 | 0.6389 | 0.7667 | 0.6970 | 0.7162 | 0.7833 | 0.8087 | 0.7833 | 0.7788 | 0.8852 |
| | CLIP (16) + LoRA | - | - | - | - | - | 0.7667 | 0.8307 | 0.7667 | 0.7548 | 0.9502 |
| | CLIP (32) + LoRA | - | - | - | - | - | 0.7467 | 0.8153 | 0.7467 | 0.7321 | 0.9483 |
| Lung | ResNet-50 [22] | 0.9867 | 0.9869 | 0.9867 | 0.9867 | 0.9996 | 0.8600 | 0.8642 | 0.8600 | 0.8588 | 0.9555 |
| | DenseNet-121 [23] | 0.9689 | 0.9689 | 0.9689 | 0.9689 | 0.9972 | 0.8689 | 0.8677 | 0.8689 | 0.8654 | 0.9704 |
| | EfficientNet-B0 [24] | 0.9667 | 0.9667 | 0.9667 | 0.9666 | 0.9977 | 0.8511 | 0.8509 | 0.8511 | 0.8502 | 0.9541 |
| | Swin-B [25] | 0.9711 | 0.9722 | 0.9711 | 0.9711 | 0.9987 | 0.8844 | 0.8845 | 0.8844 | 0.8844 | 0.9721 |
| | MobileNetV2 [26] | 0.9667 | 0.9666 | 0.9667 | 0.9666 | 0.9976 | 0.8178 | 0.8446 | 0.8178 | 0.8092 | 0.9532 |
| | ConvNet [27] | 0.8978 | 0.9027 | 0.8978 | 0.8985 | 0.9768 | 0.7622 | 0.7890 | 0.7622 | 0.7586 | 0.9047 |
| | ResNet-101 [22] | 0.9644 | 0.9647 | 0.9644 | 0.9644 | 0.9979 | 0.8711 | 0.8740 | 0.8711 | 0.8714 | 0.9630 |
| | MaxViT [28] | 0.9822 | 0.9822 | 0.9822 | 0.9822 | 0.9995 | 0.8911 | 0.8916 | 0.8911 | 0.8913 | 0.9750 |
| | DeiT-L [29] | 0.9844 | 0.9845 | 0.9844 | 0.9844 | 0.9993 | 0.8667 | 0.8688 | 0.8667 | 0.8674 | 0.9589 |
| | DeiT-B [29] | 0.9689 | 0.9689 | 0.9689 | 0.9689 | 0.9978 | 0.8933 | 0.8924 | 0.8933 | 0.8926 | 0.9725 |
| | VOLO-d1 [30] | 0.9844 | 0.9845 | 0.9844 | 0.9844 | 0.9991 | 0.8889 | 0.8874 | 0.8889 | 0.8877 | 0.9705 |
| | CLIP (16) [31] | 0.9378 | 0.9379 | 0.9378 | 0.9376 | 0.9885 | 0.8267 | 0.8261 | 0.8267 | 0.8263 | 0.9501 |
| | CLIP (32) [31] | 0.9133 | 0.9161 | 0.9133 | 0.9136 | 0.9865 | 0.8467 | 0.8479 | 0.8467 | 0.8407 | 0.9612 |
| | CLIP (16) + LoRA | - | - | - | - | - | 0.8578 | 0.8560 | 0.8578 | 0.8551 | 0.9598 |
| | CLIP (32) + LoRA | - | - | - | - | - | 0.8733 | 0.8759 | 0.8733 | 0.8711 | 0.9699 |

**TABLE 4.** Intra-class perceptual diversity (LPIPS ↑) and distribution coverage (PRD Recall ↑) across few-shot settings. Higher LPIPS indicates greater perceptual variability. PRD precision reflects fidelity, while PRD recall reflects distribution coverage.

| Dataset | Shots | LPIPS (↑) | PRD Precision (↑) | PRD Recall (↑) |
|---|---|---|---|---|
| Colon | 5 | 0.338 | 0.84 | 0.71 |
| | 10 | 0.356 | 0.87 | 0.78 |
| | 20 | 0.372 | 0.90 | 0.83 |
| | 50 | 0.381 | 0.92 | 0.88 |
| Lung | 5 | 0.327 | 0.83 | 0.70 |
| | 10 | 0.349 | 0.86 | 0.76 |
| | 20 | 0.368 | 0.89 | 0.82 |
| | 50 | 0.379 | 0.91 | 0.87 |

Although the synthetic histopathological images generated in this work demonstrate strong visual realism and support improved classification performance, formal validation by expert medical pathologists are not conducted. The focus of this work is on evaluating the utility of synthetic data for mitigating data scarcity and enhancing machine learning model generalization, rather than on establishing clinical diagnostic validity. Image quality is therefore assessed using distributional similarity metrics, qualitative inspection guided by known histopathological features, and downstream classification performance on real, unseen test data. We acknowledge this limitation and note that incorporating blinded evaluation by expert pathologists is an important direction for future work.

### 2) EVALUATION OF CLASSIFICATION MODELS ON REAL AND SYNTHETIC DATA

We apply a variety of deep learning architectures, including convolutional neural networks and transformer-based models, to evaluate their convergence behavior and overall performance. By comparing both convolutional and non-convolutional transformer models, we aim to investigate their effectiveness in learning from real and synthetic data. This approach allows us to assess how different model structures handle the complexities of the dataset and to identify which architectures are most suitable for the task of cancer classification.

Initially, all models are fine-tuned exclusively on the real dataset, and their performance is assessed using their corresponding test set. The classification results of the models are summarized in Table 3 (i.e., left half of the table). Among the models evaluated on the real dataset, the top three models, including DeiT-L, VOLO-d1, and ResNet-50, perform better than other models, where DeiT-L and VOLO-d1 for colon dataset, ResNet-50, DeiT-L, and VOLO-d1 for lung dataset, achieved the highest accuracy scores as highlighted. Fig. 5 for colon, while 6 and 7 for lung dataset, illustrate the

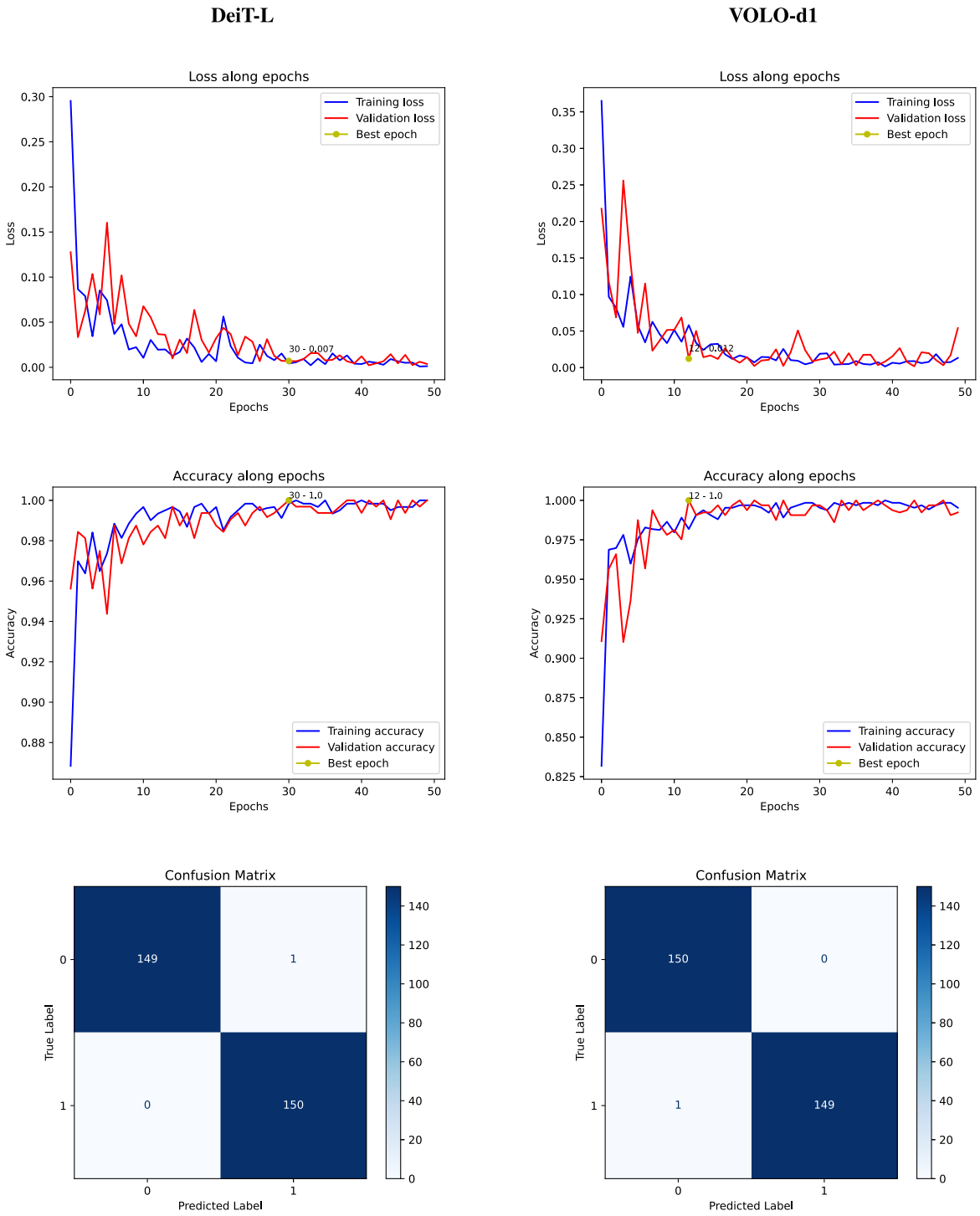



**FIGURE 5.** Performance visualization of the models; DeiT-L and VOLO-d1 selected based on their top ranking on real colon data. Training and evaluation are shown using loss, accuracy, and a confusion matrix for datasets with sufficient annotated samples.

convergence behavior of the leading models across training and evaluation phases of the real datasets of colon and lung, respectively reflecting Table 3. Specifically, we present the trajectories of accuracy and loss to capture the models' learning dynamics, alongside the corresponding confusion matrices that provide a detailed view of class-wise predictive performance. Furthermore, the CLIP model demonstrated comparatively lower accuracy, which can be attributed to the limited use of prompt engineering in this evaluation.

To further investigate the impact of the characteristics of the training data, we repeated the same set of experiments as summarized in Table 3 (right half of the table), using

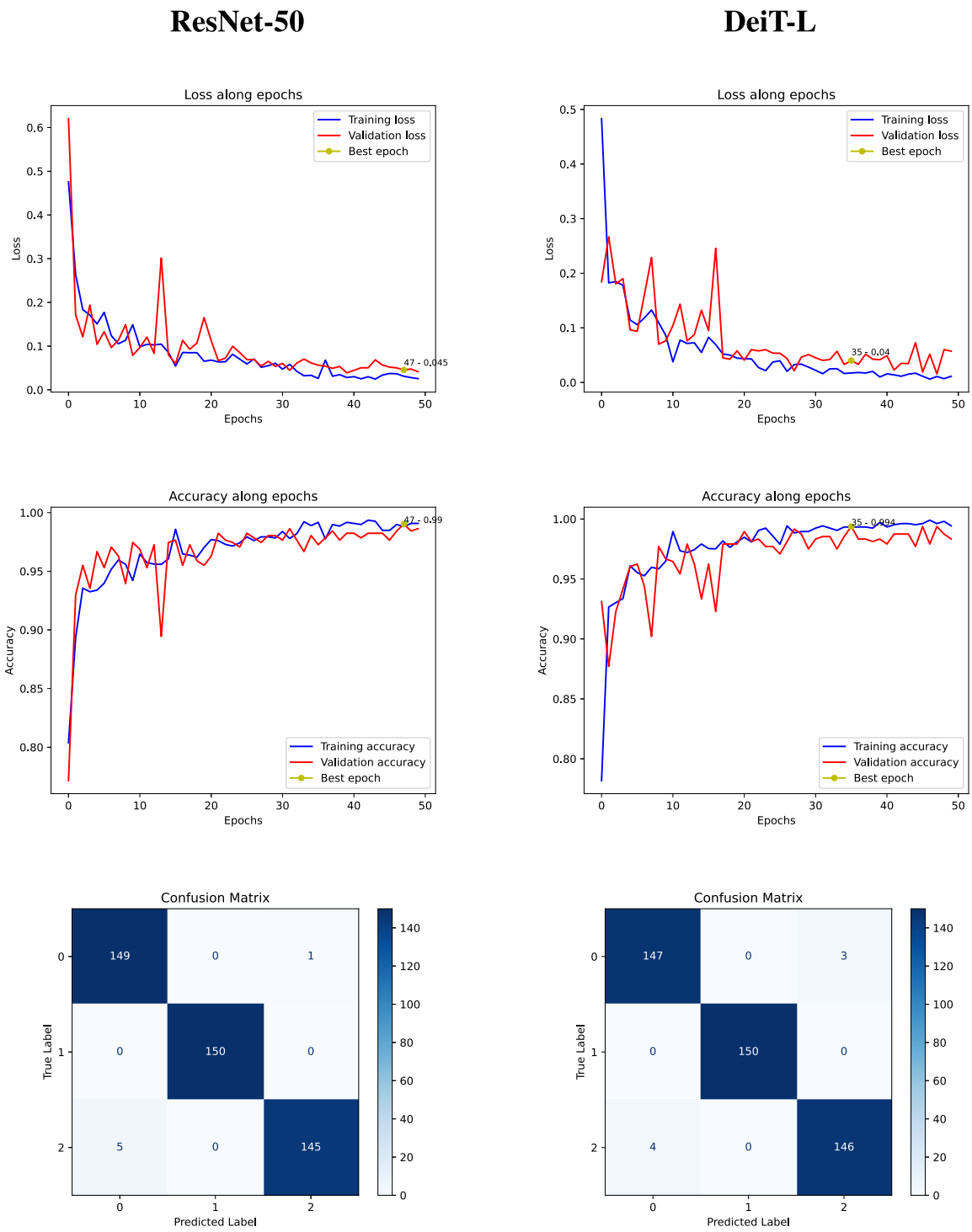
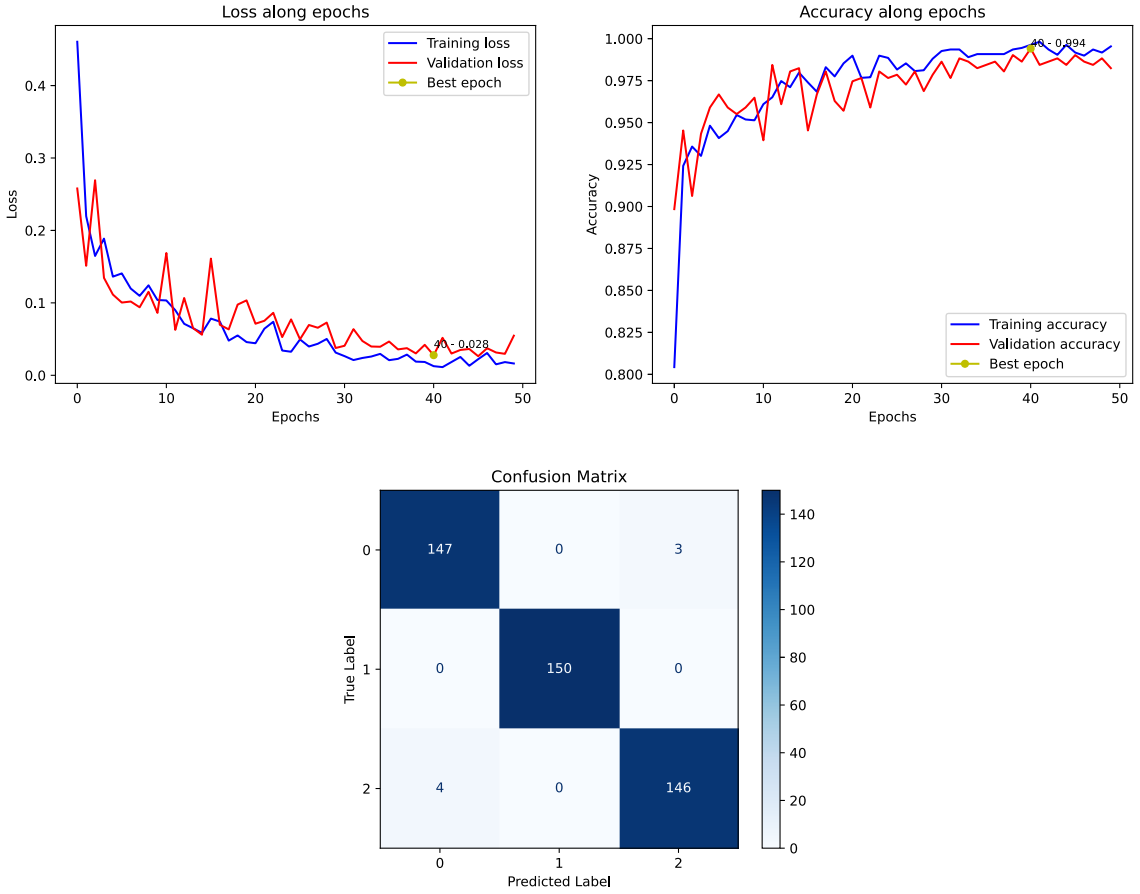



**FIGURE 6.** **Convergence analysis of the top-performing models; ResNet-50 and DeiT-L on the real lung dataset, presented through training/validation loss, accuracy, and confusion matrices under sufficient labeled data conditions.**

the synthetic dataset generated by the Stable Diffusion model guided by 50 shots. All models, including both convolutional architectures and non-convolutional models, are trained on the synthetic data and subsequently evaluated on the real test set. The classification results, presented in Table 3, demonstrate a remarkable performance across most models. This suggests that the synthetic dataset provides complementary features or representations of the target classes, leading to generalization on the synthetic dataset. Notably, even CLIP shows improved accuracy under this




(a) VOLO-d1

**FIGURE 7.** Convergence analysis of the top-performing model VOLO-d1 on the real lung dataset, presented through training/validation loss, accuracy, and confusion matrices, under conditions with sufficient labeled data.

setup, indicating that a synthetic dataset may help mitigate limitations arising from prompt dependency.

Based on the performance results for the synthetic dataset presented in Table 3, the top-performing models are selected for further evaluation and convergence analysis on synthetic datasets. Their convergence behavior is analyzed in terms of training and validation loss, accuracy, and confusion matrix across epochs. The results of this analysis are summarized in Figs. 8 and 9 for synthetic colon and lung datasets, respectively, which clearly indicate that deep learning models benefit significantly from the availability of a large synthetic dataset, showing competitive performance compared to that on real data.

Furthermore, we increase the amount of synthetic data to analyse the performance of the leading models from Table 3. We generate synthetic data up to 5000 samples per class, thus, a total of 10, 000 images for colon dataset and 15000 images for lung dataset. We use 4000 samples for training and 1000 for validation of the model, where test set remains the same as in previous experiments taken from the real dataset. The results of the top performing models evaluated on increased synthetic datasets are summarized in Table 5 showing improved performance very close to the performance of the models trained on the real data. We also fine-tune the CLIP model, leveraging its joint image and text encoding capabilities to enhance both generalizability and classification performance. Unlike earlier configurations where prompt engineering was limited, this approach fully utilizes CLIP's ability to integrate semantic information from class descriptions using comprehensive prompts. The performance of the fine-tuned CLIP model on the increased dataset, together with the top-performing models is presented in Table 5, where top results of the model for each dataset are highlighted.

We perform experiments by incorporating the synthetic training set with real few-shot samples (5, 10, 20, 50), and the results are reported in Table 6. While DenseNet-121 and CLIP (16) show minimal gains, DeiT-L achieves a clear improvement on the colon dataset, and CLIP (32) exhibits significant advances on the lung dataset. These results demonstrate that real samples provide essential domain-specific cues and fine-grained textural information that synthetic data alone may not capture. To assess the effectiveness of this strategy, the top-performing results

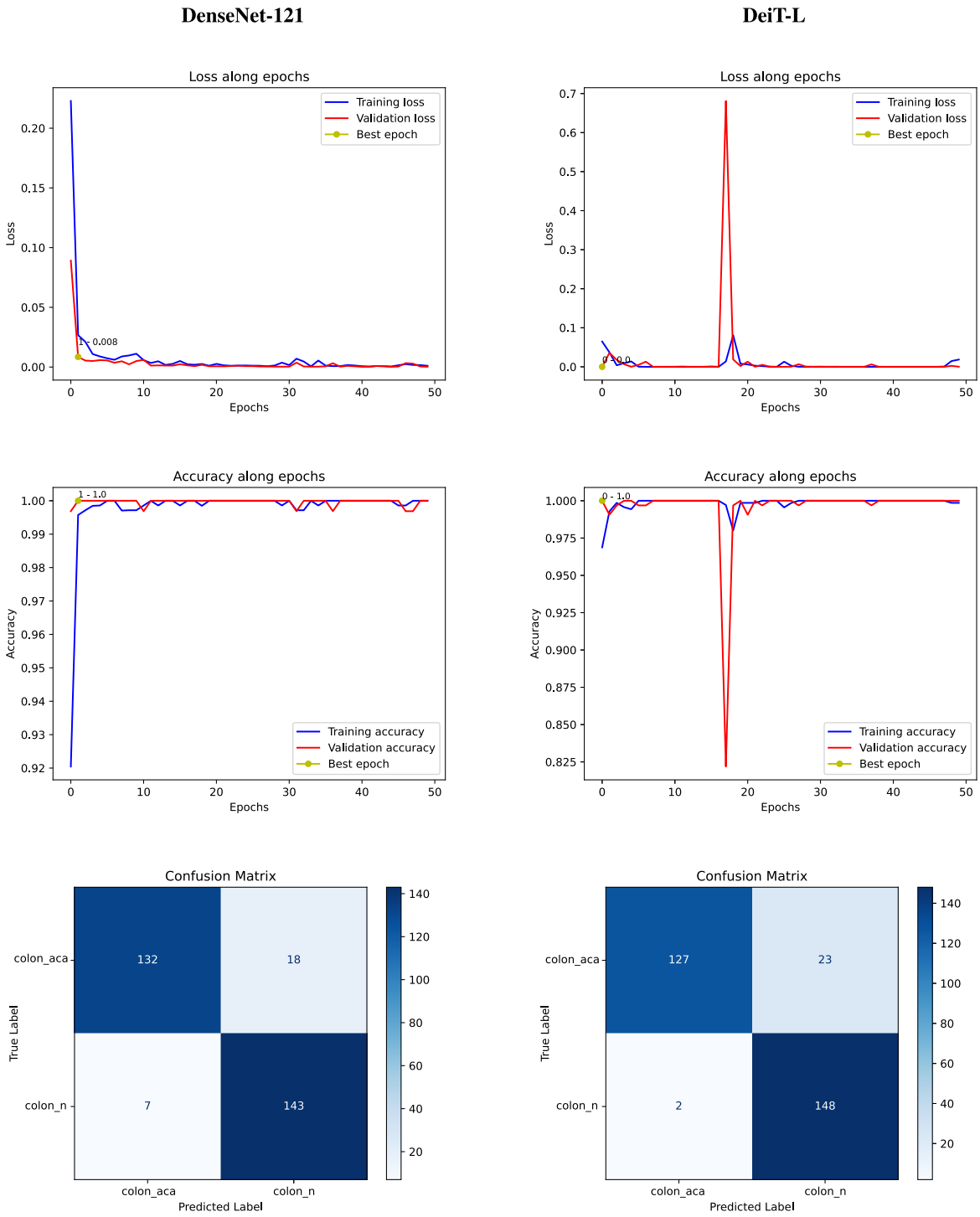



**FIGURE 8.** Convergence analysis of the models selected based on their top performance on the synthetic colon dataset, visualized using loss, accuracy, and confusion matrix.

are compared against the performance of the same models trained and evaluated on real few-shot samples alone within the same Table 6. Notably, the combined use of synthetic and real data consistently outperforms training with only a real few-shot samples, establishing synthetic data as a powerful complement for bridging the distributional gap and improving model robustness. These findings confirm that incorporating a small number of real samples into synthetic data is an effective strategy for enhancing classification performance.

Importantly, the hybrid "few real + synthetic" configuration consistently outperforms few-shot real-only training

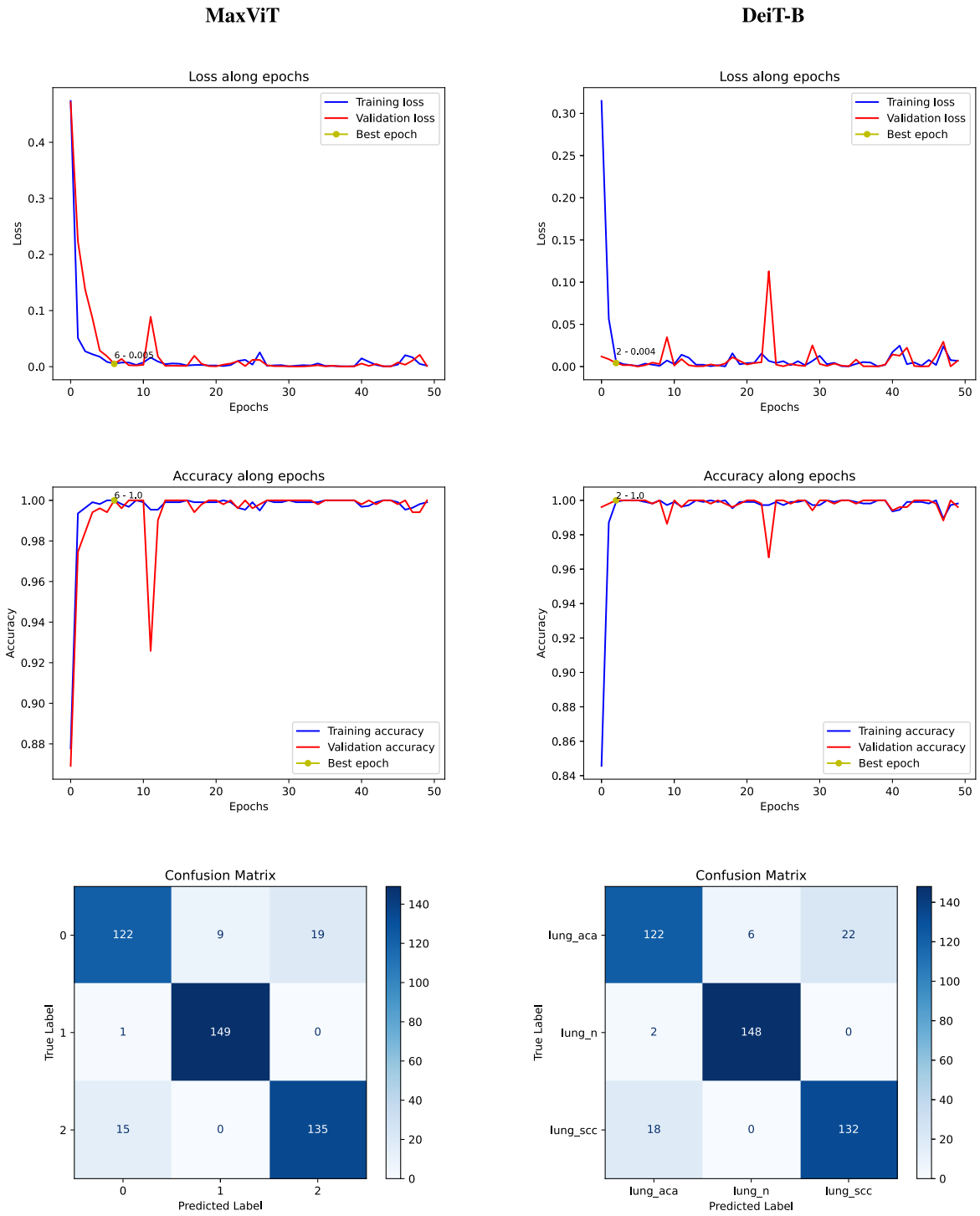



**FIGURE 9.** **Convergence analysis of the models selected based on their top ranking on the synthetic lung dataset, visualized using loss, accuracy, and confusion matrix.**

across both datasets (see Table 6). Analysis of the corresponding confusion matrices indicates a reduction in false negatives for malignant classes, suggesting improved sensitivity under limited-data conditions. Furthermore, gains in AUROC demonstrate enhanced class separability, while improvements in F1-score reflect a more balanced trade-off between precision and recall. These findings indicate that incorporating a small number of real samples provides essential domain-specific cues that complement the variability introduced by synthetic data, resulting in improved robustness.

While increasing the synthetic dataset to 5,000 samples per class led to improved performance, this work does not claim indefinite scalability. Performance saturation or

**TABLE 5.** Classification performance of the relatively superior models evaluated on increased synthetic samples (i.e., 4000 samples per class) of colon and lung training datasets, is presented. The performance of CLIP is further enhanced by applying LoRA-based attention fine-tuning to both the text (i.e., prompt) and image modalities.

| Dataset | Model | Accuracy | Precision | Recall | F1-score | AUROC |
|---|---|---|---|---|---|---|
| colon | DenseNet-121 [23] | 0.9367 | 0.9456 | 0.9267 | 0.9360 | 0.9826 |
| | DeiT-L [29] | 0.9500 | 0.9655 | 0.9333 | 0.9492 | 0.9895 |
| | CLIP (16) + LoRA | 0.8600 | 0.7872 | 0.9867 | 0.8757 | 0.9770 |
| | CLIP (32) + LoRA | 0.8633 | 0.8011 | 0.9667 | 0.8761 | 0.9697 |
| lung | MaxViT [28] | 0.9044 | 0.9046 | 0.9044 | 0.9026 | 0.9807 |
| | DeiT-B [29] | 0.8933 | 0.8924 | 0.8933 | 0.8915 | 0.9772 |
| | CLIP (16) + LoRA | 0.9067 | 0.9080 | 0.9067 | 0.9065 | 0.9793 |
| | CLIP (32) + LoRA | 0.8800 | 0.8779 | 0.8800 | 0.8781 | 0.9732 |

degradation may occur at larger scales due to reduced sample diversity, noise accumulation, or mode repetition in the generative process. We did not observe such effects within the evaluated range; however, systematic analysis beyond this scale is left for future work. Synthetic labels are assigned directly from generation prompts and are therefore consistent by design, although visual ambiguity may still arise as dataset size increases.

### E. ABLATION STUDY

To assess the individual contribution of each component in our proposed approach, we performed a detailed ablation study. Through this analysis, we aim to highlight the importance of each design choice and provide a clearer understanding of the factors driving the model's effectiveness.

#### 1) IMPACT OF FEW-SHOTS ON SYNTHETIC DATA

In real-world datasets, it is common to encounter images that are visually similar but captured from different angles or with variations in the posture of the object. This inherent similarity underscores the critical role of few-shot samples in guiding the synthetic data generation process. When a limited number of reference samples (e.g., two-shot scenarios) are used, there is a higher likelihood that the generated data will lack variability, primarily due to the homogeneity of the guiding examples. Such limited diversity in the reference images can lead to synthetic outputs that do not adequately capture the full range of visual variations present in natural settings. Conversely, incorporating a larger set of diverse few-shot examples significantly enhances the diversity of the generated data (see Fig. 2). This approach facilitates the creation of synthetic images that are more representative of different viewpoints, postures, and other relevant visual attributes; thus, enriching the dataset and improving its utility for downstream tasks such as model training and evaluation.

Moreover, the FID metric is also affected by the limited diversity of synthetic datasets generated from few-shot samples. When the guiding few-shot samples lack sufficient variations, often due to their visual similarity, the resulting synthetic data exhibit restricted diversity. This discrepancy becomes evident when comparing the synthetic dataset to the real dataset, which typically contains a broader range of unseen and diverse examples. As a result, the FID score may be relatively high, indicating poor alignment between the distributions of real and generated data. However, as the number of few-shot samples increases and incorporates greater variability (see Fig. 3 and 4), the FID score decreases. This reduction signifies that the synthetic dataset better approximates the distribution of the real data, thus supporting the improved generalizability and performance of downstream classification models.

#### 2) IMPACT OF SYNTHETIC DATA ON MODEL PERFORMANCE

In scenarios where a real dataset is limited, relying solely on real samples for training classification models often results in suboptimal performance and generalizability due to insufficient coverage of the data distribution. Conversely, when a sufficiently large and diverse synthetic dataset is generated, particularly through a well-guided few-shot synthesis process, it can significantly enhance model training. Our experiments demonstrate that synthetic data not only compensate for the scarcity of real data but also enable effective fine-tuning of classification models. As illustrated in Table 6, models fine-tuned with synthetic data show competitive performance compared to real data. This suggests that high-quality, diverse synthetic datasets can serve as a valuable alternative in low-resource settings, effectively supporting model generalizability and performance when real data availability is constrained. To boost classification accuracy, we fine-tune the CLIP model by leveraging its capability for prompt-based text encoding, which enables more effective integration of semantic context during training. This enhancement allows the model to better capture complex relationships between visual features and class labels. The final quantitative comparisons between the top-performing models and the fine-tuned CLIP model are presented in Table 6. The results clearly demonstrate that the CLIP-based approach improves classification performance, highlighting the benefit of incorporating multimodal knowledge, particularly when the synthetic dataset is generated from limited examples.

To evaluate the effect of the number of few-shot samples used to guide the synthetic data generation, we perform a

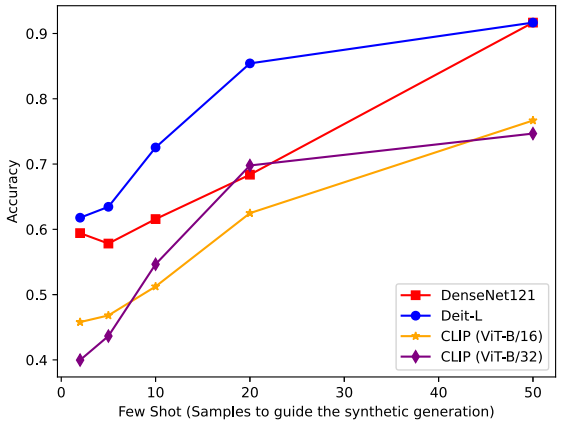
**TABLE 6.** Classification performance of the relatively top-performing models evaluated on real few-shots (*X*) alone (left half of the table), as well as increased synthetic samples and a limited number of real instances (i.e., 4000 synthetic samples +*X* real samples) from each class of the datasets used for training (right half of the table), where *X* represents 5, 10, 20 and 50 as given in the second column.

| Dataset | Shots | Model | Real train / Real test | | | | | (Real Shots + Synthetic) train / Real test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | Precision | Recall | F1-score | AUROC | Accuracy | Precision | Recall | F1-score | AUROC |
| Colon | 5 | DenseNet-121 [23] | 0.7425 | 0.8078 | 0.6812 | 0.7203 | 0.7464 | 0.9400 | 0.9714 | 0.9067 | 0.9379 | 0.9788 |
| | | DeiT-L [29] | 0.8342 | 0.8312 | 0.9357 | 0.9028 | 0.9436 | 0.9600 | 0.9423 | 0.9800 | 0.9608 | 0.9955 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.8533 | 0.7944 | 0.9533 | 0.8667 | 0.9714 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.9267 | 0.9267 | 0.9267 | 0.9267 | 0.9742 |
| | 10 | DenseNet-121 [23] | 0.7800 | 0.8524 | 0.7000 | 0.7696 | 0.7705 | 0.9560 | 0.9654 | 0.9648 | 0.9593 | 0.9793 |
| | | DeiT-L [29] | 0.8633 | 0.8728 | 0.9533 | 0.9630 | 0.9863 | 0.9659 | 0.9509 | 0.9813 | 0.9621 | 0.9962 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.8765 | 0.8555 | 0.8871 | 0.8822 | 0.9728 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.9270 | 0.9315 | 0.9214 | 0.9296 | 0.9768 |
| | 20 | DenseNet-121 [23] | 0.8533 | 0.8658 | 0.8400 | 0.8527 | 0.8907 | 0.9611 | 0.9689 | 0.9785 | 0.9696 | 0.9817 |
| | | DeiT-L [29] | 0.8700 | 0.8796 | 0.8600 | 0.8697 | 0.8948 | 0.9721 | 0.9577 | 0.9820 | 0.9638 | 0.9970 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.8992 | 0.9388 | 0.9156 | 0.9090 | 0.9748 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.9311 | 0.9466 | 0.9310 | 0.9351 | 0.9782 |
| | 50 | DenseNet-121 [23] | 0.8667 | 0.8861 | 0.8467 | 0.8660 | 0.8952 | 0.9700 | 0.9608 | 0.9800 | 0.9703 | 0.9956 |
| | | DeiT-L [29] | 0.8767 | 0.8799 | 0.8733 | 0.8766 | 0.8975 | 0.9767 | 0.9673 | 0.9867 | 0.9769 | 0.9980 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.9333 | 0.9779 | 0.8867 | 0.9301 | 0.9813 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.9467 | 0.9527 | 0.9400 | 0.9463 | 0.9867 |
| Lung | 5 | MaxViT [28] | 0.6678 | 0.6811 | 0.6678 | 0.6706 | 0.7935 | 0.9000 | 0.9003 | 0.9000 | 0.8994 | 0.9832 |
| | | DeiT-B [29] | 0.7256 | 0.7307 | 0.7256 | 0.7262 | 0.8231 | 0.9067 | 0.9061 | 0.9067 | 0.9057 | 0.9777 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.8933 | 0.8935 | 0.8933 | 0.8921 | 0.9714 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.8822 | 0.8860 | 0.8822 | 0.8826 | 0.9708 |
| | 10 | MaxViT [28] | 0.6875 | 0.6812 | 0.6576 | 0.6645 | 0.8026 | 0.9059 | 0.9078 | 0.9019 | 0.9013 | 0.9840 |
| | | DeiT-B [29] | 0.7586 | 0.7624 | 0.7713 | 0.7628 | 0.8390 | 0.9088 | 0.9092 | 0.9088 | 0.9085 | 0.9793 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.9123 | 0.9035 | 0.9018 | 0.9012 | 0.9747 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.9058 | 0.9080 | 0.9022 | 0.9042 | 0.9789 |
| | 20 | MaxViT [28] | 0.7389 | 0.7461 | 0.7389 | 0.7360 | 0.8399 | 0.9081 | 0.9158 | 0.9048 | 0.9057 | 0.9861 |
| | | DeiT-B [29] | 0.7967 | 0.7996 | 0.7967 | 0.7952 | 0.8607 | 0.9118 | 0.9138 | 0.9129 | 0.9118 | 0.9824 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.9188 | 0.9147 | 0.9194 | 0.9142 | 0.9809 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.9195 | 0.9209 | 0.9270 | 0.9281 | 0.9862 |
| | 50 | MaxViT [28] | 0.8256 | 0.8257 | 0.8256 | 0.8255 | 0.8648 | 0.9089 | 0.9231 | 0.9089 | 0.9083 | 0.9874 |
| | | DeiT-B [29] | 0.7989 | 0.7989 | 0.7989 | 0.7986 | 0.8570 | 0.9156 | 0.9172 | 0.9156 | 0.9152 | 0.9832 |
| | | CLIP (16) + LoRA | - | - | - | - | - | 0.9378 | 0.9406 | 0.9378 | 0.9380 | 0.9896 |
| | | CLIP (32) + LoRA | - | - | - | - | - | 0.9511 | 0.9509 | 0.9511 | 0.9509 | 0.9926 |

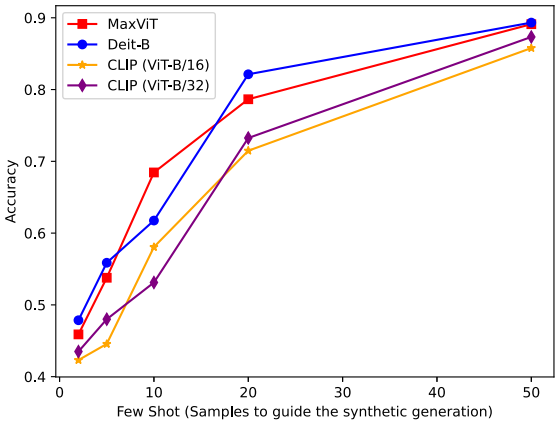
classification task using models with higher performance (see Table 6), including DenseNet-121, DeiT-L for colon dataset, where MaxViT and DeiT-B for lung dataset, and the CLIP model for both datasets. As illustrated in Fig. 10a for colon dataset and Fig. 10b for the lung dataset, all models show a clear upward trend in accuracy as the number of shots increases. This performance gain is primarily due to the increased diversity and coverage of the support set, which enhances the quality of the synthetic data generated for training. A more varied few-shot input allows the generative process to better capture the underlying data distribution, resulting in more informative synthetic samples that support improved generalization. While traditional architectures like DenseNet-121 benefit steadily from additional shots, vision transformer-based models such as MaxViT and CLIP show even greater improvements, likely due to their stronger capacity to exploit the added contextual diversity provided by larger few-shot sets and sufficiently available synthetic data for fine-tuning.

The ablation study presented in Fig. 11a and Fig. 11b demonstrates a clear upward trend in classification accuracy for both datasets across all evaluated models as the number of synthetic training samples increases from 350 to 4000 per class. This performance gain reinforces the effectiveness of using additional synthetic data to improve generalization in few-shot learning settings. In particular, models based on pre-trained vision transformers, particularly DeiT-L, DeiT-B, and CLIP, consistently outperform convolution-based architectures. MaxViT also shows strong improvements with scale, highlighting the value of transformer-based representations. Although minor fluctuations are observed, probably due to sample diversity and randomness in synthetic generation, the overall pattern confirms that augmenting the training set with more synthetic examples leads to more robust and accurate models. This underscores the critical role of both model architecture and data scale in low-resource medical image classification.

Both Figs. 10 and 11 highlight the positive correlation between data availability (i.e., real or synthetic) and improved model performance, emphasizing the importance of guided generation and data scale in rare disease classification.

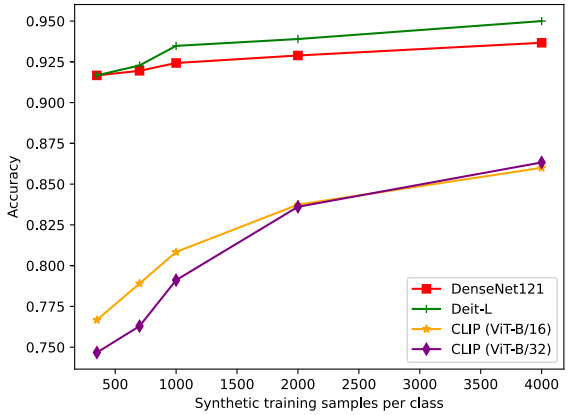



(a) Accuracy across models trained on the synthetic colon dataset with increasing few-shot real samples used to guide synthetic data generation.

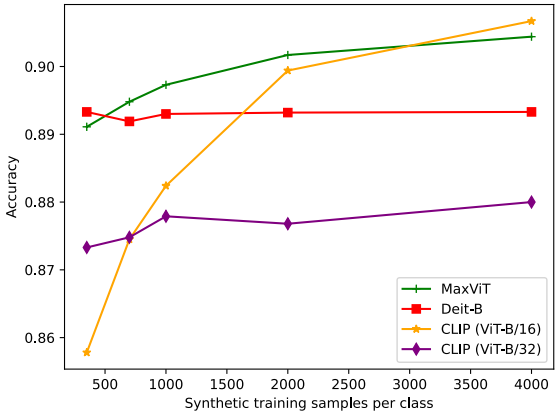



(b) Accuracy across models trained on the synthetic lung dataset with increasing few-shot real samples used to guide synthetic data generation.

**FIGURE 10.** Model performance comparison under varying data conditions. (a) shows the classification accuracy of different models trained on the synthetic colon dataset as the number of few-shot real samples used to guide synthetic generation increases. (b) shows the classification accuracy of different models trained on the synthetic lung dataset as the number of few-shot real samples used to guide synthetic generation increases.




(a) Accuracy across models evaluated on the colon dataset with increasing synthetic training samples.




(b) Accuracy across models evaluated on lung dataset with increasing synthetic training samples.

**FIGURE 11.** Model performance comparison under varying data conditions. (a) presents accuracy trends as the quantity of synthetic colon training data increases. (b) presents accuracy trends as the quantity of synthetic lung training data increases.

**TABLE 7.** Cross-dataset external validation on CRC [32].

| Training Data | Model | Accuracy | Precision | Recall | F1-score | AUROC |
|---|---|---|---|---|---|---|
| Real (LC25000) | DenseNet-121 | 0.9212 | 0.9203 | 0.9374 | 0.9189 | 0.9385 |
| | DeiT-L | 0.9238 | 0.9255 | 0.9311 | 0.9124 | 0.9195 |
| | CLIP (16) + LoRA | 0.9012 | 0.8978 | 0.8905 | 0.8868 | 0.8942 |
| | CLIP (32) + LoRA | 0.9145 | 0.9014 | 0.9088 | 0.9072 | 0.9045 |
| Synthetic (LC25000) | DenseNet-121 | 0.8415 | 0.8325 | 0.8368 | 0.8343 | 0.8422 |
| | DeiT-L | 0.8455 | 0.8398 | 0.8371 | 0.8289 | 0.8417 |
| | CLIP (16) + LoRA | 0.8725 | 0.8681 | 0.8622 | 0.8712 | 0.8808 |
| | CLIP (32) + LoRA | 0.8818 | 0.8852 | 0.8748 | 0.8787 | 0.8804 |

### 3) EXTERNAL VALIDATION

To further assess cross-dataset generalizability, we evaluate the trained colon classification models on an independent external colorectal histopathology dataset CRC [32], which contains H&E-stained image patches acquired from a different patient cohort and scanning pipeline than LC25000. No retraining or fine-tuning is performed on the external dataset. Images are resized and normalized using the

**TABLE 8.** A summary of components used to construct the prompts used for the stable diffusion model to generate corresponding synthetic data.

| Component | Description | Required |
|---|---|---|
| Imaging modality | Histopathology slide description | ✓ |
| Tissue type / condition | Organ and pathology class | ✓ |
| Staining protocol | e.g., H&E staining | ✓ |
| Magnification level | Microscopy scale (e.g., 40×) | ✓ |
| Morphological features | Class-specific cellular or structural characteristics | ✓ |
| Tissue organization | Architecture or structural arrangement | Optional |
| Color characteristics | Nuclear/cytoplasmic staining properties | Optional |
| Image quality attributes | Resolution, focus, lighting | ✓ |
| Style specification | Scientific/medical imaging style | ✓ |
| Resolution / format | Aspect ratio, pixel resolution | Optional |

**TABLE 9.** Attributes to construct prompt generate samples of the adenocarcinoma class of the colon dataset.

| Attribute | Example |
|---|---|
| TISSUE_TYPE | human colon adenocarcinoma tissue |
| STAINING_METHOD | hematoxylin and eosin (H&E) |
| MAGNIFICATION | 40x |
| MORPHOLOGICAL_FEATURES | irregular glandular structures and hyperchromatic nuclei |
| TISSUE_ARCHITECTURE | loss of normal crypt organization |
| COLOR_CHARACTERISTICS | pink eosinophilic cytoplasm and blue-purple nuclei |
| RESOLUTION | 1024 x1024 |

same preprocessing pipeline as the LC25000 experiments. To ensure label compatibility, CRC classes are mapped to a binary tumor vs. non-tumor setting consistent with our colon classification task. This experimental design enables evaluation of true cross-dataset generalization under domain shift conditions, thereby providing additional evidence of robustness beyond single-dataset validation.

The cross-dataset evaluation results are summarized in Table 7. Despite the domain shift between LC25000 and CRC, the top-performing models retain strong discriminative performance without retraining, indicating that the learned morphological representations generalize beyond the original dataset. Notably, models trained on synthetic data exhibit competitive performance compared to those trained on real data, further supporting the utility of diffusion-generated samples in enhancing robustness under cross-institutional variability. Interestingly, the performance gap between real-trained and synthetic-trained models narrows under external domain shift, suggesting that synthetic data may implicitly introduce variability that improves generalization.

## IV. LIMITATIONS AND FUTURE DIRECTIONS

Despite the promising results, several limitations of this study should be acknowledged. First, although synthetic data significantly improved classification performance under data-scarce conditions, the generated images may not capture the full biological and morphological diversity observed in real-world clinical settings, particularly across different institutions, staining protocols, and scanning devices. Second, the evaluation of synthetic image quality relied partly on FID, which is computed using features extracted from models trained on natural images. While FID provides a useful relative comparison across experimental settings, it may not fully reflect clinically relevant histopathological characteristics. Third, although the proposed approach demonstrates strong generalization when evaluated on real test data, the work is limited to two cancer types and a single public dataset. Additional validation on multi-center datasets and other rare or underrepresented diseases is necessary to further establish clinical robustness and applicability. Finally, the use of carefully designed prompts and manual selection of high-quality few-shot samples introduces a degree of expert intervention, which may limit scalability. Future work will focus on automated prompt optimization, broader disease coverage, and clinical expert–guided evaluation of synthetic image realism.

While synthetic data generation may reduce the need to share raw patient images, the proposed method does not establish formal privacy guarantees. The diffusion model is fine-tuned using real samples without differential privacy constraints, and potential risks such as memorization or reconstruction are not empirically evaluated. Therefore, claims regarding privacy should be interpreted cautiously. Future work will incorporate formal privacy-preserving mechanisms and adversarial robustness testing to assess privacy leakage risks.

## V. CONCLUSION

In this paper, we propose a synthetic data generation framework to improve cancer disease classification in low-data medical imaging scenarios. Addressing the challenges of

limited annotated samples and high visual similarity across clinical images, our approach leverages diverse few-shot examples to guide diffusion-based synthetic image generation. We demonstrated that both the diversity and the number of few-shot samples significantly influence the variability and realism of the generated dataset. Qualitative and quantitative evaluations, including FID analysis, showed that increasing few-shot diversity leads to synthetic images that better approximate the distribution of real clinical data. Importantly, classification models trained solely on synthetic data achieved competitive performance, approaching that of models fine-tuned on real datasets. Moreover, synthetic data augmentation improved convergence behavior and generalization performance. To further enhance classification, we fine-tuned a CLIP-based multimodal framework using prompt-guided text encoding, enabling the integration of semantic disease descriptions with visual representations. The transformer-based multimodal approach outperformed competitive convolution-based baselines, highlighting the effectiveness of combining synthetic visual data with multimodal representation learning. Overall, the proposed framework offers a practical and scalable strategy for improving colon and lung cancer classification in data-constrained environments. The integration of few-shot–driven synthetic image generation with multimodal fine-tuning demonstrates strong potential for broader application across medical imaging domains where annotated data remain scarce.

Among all evaluated settings, the hybrid approach combining a few real samples with synthetic data emerges as the most practically actionable strategy. This configuration achieves consistent improvements in AUROC and F1-score while reducing class-wise confusion, particularly for malignant categories. Such a setup reflects realistic clinical constraints and provides a feasible pathway for deploying deep learning models in data-scarce environments.

## APPENDIX A GENERALIZED PROMPT DESIGN TEMPLATE FOR SYNTHETIC HISTOPATHOLOGY GENERATION

To improve reproducibility and enable practical reuse of the proposed synthetic data generation pipeline, we provide a generalized template summarizing the prompt construction strategy used in this work. The template abstracts the core components required to generate domain-specific histopathological images using diffusion models.

Each prompt consists of specific components as given in Table 8.

The general prompt can be formulated as follows.

> "Ultra-high-resolution histopathology slide of [TISSUE_TYPE/CONDITION], stained with [STAINING_METHOD], microscopic view at [MAGNIFICATION], showing [MORPHOLOGICAL_FEATURES], [TISSUE_ARCHITECTURE], realistic [COLOR_CHARACTERISTICS], scientific medical imaging style, sharp focus, pathology laboratory quality, resolution [RESOLUTION]".

Following the template described above, a colon adenocarcinoma example is given in Table 9.

The following principles support prompt construction:

- **Domain specificity:** Include medically meaningful morphological descriptors.
- **Visual fidelity:** Specify imaging conditions and quality attributes.
- **Class discrimination:** Emphasize distinguishing features between classes.
- **Consistency:** Maintain standardized imaging parameters across classes.
- **Adaptability:** Allow substitution of organ, pathology, or modality.

## REFERENCES

[1] K. S. Wang et al., "Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence," *BMC Med.*, vol. 19, no. 1, p. 76, Dec. 2021, doi: 10.1186/s12916-021-01942-5.

[2] H. Ding, Y. Feng, X. Huang, J. Xu, T. Zhang, Y. Liang, H. Wang, B. Chen, Q. Mao, W. Xia, X. Huang, L. Xu, G. Dong, and F. Jiang, "Deep learning-based classification and spatial prognosis risk score on whole-slide images of lung adenocarcinoma," *Histopathology*, vol. 83, no. 2, pp. 211–228, Aug. 2023, doi: 10.1111/his.14918.

[3] N. Shahadat, R. Lama, and A. Nguyen, "Lung and colon cancer detection using a deep AI model," *Cancers*, vol. 16, no. 22, p. 3879, Nov. 2024, doi: 10.3390/cancers16223879.

[4] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso, "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, no. 1, p. 119, Sep. 2020, doi: 10.1038/s41746-020-00323-1.

[5] Y. Sun, W. Tan, Z. Gu, R. He, S. Chen, M. Pang, and B. Yan, "A data-efficient strategy for building high-performing medical foundation models," *Nature Biomed. Eng.*, vol. 9, no. 4, pp. 1–13, Mar. 2025, doi: 10.1038/s41551-025-01365-0.

[6] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, and D. I. Fotiadis, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 2892–2910, Dec. 2024, doi: 10.1016/j.csbj.2024.07.005.

[7] A. S. Coyner, J. S. Chen, K. Chang, P. Singh, S. Ostmo, R. V. P. Chan, M. F. Chiang, J. Kalpathy-Cramer, and J. P. Campbell, "Synthetic medical images for robust, privacy-preserving training of artificial intelligence," *Ophthalmology Sci.*, vol. 2, no. 2, Jun. 2022, Art. no. 100126, doi: 10.1016/j.xops.2022.100126.

[8] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22500–22510, doi: 10.1109/CVPR52729.2023.02155.

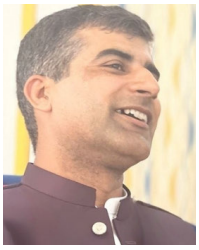
[9] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," 2023, *arXiv:2307.01952*.

[10] W. Chan, E. Denton, D. Fleet, K. Ghasemipour, R. G. Lopes, J. Ho, B. K. Ayan, L. Li, M. Norouzi, C. Saharia, T. Salimans, S. Saxena, and J. Whang, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 36479–36494.

[11] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, and A. Ramesh, "Improving image generation with better captions," *Comput. Sci.*, vol. 2, no. 3, p. 8, 2023. [Online]. Available: https://cdn.openai.com/papers/dall-e-3pdf

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685, doi: 10.1109/CVPR52688.2022.01042.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[14] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. D. Costa, V. Fernández, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *Proc. MICCAI Workshop Deep Generative Models*, 2022, pp. 117–126.
[15] K. Tajbakhsh, A. Neels, E. Fadeeva, J. C. Larsson, O. Stanowska, A. Perren, and R. Zboray, "A comprehensive study of laboratory-based micro-CT for 3D virtual histology of human FFPE tissue blocks," *IEEE Access*, vol. 12, pp. 78304–78316, 2024, doi: 10.1109/ACCESS.2024.3407733.
[16] B. Sheng, P. A. Keane, Y.-C. Tham, and T. Y. Wong, "Synthetic data boosts medical foundation models," *Nature Biomed. Eng.*, vol. 9, no. 4, pp. 443–444, Apr. 2025, doi: 10.1038/s41551-025-01375-y.
[17] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, CNNs and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023, doi: 10.1109/ACCESS.2023.3266093.
[18] J. M. Kim, J. Bader, S. Alaniz, C. Schmid, and Z. Akata, "DataDream: few-shot guided dataset generation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 252–268, doi: 10.48550/arxiv.2407.10910.
[19] J. Y. Choi, J. R. Park, I. Park, J. Cho, A. No, and E. K. Ryu, "Simple drop-in LoRA conditioning on attention layers will improve your diffusion model," 2024, *arXiv:2405.03958*.
[20] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (LC25000)," 2019, *arXiv:1912.12142*.
[21] Y. Wu, F. Liu, R. Yilmaz, H. Konermann, P. Walter, and J. Stegmaier, "A pragmatic note on evaluating generative models with Fréchet inception distance for retinal image synthesis," 2025, *arXiv:2502.17160*.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[24] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.
[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
[27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986, doi: 10.1109/CVPR52688.2022.01167.
[28] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. C. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 459–479, doi: 10.1007/978-3-031-20053-3_27.
[29] H. Touvron, C. Matthieu, M. Douze, F. Massa, A. Sablayrolles, and J. Hervé, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 10347–10357.
[30] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision outlooker for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6575–6586, May 2023, doi: 10.1109/TPAMI.2022.3206108.
[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[32] J. N. Kather, N. Halama, and A. Marx, "100,000 histological images of human colorectal cancer and healthy tissue," Nat. Center Tumor Diseases, Heidelberg, Germany, Tech. Rep., 2018, doi: 10.5281/zenodo.1214455.



**HUSSAIN AHMAD MADNI** received the B.S. degree in computer system engineering from the Islamia University of Bahawalpur, Pakistan, in 2014, the M.S. degree in computer science from COMSATS University Islamabad, Pakistan, in 2018, and the Ph.D. degree in computer science and artificial intelligence from the University of Udine, Italy, in 2024. He is currently a Postdoctoral Researcher with the Department of Mathematics, Computer Science, and Physics, University of Udine. He has contributed to the development of secure and scalable learning frameworks, including Swarm-FHE and blockchain-based swarm learning, which address gradient leakage and promote robust collaboration across heterogeneous environments. His work has been published in reputable international journals and conferences in the field of AI. His research interests include artificial intelligence, federated learning, privacy-preserving machine learning, and cybersecurity, with a particular emphasis on distributed AI systems for medical imaging and healthcare applications.

**HAFSA SHUJAT** received the B.S. degree in education with a specialization in mathematics and science and the M.S. degree in educational leadership and management from International Islamic University, Islamabad, Pakistan, in 2021 and 2024, respectively. She has contributed to research studies on the analysis of M.S. education programs in Pakistan and on the impact of digital game-based learning on primary students' academic achievement, engagement, and motivation. In addition to her academic research, she has gained professional experience in both private and public educational institutions, where she taught mathematics, science, and computer studies and provided administrative support. She is currently preparing Ph.D. applications and developing research proposals while seeking academic opportunities aligned with her long-term scholarly goals. Her academic focus centers on strengthening methodological expertise, refining theoretical perspectives, and identifying doctoral programs that support advanced research, innovation, and impactful contributions to education systems globally.

**SILVIA ZOTTIN** received the B.Sc. degree in computer science from the Ca' Foscari University of Venice, in 2018, and the M.Sc. degree (summa cum laude) in multimedia communication and information technologies from the University of Udine, in 2021, with a specialization in artificial intelligence and industrial automation. Currently, she is pursuing the Ph.D. degree in computer science and artificial intelligence. She is with the Artificial Vision and Machine Learning (AVML) Laboratory. She continues her research work with the University of Udine, initially with a Research Fellowship. Her main research interests include computer vision, anomaly detection, image segmentation, and machine learning in contexts with limited data. She has published several papers in national and international journals and conferences. She collaborates with the Artificial Intelligence for Cultural Heritage (AI4CH) Research Group of the University of Udine, particularly focusing on document layout analysis in ancient manuscripts.

**AXEL DE NARDIN** received the M.Sc. (summa cum laude) and Ph.D. degrees in computer science, in 2020 and 2024, respectively. He is currently an Assistant Professor with the Artificial Vision and Machine Learning (AVML) Laboratory, University of Udine. His main research interests include semantic segmentation, anomaly detection, and pattern recognition, with a particular focus on low data settings. Since 2021, he has published several papers in international venues, touching a wide range of application fields, such as industrial quality inspection, medical imaging, and cultural heritage. Since 2022, he has been a member of the Artificial Intelligence for Cultural Heritage (AI4CH) Research Group, which fosters interdisciplinary collaboration between humanities and computer science experts to develop novel techniques that leverage the use of AI systems for the analysis and preservation of cultural heritage.

**GIAN LUCA FORESTI** (Senior Member, IEEE) is currently a Full Professor of computer science with the Department of Mathematics, Computer Science, and Physics, University of Udine. He is the Director of the Artificial Intelligence for Cultural Heritage (AI4CH) Center and the Artificial Vision, Real-Time System (AVIRES) Laboratory and Artificial Vision and Machine Learning Laboratory (AVML). He is PI of several research projects in the field of AI (machine learning for segmentation of ancient manuscripts), cybersecurity (anomaly detection in computer networks), and computer vision (autonomous and cooperative aerial and underwater systems). He has been the Finance Chair of the 11th IEEE Conference on Image Processing (ICIP05), the General Chair of the 8th IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS11), the General Chair of the 16th International Conference on Image Analysis and Processing (ICIAP11), and the General Chair of the 22th International Conference on Image Analysis and Processing (ICIAP23). He has an H-index of more than 50 with more than 10000 citations. He is an IAPR and AAAI Fellow Member, and National Delegate for the Ministry of Defence in the NATO-STO Information System Technology (IST) Panel.