algorithms

MDPI

Article

Text-Guided Synthesis in Medical Multimedia Retrieval: A Framework for Enhanced Colonoscopy Image Classification and Segmentation

Ojonugwa Oluwafemi Ejiga Peter [1], Opeyemi Taiwo Adeniran [2], Adetokunbo MacGregor John-Otumu [3], Fahmi Khalifa [2] and Md Mahmudur Rahman [1,*]

[1] Department of Computer Science, School of Computer, Mathematical and Natural Sciences, Morgan State University, Baltimore, MD 21251, USA; ojeji1@morgan.edu
[2] Department of Electrical and Computer Engineering, Morgan State University, Baltimore, MD 21251, USA; opade7@morgan.edu (O.T.A.); fahmi.khalifa@morgan.edu (F.K.)
[3] Department of Information Technology, Federal University of Technology Owerri, Owerri 460116, Imo State, Nigeria; adetokunbo.johnotumu@futo.edu.ng
* Correspondence: md.rahman@morgan.edu

Abstract: The lack of extensive, varied, and thoroughly annotated datasets impedes the advancement of artificial intelligence (AI) for medical applications, especially colorectal cancer detection. Models trained with limited diversity often display biases, especially when utilized on disadvantaged groups. Generative models (e.g., DALL-E 2, Vector-Quantized Generative Adversarial Network (VQ-GAN)) have been used to generate images but not colonoscopy data for intelligent data augmentation. This study developed an effective method for producing synthetic colonoscopy image data, which can be used to train advanced medical diagnostic models for robust colorectal cancer detection and treatment. Text-to-image synthesis was performed using fine-tuned Visual Large Language Models (LLMs). Stable Diffusion and DreamBooth Low-Rank Adaptation produce images that look authentic, with an average Inception score of 2.36 across three datasets. The validation accuracy of various classification models Big Transfer (BiT), Fixed Resolution Residual Next Generation Network (FixResNeXt), and Efficient Neural Network (EfficientNet) were 92%, 91%, and 86%, respectively. Vision Transformer (ViT) and Data-Efficient Image Transformers (DeiT) had an accuracy rate of 93%. Secondly, for the segmentation of polyps, the ground truth masks are generated using Segment Anything Model (SAM). Then, five segmentation models (U-Net, Pyramid Scene Parsing Network (PSNet), Feature Pyramid Network (FPN), Link Network (LinkNet), and Multi-scale Attention Network (MANet)) were adopted. FPN produced excellent results, with an Intersection Over Union (IoU) of 0.64, an F1 score of 0.78, a recall of 0.75, and a Dice coefficient of 0.77. This demonstrates strong performance in terms of both segmentation accuracy and overlap metrics, with particularly robust results in balanced detection capability as shown by the high F1 score and Dice coefficient. This highlights how AI-generated medical images can improve colonoscopy analysis, which is critical for early colorectal cancer detection.

Keywords: medical imaging synthesis; polyp detection; text-to-image generation; image segmentation; generative AI; medical image synthesis; colorectal cancer detection; data augmentation; synthetic colonoscopy images; text-to-image generation; DreamBooth; Stable Diffusion; Low-Rank Adaptation (LoRA); polyp segmentation; Feature Pyramid Network; Vision Transformer; Segment Anything Model; medical diagnostic models; healthcare AI

Academic Editor: Edward Rolando Núñez-Valdez

Received: 1 January 2025
Revised: 6 March 2025
Accepted: 6 March 2025
Published: 9 March 2025

Citation: Ejiga Peter, O.O.; Adeniran, O.T.; John-Otumu, A.M.; Khalifa, F.; Rahman, M.M. Text-Guided Synthesis in Medical Multimedia Retrieval: A Framework for Enhanced Colonoscopy Image Classification and Segmentation. *Algorithms* **2025**, *18*, 155. https://doi.org/10.3390/a18030155

1. Introduction

Artificial intelligence (AI) has initiated substantial transformations in the analysis of medical imaging. In gastroenterology, the scientific study of the health and diseases of the stomach, small intestine, colon, and rectum, AI-assisted colonoscopy has shown the potential to improve the diagnosis of colon cancer and reduce the incidence of colorectal cancer (CRC) [1]. The efficacy of artificial intelligence models is fundamentally dependent on the quantity and quality of training image data for colon cancer. A fundamental drawback in the advancement of successful models for medical imaging and image interpretation is the lack of sufficient training and testing data. The challenges of generating large datasets include privacy concerns, invasive methodologies, and the labor-intensive nature of professional data annotation [2].

The absence of extensive, varied, and thoroughly annotated datasets hinders the training and development of sophisticated AI-assisted colonoscopy technologies [3,4]. The inaccessibility of data in this context leads to the creation of severely defective AI models or algorithms with intrinsic bias when used for vulnerable populations [5]. The scarcity of ample high-quality training data undermines AI's efficacy in improving the identification and prevention of colorectal cancer, particularly among groups facing health care disparities. Like prior research, there has been an effort to utilize generative adversarial networks (GANs) for the synthesis of text-to-image mapping, among other methodologies [6–8]. Recent evaluations have emphasized the refinement of large language models based on vision, including Stable Diffusion (SD), Contrastive Language-Image Pre-training (CLIP), and DreamBooth + Low-Rank Adaptation (DB+LoRa), for the creation of medical images [9]. However, the application of these algorithms to generate a comprehensive array of high-quality colonoscopy images remains unexplored. The issue is the lack of a robust framework for generating synthetic colonoscopy images to enhance training data for more effective diagnostic algorithms.

Researchers are exploring various approaches (GANs, VQGAN, CLIP, Bootstrapping Language-Image Pre-training (BLIP)) to address these issues by improving and diversifying medical imaging databases [10–12]. This study examines the use of artificial intelligence for text-to-image synthesis and intelligent data augmentation to improve AI models and promote demographic equity. Inception score (IS) and Fréchet Inception Distance (FID) are utilized to evaluate synthetic images generated by fine-tuning the SD, CLIP and DB + LoRa models. In addition, advanced classification models such as BiT, ViT, FixResNeXt, EfficientNet, and DeiT are employed to train the generated images. Assessments of these models are based on standard metrics, including the F1 score, accuracy, precision, recall, and Area Under Curve (AUC), among others. All of this manifests itself as a structured plan to improve the standards of imaging and medical diagnosis.

2. Related Work

In recent years, artificial intelligence has demonstrated potential in medical image processing, improving the precision and efficacy of various diagnostic procedures, including colonoscopy [1]. Recent advances in AI for medical imaging have shown promising developments in various approaches. Pengxiao et al. [13] introduced X. Latent-based Diffusion Model for Long-tailed Recognition (LDMLR), which addresses long-tailed recognition challenges through latent-based diffusion models. Du et al. [14] developed Adaptive Refinement Semantic Diffusion Models (ArSDM), focusing on the scarcity of annotated medical image data through adaptive refinement semantic diffusion. Earlier work by Ku et al. [15] established TextControlGAN to improve image quality with text control. In 2024, Ejiga Peter et al. [9] conducted a comparative analysis of multiple approaches, including CLIP

and Stable Diffusion, while Iqbal et al. [16] developed a Conditional Generative Adversarial Network (GAN) for facial expressions.

In the specific domain of gastrointestinal imaging, several notable contributions emerged. Shin et al. [17] and Qadir et al. [18] implemented Convolutional Neural Network (CNN)-based approaches for polyp detection, while Dong et al. [19] introduced a transformer-based solution. Clinical applications have shown remarkable progress, with Repici et al. [20] demonstrating the efficacy of GI-Genius in colonoscopy and Kudo et al. [21] developing EndoBRAIN for polyp detection. Zhou et al. [22] contributed ENDOANGEL for real-time assessment, while Mahmood et al. [23] focused on depth estimation in endoscopy. Goceri [24] provided a comprehensive review of data augmentation techniques across different medical imaging modalities. This work investigates generative artificial intelligence in the context of colonoscopy imaging interpretation, specifically focusing on text-to-image generation and advanced data augmentation techniques. Early detection during colonoscopy is crucial to avert colon cancer, a major global health problem [5].

Kim et al. [3] investigated the generation of synthetic image data within a scientific framework, specifically focusing on medical image diagnostics and laboratory techniques, using the OpenAI DALL-E 3. However, the model exhibited low precision due to intrinsic bias within the system. They additionally examined the ethical implications of producing synthetic data. This research was constrained by the lack of an iterative generation process, which included erroneous or extraneous features, duplication of existing images, and reliance on proxy features. In a similar vein, Yang et al. [25] created an AI-generated image model to serve as a data source for synthetic data. Upon thorough examination, Yang et al. [25] employed CovNets and transformer-based models. Convolutional neural network models, such as ResNet-50 and ResNet-101, demonstrate substantial improvements. Transformers demonstrate improvements in ViT-S/16, DeiT-S, and DeiT-B. The larger versions typically exhibit superior overall performance. Our strategy emphasizes text-to-image synthesis utilizing fine-tuned Large Language models for the development of synthetic colonoscopy images and Classification models for polyp detection, facilitating the early identification of colorectal cancer.

Cao et al. [26] examined AI-generated content (AIGC) and its historical evolution from Generative Adversarial Networks (GANs) to contemporary ChatGPT, emphasizing the unimodality and multimodality of generative AI models. Bandi et al. [27] conducted a survey of various GAN-based designs for AIGC and examined the fundamental needs for the development of AIGC, including the input–output format of the model, the evaluation metrics, and the associated obstacles. However, the analyses presented in both [26] and Bandi et al. [27] were purely theoretical and devoid of practical application; therefore, our research transcends qualitative examination by incorporating the practical implementation of AIGC with cutting-edge models such as SD. Bendel [28] examined image synthesis (image production) from an ethical perspective. Elements of creation can also facilitate destruction, which is why Bendel [28] advocated the ethical application of generative AI in image synthesis. The significant potential of image synthesis models entails a considerable duty to guarantee ethical utilization. Bendel [28] used the Midjourney, SD, and DALL-E2 models to analyze the topic from the point of view of risk and opportunity. In a comparable context, Ray et al. [28] addressed the meticulous supervision required to realize the potential of generative AI in precise medical diagnostics and neurointerventional surgery while mitigating the risk of encountering pitfalls. Both works lack practical application and are purely theoretical. However, ethical considerations were implemented during the construction and testing of the models in our research.

Derevyanko et al. [29] conducted a comparative examination of SD, Midjourney, and DALL-E, advocating for their implementation due to their educational use. The study

demonstrated that neural network applications can improve the teaching of students in design-related fields. Mahmood et al. Al [23] developed a depth estimator utilizing an extensive dataset of synthetic images produced through a precise forward model of an endoscope and an anatomically accurate colon. Their studies revealed the following enhancements in porcine colon data: an 88% improvement in Structural Similarity Index Measure (SSIM) for depth estimation, a 48% improvement in SSIM for phantom colon data depth estimation, and approximately a 125% improvement compared to the prior dictionary learning method. Additional validation of various medical imaging tasks and modalities would enhance the broader application.

Iqbal et al. [16] developed a synthetic image generation model using Conditional GAN for single-sample face images. The research achieved an initial accuracy of 76% on individual neutral images and a final accuracy of 99% after the fine-tuning of synthetic expressions, representing a 23% improvement over the original accuracy recorded. This method presents possibilities to address the difficulty of Single Sample Per Person (SSPP) in facial recognition systems. However, it is limited solely to variations in expression and can necessitate the integration of additional motions to capture other forms of facial alterations for applicability in different contexts.

Du et al. [14] used the Adaptive refinement semantic diffusion model for the synthesis of colonoscopy images. Polyp segmentation exhibited the greatest enhancement with PraNet (6.0% mDice, 5.7% mIoU), succeeded by SANet and Polyp-PVT. CenterNet demonstrated the most significant gain in average precision (9.1%) and the greatest improvement in the F1 score (6.1%) for the diagnosis of polyps. The research indicates that the use of synthetic images generated by ArSDM for training significantly enhances the results in both polyp segmentation and detection tasks. The methodology effectively addresses weakly annotated medical imaging data.

Ku et al. [15] present a novel methodology for text-to-image synthesis that improves image quality and adherence to textual descriptions compared to previous methods. Compared to the Caltech-UCSD Birds-200-2011 (CUB) dataset, the model improved the Inception score by 17.6% and decreased the Fréchet Inception Distance by 36.6% relative to the GAN-INT-CLS model based on cGAN. The model is limited to bird images at a relatively low resolution (128×128) and was assessed only on a single dataset, CUB.

Ejiga Peter et al. [9] employed three methodologies for medical image synthesis: DB + LoRA, fine-tuned SD, and CLIP. They analyzed numerous advanced strategies and tackled the need for dynamic generation of medical images from textual descriptions. The models were evaluated using single-center and multicenter datasets. SD beat other Fréchet Inception Distance techniques. The highest multicenter score of 0.064 indicated excellent image quality. DB + LoRA outperformed CLIP (1.568) and SD (2.362) in the initial score. SD produced diversified and high-quality images efficiently, beating FID and competing in IS. Their analysis did not evaluate the images generated clinically or diagnostically. The research shows the potential of AI-generated medical imaging, but it also stresses the need for clinical trials and ethical considerations.

Sanchez et al. [30] explored how deep learning has improved colonoscopy polyp identification. The CNN is the most popular architecture for end-to-end applications; hybrid methods are less popular. The biggest issues are large, annotated datasets and inconsistent evaluation metrics. Their research shows that these methods can improve the identification of adenoma but require clinical validation. Standardization of assessment methods, semisupervised learning, elimination of Histogram of Displacement (HD) colonoscopy images, and improved access to large public datasets can result.

Goceri [24] conducted an extensive evaluation of data enhancement methodologies for medical imaging. The efficacy of augmentation techniques varies according to the

specific medical imaging modality used. Shearing and translation proved to be most effective for brain Magnetic Resonance Imaging (MRI), lung Computed Tomography (CT), and breast mammography, but noise-based methods often faltered. The most precise components of brain MRI were translation, shearing, and rotation, with an accuracy of 88.24%. Translation by shearing exhibited optimal performance on lung CT at 85.74%. Translation, shearing, and rotation were the most beneficial for breast mammography (83.34%). The results for color shift, sharpening, and contrast modification were most pronounced in the classification of the eye fundus, with a precision of 88.24%. Shearing and noise performed poorly, but the combinations performed better. Fundus images had the greatest improvement in color and contrast. GAN-based methods are more diverse, yet suffer from gradient vanishing [31,32]. The assessment stresses image categorization and the selection of task-related augmentation techniques.

Wang et al. [33] used deep learning in real time to identify colonoscopy polyps. In addition to 27,113 images and videos, 1290 patient records confirmed findings. The approach performed well in datasets, with per-image sensitivity from 88.24% to 94.38% and specificity from 95.40% to 95.92%. Please note that the confirmed polyp videos were 100% sensitive. The gadget can handle 25 frames per second with low latency for real-time clinical applications. This method allows endoscopists to test their polyp-detecting skills.

This pilot study by Misawa et al. [34] evaluates a computer-aided detection (CADe) system for the real-time identification of polyps during colonoscopy. Although previous CADe systems were capable of recognizing over 90% static images and 70% videos, the extract failed to include the performance characteristics of the system. Expanding datasets and including video-based analysis could improve the practical importance of this research. The absence of performance metrics in the passage is concerning. The initiative aims to reduce human error in polyp detection, thus improving adenoma diagnosis and perhaps decreasing the rates of colorectal cancer in intervals.

Dong et al. [19] used Polyp-olyp Pyramid Vision Transformer(PVT) and a transformer encoder that incorporates Context Feature Module (CFM), Cross-interaction Module (CIM), and SAM segment polyps, achieving a score of 0.900 Dice and 0.833 IoU on the Endoscene dataset, surpassing previous approaches in five datasets. The models appear to be resilient to variations in appearance, tiny objects, and rotation, demonstrating strong performance on hitherto unseen data. The limits of polyps are difficult to see due to overlapping light and shadow, as well as numerous false positives from the reflection point. The model surpasses polyp segmentation techniques in recommended colonoscopy applications.

Guo et al. [35] introduced Dilated ResFCN and SE-Unet, two innovative fully convolutional neural network (FCNN) architectures for the segmentation of colonoscopy polyps. The average Dice score was 0.8343, the standard deviation was 0.1837, and only three polyps were overlooked. The strategies mentioned above secured victories in the Gastrointestinal Image ANAlysis (GIANA) tournaments in 2017 and 2018. The study was beneficial; however, the temporal dependencies of colonoscopy records should enhance future research.

Qadr et al. [18] used Mask R-CNN to identify and segment colonoscopy polyps to reduce the physician 25% missed rate. Multiple CNN feature extractors and additional training data were evaluated to construct Mask R-CNN. An ensemble method for enhancement was proposed. The segmentation of the leading model was performed on the 2015 Medical Image Computing and Computer Assisted Intervention (MICCAI) dataset, with the following results: recall 72.59%, accuracy 80%, Dice coefficient 70.42%, and Jaccard index 61.24%. The study elucidates the trade-offs between model complexity and dataset quality for better polyp identification and segmentation. The results are promising, but automated polyp recognition should be improved.

Borgli et al. [36] tested CNN architectures for gastrointestinal endoscopy using 110,079 images and 374 movies for multiclass classification. Size, diversity, segmentation masks, and bounding boxes help, but class imbalance and interobserver variability hurt. ResNet-152 and DenseNet-161 achieved amazing results with microaveraged F1 scores of 0.910, macroaverages of 0.617, and Matthews Correlation Coefficients of 0.902. Despite its appeal, the model struggled to distinguish esophagitis from ulcerative colitis. The material is suitable for AI-assisted gastrointestinal endoscopic diagnostic tools; however, classification accuracy needs to be improved.

Kudo et al. [21] used a large training dataset to test EndoBRAIN, an AI system that detects colorectal polyps from endocytoscopic images. The model outperformed the trainee and professional endoscopists in stained image analysis with 96.9% sensitivity, 100% specificity, and 98% precision. Size and various imaging modalities are benefits, but retrospective design and Japanese center concentration are weaknesses. EndoBRAIN's 100% positive and 94.6% negative predictive scores improve colorectal polyp diagnosis. More prospective trials are needed to prove that AI can improve endoscopic diagnostics.

Zhou et al. [22] developed ENDOANGEL, a deep convolutional neural network-based method to assess the quality of intestinal preparation through precise image and video analysis. This model surpassed human endoscopists, achieving 93.33% accuracy on conventional images and 80.00% accuracy on bubble images. The video examination demonstrated an accuracy of 89.04%. The benefits of ENDOANGEL include an objective and consistent evaluation, immediate colonoscopy scoring, and enhanced image and video quality. The study had limitations, including the use of retrospective data collection for training and restricted video assessment. The reproducible evaluation of intestinal preparation as a continuous variable during colonoscopy withdrawal is an innovative approach that enhances the potential to improve colonoscopy results and standardize the evaluation of stool preparation quality.

The literature review shows that AI for colonoscopy analysis, particularly polyp detection and categorization, has advanced. However, significant gaps and constraints remain due to the bias of the models resulting from the reduced variability of the training dataset [3,25]. Some studies have examined novel AI-generative systems for the generation of medical images [9,14]. The lack of comprehensive systems that incorporate segmentation, classification, and image creation for colonoscopy analysis [14,25] and the lack of studies on artificial data augmentation and model fairness among demographic groups [30] are major problems. The study is expected to contribute significantly to AI-assisted colonoscopy analysis, as these prospects fit with the research objectives and questions.

3. Materials and Methods

The research technique employs a complete framework for the interpretation of medical images, with a specific emphasis on colonoscopy imaging. This framework incorporates three fundamental components: image production and augmentation, classification, and segmentation, utilizing cutting-edge deep learning models to establish a resilient system for colonoscopy image processing and analysis. Our methodology integrates contemporary artificial intelligence tools with conventional medical imaging practices to improve the detection and categorization of polyps in colonoscopy images, as seen in Figure 1. Originally trained on the LAION-5B dataset [37], the SD model offers a wealth of information from a wide spectrum of image-text combinations. Comprising 5.85 billion CLIP-filtered image-text pairs (2.3 billion English samples, 2.2 billion from more than 100 additional languages) and 1 billion with language-agnostic text, LAION5B is a vast collection.

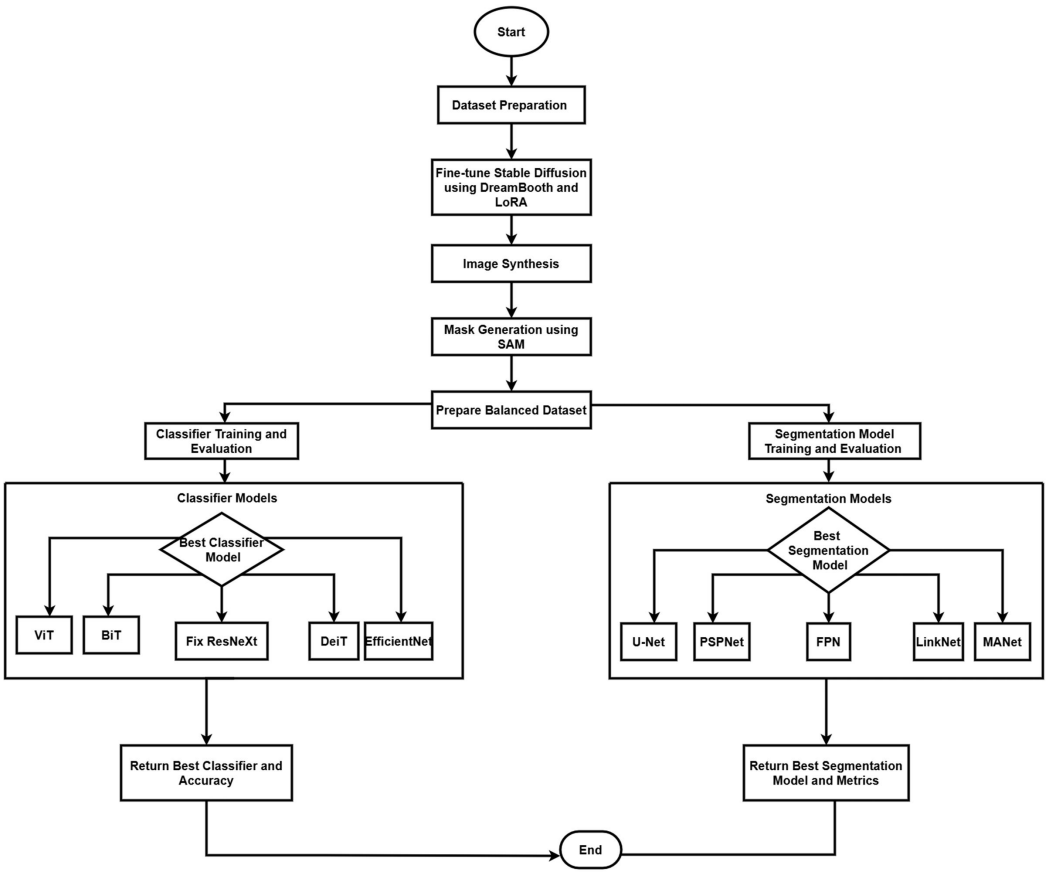

Figure 1. Comprehensive flow diagram of the research methodology, illustrating the sequential process from dataset preparation through model development and evaluation.

3.1. Dataset Description

The study uses the Conference and Labs of the Evaluation Forum (CLEF) dataset for the image generation task, which contains 2000 colonoscopy images paired with 20,242 training prompts and 5000 non-image test prompts to train and evaluate CLIP, SD, and DB LoRa. Research evaluates the ability of models to generate medical images from textual descriptions [9]. This study uses a dataset of 15,428 images classified into three main groups. The initial dataset comprises 5914 images obtained from reputable medical imaging repositories that include the CLEF Image [38], Cerebral Palsy-Comprehensive Health Index of Limb Disorders (CP-CHILD) [39], and binary polyp data [40]. To enhance the diversity and robustness of the dataset, we sampled 1800 synthetic images using fine-tuned Stable Diffusion and created 7714 augmented images through various data augmentation methods. The Classification Model Dataset contains 2946 original polyp and 2967 polyp images, supplemented with 900 polyp and non-polyp synthetic images. When combined with the original dataset, this creates an augmented dataset of 3846 non-polyp and 3867 polyp images, enhancing the diversity and balance of the training data. The complete dataset is partitioned using a 70-20-10 split ratio for training, testing, and validation, respectively, ensuring an equitable distribution of polyp and non-polyp images in each segment to promote impartial model training and assessment. The 70:20:10 split allocates 70% for training data to learn patterns, 20% for validation to adjust hyperparameters and prevent overfitting, and 10% for testing to evaluate the performance of the final model. This balanced ratio ensures sufficient training while maintaining independent validation and testing sets.

For the image segmentation task, the dataset contains 1824 images distributed in four distinct folders. The real-world data are stored in two folders: *real_images*, which contains 612 original colonoscopy images, and *real_masks*, which contain their corresponding

612 segmentation masks. The synthetic data, generated using Fine-tuned Stable Diffusion and DreamBooth LoRa, are organized in *mdpi_syn* with 300 synthetic colonoscopy images and *mdpi_masks* with their 300 corresponding segmentation masks.

The image generation pipeline uses stable diffusion 1.5, obtained by Hugging Face (Hugging Face, New York, USA), as its principal model. The system employs two advanced fine-tuning methodologies: DreamBooth (DB) for subject-specific customization and Low-Rank Adaptation (LoRA) for efficient parameter modifications. Our training parameters used a 4-step batch size with 8-bit Adam optimization, integrating gradient checkpointing and bfloat16 precision. The training procedure executes for 1000 iterations utilizing logit normal weighting to guarantee optimal convergence. To improve contextual awareness and the quality of image formation, we used the Contrastive Language Image Pre-training (CLIP) method, using its zero shot transfer abilities and natural language supervision attributes. See Figure 2.

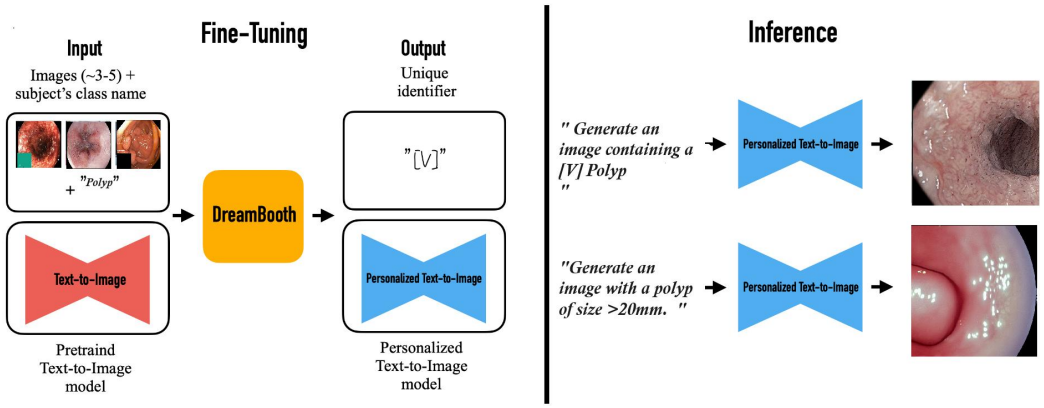

Figure 2. DB fine-tunes text-to-image model with few polyp images to generate personalized colonoscopy images.

3.2. Training Pipeline

The image processing pipeline starts with a thorough standardization procedure, in which all images are resized to 256 × 256 pixels with LANCZOS resampling, subsequently converting them to RGB color space. Resizing to 256 × 256 ensures consistent memory usage and fixed network inputs while preserving features for subsequent 224 × 224 random cropping during training. All images are stored in PNG format to preserve image quality through lossless compression. The training set is subjected to further augmentation techniques, including color jittering, random rotation of ±15 degrees, random horizontal flipping, and random scaled cropping to 224 × 224 pixels. The model sees slightly different versions of the same image during training, improving its robustness and generalization capabilities. All images are subsequently normalized using standard ImageNet normalization parameters and processed in batches via PyTorch python library DataLoader with a batch size of 32 for optimal training efficiency. Pytorch was created from Facebook's AI Research lab (FAIR), Menlo Park, California, USA. PyTorch 2.2.x was used in this research.

We executed and assessed five cutting-edge models for classification tasks: Vision Transformer (ViT), Big Transfer (BiT), FixResNeXt, Data-efficient Image Transformer (DeiT) and EfficientNet. Each model possesses distinct advantages for the classification challenge, with transformers that deliver strong feature extraction skills and convolutional networks that facilitate effective spatial relationship processing. The segmentation component of our approach employs five distinct architectures, U-Net, PSPNet, Feature Pyramid Network (FPN), LinkNet, and MANet. The models were chosen for their demonstrated efficacy in

segmenting medical images and their ability to address the unique problems posed by colonoscopy images.

The training methodology utilizes the AdamW optimizer with weight decay and Cross Entropy Loss as a loss function. To avoid overfitting and guarantee optimal model performance, we established an early stopping mechanism based on the plateau of validation loss, consistently retaining the best validation model weights during the training phase. This method guarantees that our models preserve their generalizability while attaining superior performance on the designated task. The binary cross-entropy loss function for classification is defined as

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{1}$$

where N represents the number of samples, y_i is the true label, and $\hat{y}_i$ is the predicted probability for the positive class. For the LoRA adaptation process, the weight matrix transformation is computed as:

$$W = W_0 + BA = W_0 + \Delta W \tag{2}$$

where W_0 represents the original weight matrix, B and A are low-rank decomposition matrices and ΔW is the update matrix. The Dice coefficient, used for evaluating segmentation performance, is calculated as:

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{3}$$

where X and Y represent the predicted segmentation masks and ground truth, respectively, TP denotes true positives, FP denotes false positives and FN denotes false negatives. Our evaluation approach integrates many metrics to thoroughly evaluate distinct facets of the system's performance. The quality of images is assessed by the Inception score (IS) and Fréchet Inception Distance (FID), which offer quantifiable metrics for the caliber of synthetic image synthesis. Classification performance is evaluated using many metrics, such as accuracy, precision, recall (sensitivity), F1 score, and area under the ROC curve (AUC-ROC). For segmentation tasks, we assess quality by Intersection over Union (IoU), Dice coefficient, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM), guaranteeing a comprehensive evaluation of segmentation precision and dependability [41–47]. The CLEF data comprise text-image pairs, as seen in Figure 3.

Figure 3 displays image-text pairs in which a number of descriptive text prompts are paired with each colonoscopy image. In addition to having distinct labels (such as 064086616.png and 040086376.png), each image features a set of text questions that offer several approaches to describing or creating comparable medical situations. This pairing structure produces a comprehensive dataset with detailed descriptions of every medical image from various textual perspectives. The instructions in each image addresses a variety of topics related to the medical scene, from identifying locations and image features to describing apparent instruments and discoveries. This creates a rich image–text relationship in which a single medical image is connected to multiple relevant text descriptions, each highlighting different aspects of the same clinical scenario.

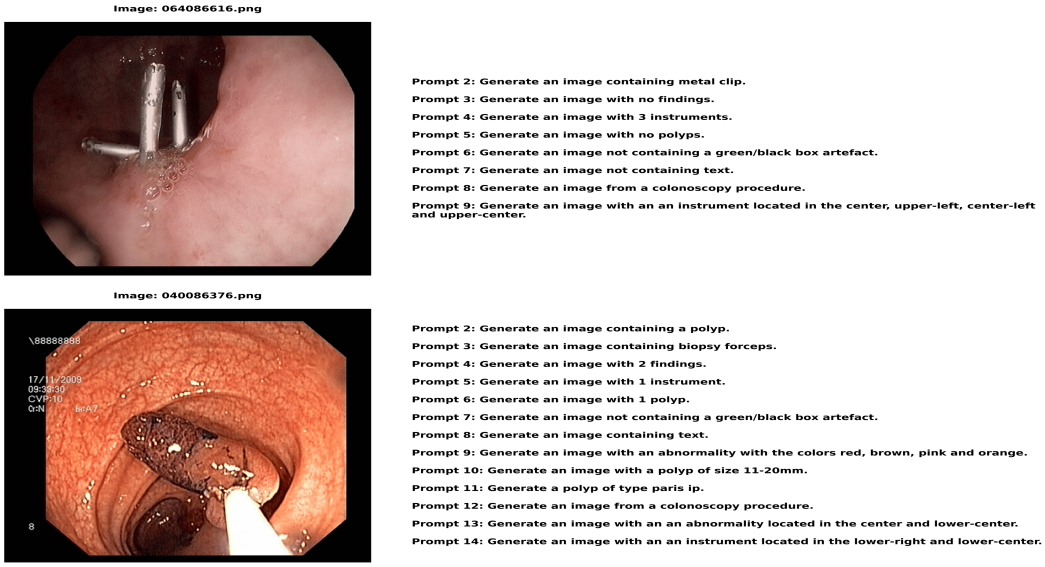

Figure 3. The dataset shows image–text pairs where each colonoscopy image is matched with multiple descriptive text prompt images.

3.3. Algorithm

The preprocessing pipeline coordinates the data preprocessing process through a strategic multihierarchical procedure. The first step is to turn the file that contains image locations and prompts into a structured framework using the pandas library by sorting out the data for polyps and non-polyps. This becomes the structural format of organized data that can feed other pre-processing and training model data stages. In the data partitioning process, random shuffling is used to have equal distribution over training, testing, and validation. The images are analyzed using the Python Imaging Library and each step of the pipeline is designed to support different classification models to achieve the best performance. The process of synthesizing colonoscopy images and training classification and segmentation models is formalized in Algorithm 1.

This algorithm presents the complete pipeline of our methodology, from data preparation through model training and evaluation. The algorithm takes as input colonoscopy images, custom prompts, and pre-trained models and outputs the best performing classifier and segmentation model along with their respective performance metrics. Each function in the algorithm represents a distinct phase in our pipeline, with the main process divided into the data preparation, synthesis, classification training, and segmentation model training stages. For reproducibility and convenience in the case of further investigations of this topic, all code, including models, training procedure scripts, and assessment schemes—is committed to the repository [48]. The datasets included in this study are available through their respective platforms, CLEF [38], CP-CHILD [39], and binary polyp data [40]. The synthesized images produced and their associated prompts can be accessed in [48]. The repository contains a detailed documentation of all implementation specifics, including hyperparameters and model setups, to facilitate the replication of our results and the advancement of the approach.

Algorithm 1 Colonoscopy Image analysis through text-to-image synthesis using generative AI

1. **Input:** Colonoscopy images, Custom prompts, Pre-trained Stable Diffusion 1.5, SAM model
2. **Output:** Best classifier C, Accuracy Acc, Best segmentation model S, Performance $Perf$
3. Dataset Preparation:
   (a) $D_{orig} \leftarrow$ PrepareDataset (images, prompts)
   (b) $G_{ft} \leftarrow$ FineTuneStable (Diffusion 1.5, D_{orig})
   (c) $D_{syn} \leftarrow$ SynthesizeImages (G_{ft}, prompts, 5000)
   (d) $D_{syn\_mask} \leftarrow$ GenerateMasks (D_{syn}, SAM)
   (e) $D_{bal} \leftarrow$ SampleBalancedDataset (D_{orig}, D_{syn_mask}, 1000)
4. Classifier Training:
   (a) **for** each M in {ViT, BiT, FixResNeXt, DeiT, EfficientNet} **do**
      i. $C_M \leftarrow$ TrainClassifier (M, D_{bal})
      ii. $Acc_M \leftarrow$ EvaluateClassifier (C_M, D_{bal})
5. Segmentation Model Training:
   (a) **for** each S in {U-Net, PSPNet, FPN, LinkNet, MANet} **do**
      i. $S_{trained} \leftarrow$ TrainSegmentation (S, D_{bal})
      ii. $Perf_S \leftarrow$ EvaluateSegmentation ($S_{trained}$, D_{bal})
6. **return** Best classifier C, Accuracy Acc, Best segmentation model S, Performance $Perf$

Helper Functions:

FineTune (G, D):	Fine-tune model using DreamBooth and LoRA
SynthesizeImages (G, prompts, n):	Generate n images using G
GenerateMasks (D, SAM):	Generate masks for images in D using SAM
TrainClassifier (M, D):	Train classifier M on dataset D
EvaluateClassifier (C, D):	Evaluate classifier C
TrainSegmentation (S, D):	Train segmentation model S on dataset D
EvaluateSegmentation (S, D):	Evaluate segmentation model S

3.4. Software and Hardware Requirements

Python 3.10 programming language was used to implement all algorithms. The code was developed and executed using Anaconda 2023.09 distribution for environment management, Google Colab Pro (subscription-based), and Vertex Workbench 1.11 cloud platforms. We utilized various libraries including accelerate 0.16.0, PyTorch 2.0.1, torchvision 0.15.2, transformers 4.25.1, huggingface hub 0.15.1, diffusers 0.14.0, xformers 0.0.20, and bitsandbytes 0.37.0. Computations were performed on high-performance hardware: g2-standard-48 machines (4 NVIDIA L4 GPUs, 48 vCPUs, 192 GB RAM) and a 40 GB NVIDIA A100 GPU. The system utilized 83.5 GB of total RAM (with 2.2 GB utilized during the experiments) and 40.0 GB of GPU RAM, with a storage capacity of 235.7 GB (100 GB data disk + 150 GB boot disk). With pretrained Stable Diffusion 1.5 and SAM models for image synthesis and segmentation, this configuration facilitated a complete colonoscopy analysis pipeline that was evaluated using a variety of classifier architectures (ViT, BiT, FixResNeXt, DeiT, EfficientNet) and segmentation models (U-Net, PSPNet, FPN, LinkNet, MANet).

4. Results

The results and a discussion of the research are presented in this chapter, which details both software and hardware specifications. Table 1 presents a comparison of performance metrics (FID scores and average inception scores) in three different models (CLIP, SD and DB + LoRa) in three different datasets (single, multi and both). The FID scores range from

0.06 to 0.12, with SD showing the best performance (lowest FID scores) in all datasets. For Inception scores (IS avg), DB + LoRa consistently achieves the highest values (2.36), followed by SD (2.33), while CLIP shows the lowest IS avg (1.57). Each model maintains consistent performance across different datasets, with only minor variations in FID scores.

Table 1. Performance comparison of different models across datasets. The evaluation metrics include FID score (lower is better) and Inception score (IS) measurements, including average (avg), standard deviation (std), and median (med) values.

Dataset	Model	FID	IS avg	IS std	IS med
single	CLIP	0.11	1.57	0.03	1.56
multi	CLIP	0.11	1.57	0.03	1.56
both	CLIP	0.12	1.57	0.03	1.56
single	SD	0.06	2.33	0.07	2.34
multi	SD	0.06	2.33	0.07	2.34
both	SD	0.07	2.33	0.07	2.34
single	DB+LoRa	0.11	2.36	0.05	2.36
multi	DB+LoRa	0.07	2.36	0.05	2.36
both	DB+LoRa	0.08	2.36	0.05	2.36

Table 2 presents a detailed evaluation of model performance using Inception Score Groups (G1-G10) in three different models (CLIP, SD, and DB+LoRa) and three dataset configurations (single, multi, and both).

Table 2. Detailed Inception score evaluation across different groups (G1-G10) for each model and dataset combination. Groups represent different aspects of image quality assessment, showing model consistency across evaluation dimensions.

Dataset	Model	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
single	CLIP	1.56	1.55	1.55	1.57	1.55	1.55	1.54	1.54	1.56	1.51
multi	CLIP	1.56	1.55	1.55	1.57	1.55	1.55	1.54	1.54	1.56	1.51
both	CLIP	1.56	1.55	1.55	1.57	1.55	1.55	1.54	1.54	1.56	1.51
single	SD	2.37	2.40	2.40	2.39	2.24	2.26	2.31	2.22	2.25	2.77
multi	SD	2.37	2.40	2.40	2.39	2.24	2.26	2.31	2.22	2.25	2.77
both	SD	2.37	2.40	2.40	2.39	2.24	2.26	2.31	2.22	2.25	2.77
single	DB+LoRa	2.35	2.27	2.39	2.41	2.31	2.33	2.34	2.36	2.46	2.38
multi	DB+LoRa	2.35	2.27	2.39	2.41	2.31	2.33	2.34	2.36	2.46	2.38
both	DB+LoRa	2.35	2.27	2.39	2.41	2.31	2.33	2.34	2.36	2.46	2.38

The scores show that CLIP consistently maintains lower values (around 1.51–1.57) across all groups, while SD and DB+LoRa achieve higher scores (ranging from 2.22–2.77). In particular, SD shows greater variability with peaks in G2–G4 ($\approx$ 2.40) and G10 (2.77), while DB+LoRa demonstrates more stable performance between groups with slight improvements in G4 (2.41) and G9 (2.46).

4.1. Image Generation Results

The synthesized images were measured using the inception score and the Fréchet Inception Distance. Among the three models, CLIP has the highest Fréchet FID score: 0.114 for 0.128 and 0.124 for three (3) datasets [9]. These higher ratings indicate poorer image quality and realism as CLIP-generated images seem less like real photos than other models. Figure 4 shows the final image. The higher FID score for the multicenter dataset indicates that CLIP struggles to create realistic images when trained on a diverse dataset from numerous medical facilities. High FID scores indicate that in this case, CLIP is not the best approach to obtain high-quality medical images. The images in Figure 4 demonstrate

CLIP's capability to generate synthetic colonoscopy images from textual descriptions, though with notable quality limitations when trained on diverse medical facility datasets.

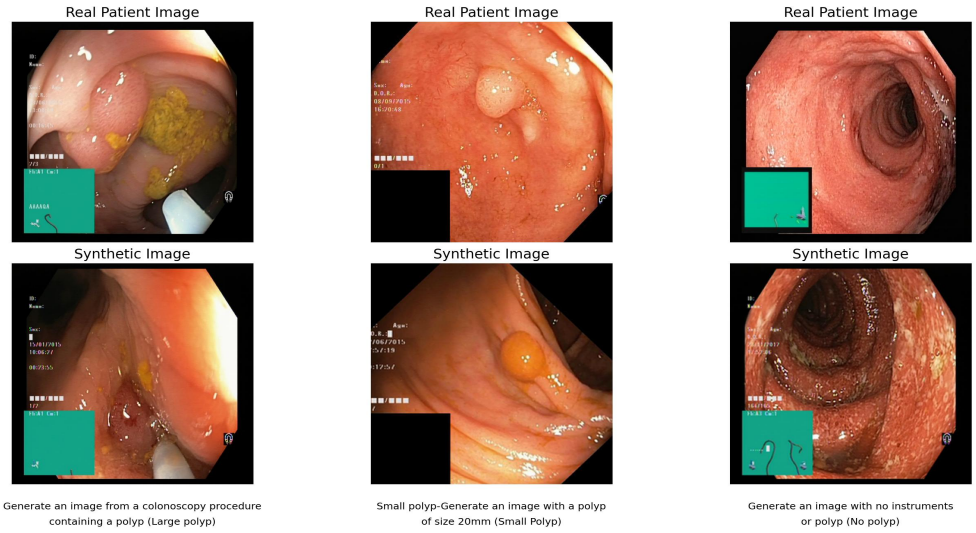

Figure 4. CLIP text-to-image-based colonoscopy image generation. Comparison between real patient colonoscopy images (**top** row) and synthetic images generated using CLIP (**bottom** row) based on text prompts. **Left**: Generation of a large polyp. **Center**: Generation of a small 20mm polyp. **Right**: Generation of a normal colon view without instruments or polyps.

The implementation combines Stable Diffusion pre-trained, DreamBooth, and LoRA for the synthesis of colonoscopy images, optimized with a resolution of 512 pixels, a learning rate of $1 \times e^{-4}$, and specialized medical prompts. This architecture efficiently handles medical imaging complexity while minimizing computational resources through LoRA's targeted modifications and DreamBooth's instance-specific adaptations. Indicating better image quality and realism, the lowest FID scores for fine-tuned SD are "0.099" (single center), "0.064" (multicenter), and "0.067" (combined datasets) [9]. The reduced score of the multicenter dataset implies better performance of healthcare facilities. Although somewhat inferior to fine-tuned DB + LoRa, SD offers a high amount of visual diversity and quality, averaging "2.33" in all datasets. The consistency of SD across datasets shows its ability to function regardless of the data source. In general, fine-tuned SD produces high-quality, diversified medical images. Figure 5 shows the synthetic images that demonstrate SD's ability to generate diverse high-quality medical images with FID scores of 0.064–0.099 across different datasets, showing consistent performance regardless of data source.

The FID scores of "0.11" (single center), "0.073" (multicenter), and "0.076" (combined datasets) obtained by fine-tuning DB + LoRa show a high level of image quality and accuracy. With a wider range of training data, the multicenter score is more favorable. However, scores are higher than for SD, which has been fine-tuned, indicating a good level of quality overall. Regardless of the data source, the Inception score of "2.36" remains constant across all datasets, indicating strong image diversity and quality [9]. Figure 6 shows the result of the text-to-image synthesis using DB and LoRa. The synthetic images demonstrate the model's ability to generate colonoscopy images in response to specific text prompts, with each pair showing the relationship between real patient images and their synthetic counterparts.

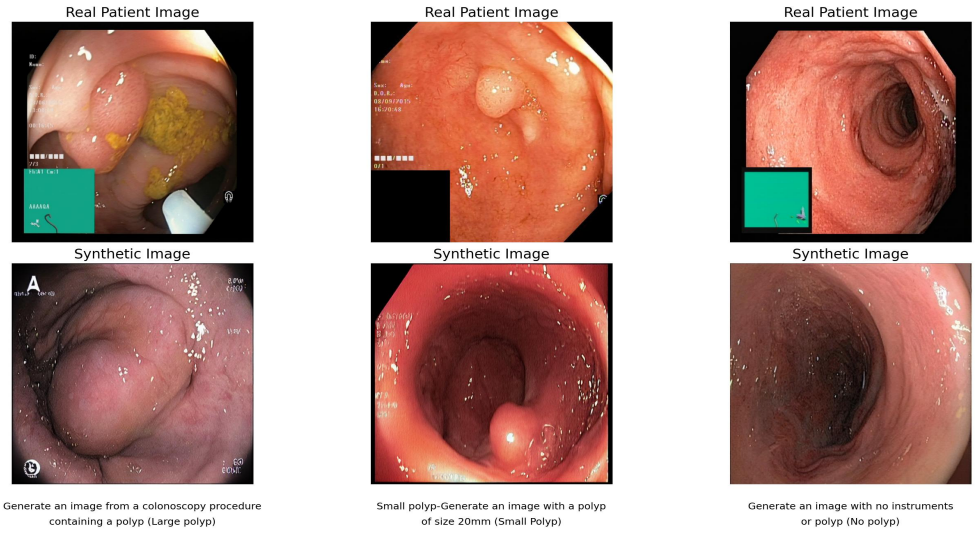

Figure 5. Fine-tuned SD text-to-image-based colonoscopy image generation. Comparison between real patient colonoscopy images (**top** row) and synthetic images generated using fine-tuned Stable Diffusion (**bottom** row). The image pairs show three scenarios: (**left**) a large polyp visualization, (**center**) a small 20mm polyp, and (**right**) normal colon without instruments or polyps.

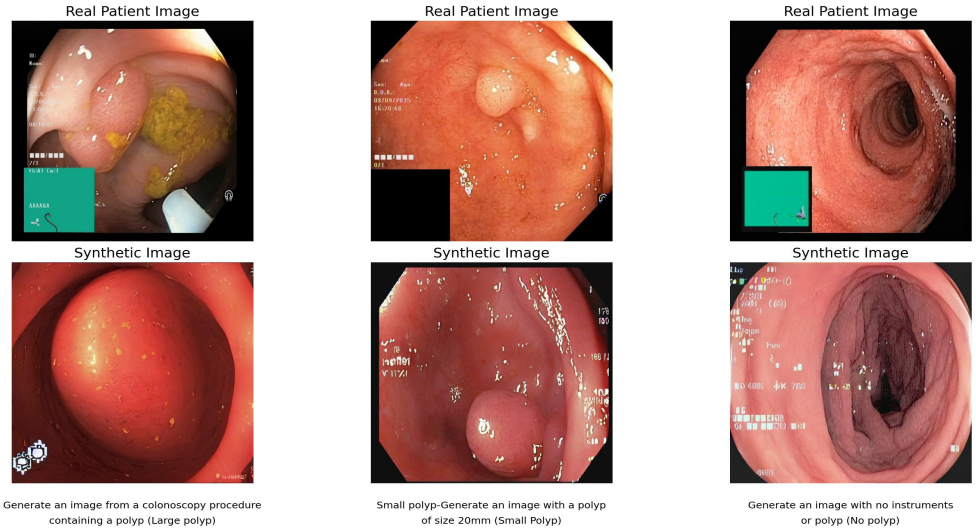

Figure 6. DB and LoRa text-to-image-based colonoscopy image generation. Comparison of real patient colonoscopy images (**top** row) with synthetic images generated using DB and LoRa models (**bottom** row). Three clinical scenarios are presented: (**left**) colonoscopy procedure showing a large polyp, (**center**) visualization of a small 20mm polyp, and (**right**) normal colon view without instruments or polyps.

In terms of image variety and value, these scores are slightly better than those of SD. Fine-tuned SD achieves higher FID scores, but fine-tuned DB + LoRA strikes a better balance between diversity and realism, potentially making it the best choice for medical imaging tasks that require high-quality and diverse image generation.

4.2. Model Comparison

Figures 7 and 8 show the evaluation of each visual generative AI model in terms of FID and IS in three datasets.

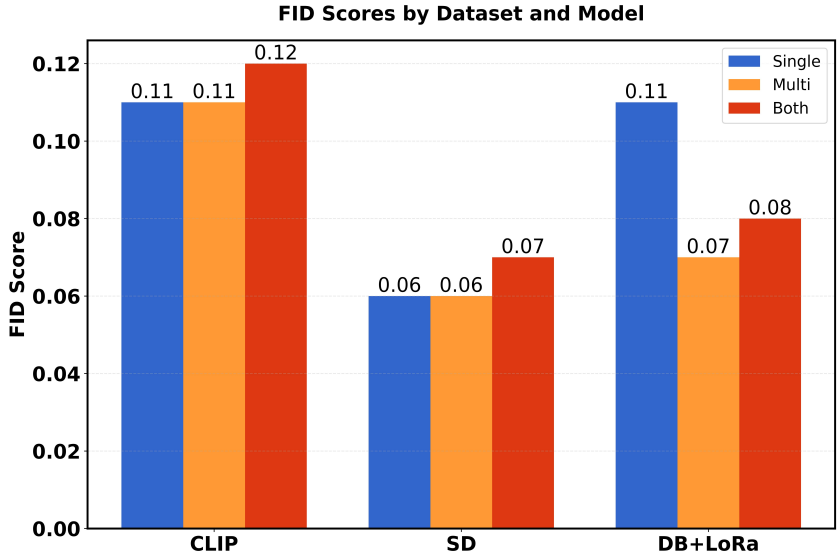

Figure 7. FID comparison across three models and three datasets.

While fine-tuned DB and LoRa produce high-quality, lifelike images, fine-tuned SD beats CLIP and fine-tuned DB in generating such images. Its superiority is shown by producing the lowest FID values among all tests. With an average Inception score of 2.36, the DB + LoRa model ranks highest followed by fine-tuned SD, with an average of 2.33. The average starting score for CLIP is the lowest. However, fine-tuned DB + LoRA achieves a better balance between diversity and realism, so it may be the perfect solution for medical imaging positions that require both high-quality and diverse image generation.

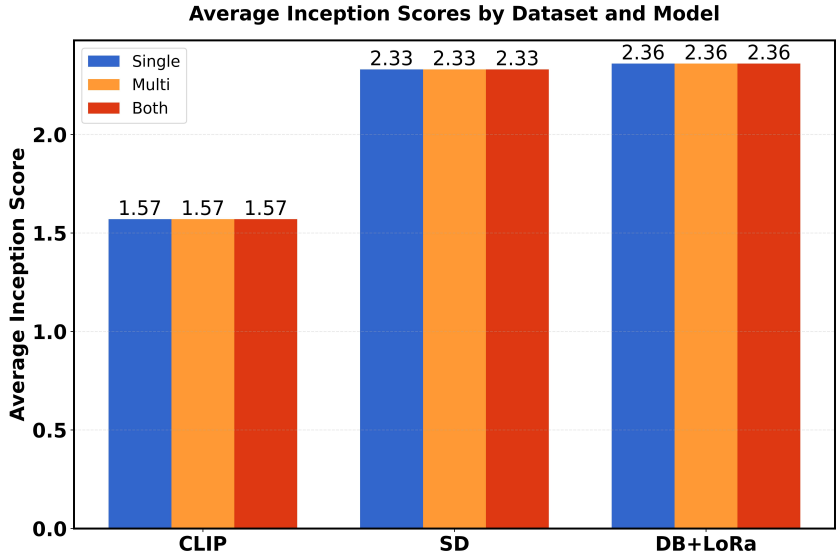

Figure 8. IS comparison across three models and three datasets.

4.3. Image Segmentation Results

This section comprises two primary tasks, namely image mask generation and image segmentation. The dataset contains 1824 images: 612 real colonoscopy images with masks and 300 synthetic images with masks. A pretrained faster R-CNN with ResNet50 backbone first classifies and localizes polyp regions using bounding boxes, which then guide the SAM model for precise mask generation. Image masks were created for shapes identified with the Segment Anything Model (SAM) [49]. The technique starts with initializing SAM with a pre-trained checkpoint and finds circles and outlines in the input image that mimic polyps.

Input points are produced along the boundary of every form found. Figure 9 shows the location of the polyp in both real and AI-generated endoscopic views, highlighting the similarity between actual and synthetic medical imaging data. Following mask generation with the Segment Anything Model (SAM), we performed manual visual screening of all masks generated. This validation process involved systematic examination of each image–mask pair to verify accuracy and alignment with anatomical boundaries. Masks that exhibited inaccuracies or did not properly delineate the target structures were excluded from further analysis, ensuring that the dataset contained only precisely segmented regions for subsequent processing. For model training and evaluation, a standard 80:20 train–test split was applied. This approach ensures robust model development by allocating 80% of the data for training and reserving 20% for independent testing, allowing for a comprehensive assessment of the performance and generalizability of the segmentation models.

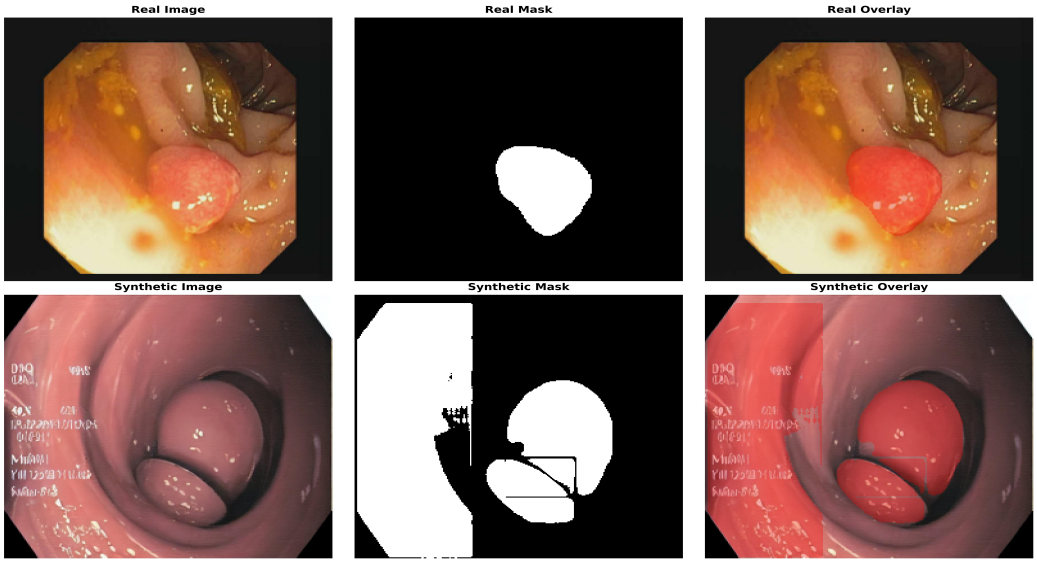

Figure 9. Comparison of real and synthetic colonoscopy images with polyp detection. Top row shows a real colonoscopy image with its corresponding binary mask and region overlay. Bottom row displays a synthetic colonoscopy image with its binary mask and region overlay.

The analyses of the five segmentation models showed that the metrics had varying effects on each model's performance. With the highest IoU (0.65), UNet showed strong performance in many metrics, particularly achieving the highest recall (0.85) and Dice coefficient (0.78). FPN demonstrated solid results with the highest precision (0.77) and PSNR (7.21). LinkNet achieved the highest F1 score (0.78) among all models, although its other metrics were comparable to its peers. PSPNet showed slightly lower overall performance, with the lowest IoU (0.63) and PSNR (6.64). MANet's performance was middle of the pack across most metrics, with values generally similar to FPN and LinkNet. Regarding structural similarity, all models showed relatively low SSIM scores ranging from 0.44 to 0.49, with UNet and FPN tied for the highest at 0.49. The PSNR values were consistently in the range of 6–7 in all models, with the FPN achieving the highest at 7.21.

Figures 10 and 11 show the segmentation findings for real and synthentic colonoscopy image polyp images.

A pink growth is shown against a crimson intestinal lining in the original photograph. The real mask displays the polyp area as white. The segmentation of the UNet is shown in black on the expected mask. Green indicates a valid identification, red indicates a missing area, and yellow indicates a false positive. The overlay contrasts the results. UNet accurately depicts the polyp's form and position but overestimates its size, especially around the edges. UNet's multiscale feature recognition and over-segmentation are shown

here. Perhaps fine-tuning or postprocessing could sharpen edges. The results of this study are shown in Table 3. Table 3 shows the Dice coefficients. The five segmentation algorithms were thoroughly evaluated using the synthetic colonoscopy dataset, yielding the following results: UNet demonstrated better overall performance, achieving the highest scores in IoU (0.65), F1 score (0.76), and Dice coefficient (0.78), as seen in Table 3.

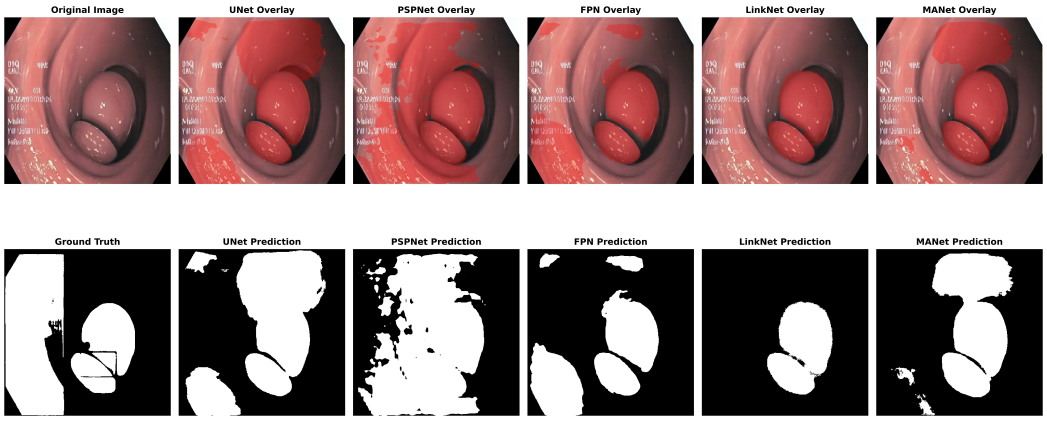

Figure 10. Comparison of polyp segmentation performance across different deep learning architectures. Top row shows overlay visualizations on the original (synthetic) colonoscopy image, while bottom row presents binary mask predictions. From left to right: original image, UNet, PSPNet, FPN, LinkNet, and MANet segmentation results, with ground truth mask for reference. The overlays (red regions) and binary predictions demonstrate varying segmentation accuracies and boundary detection capabilities across the different network architectures.

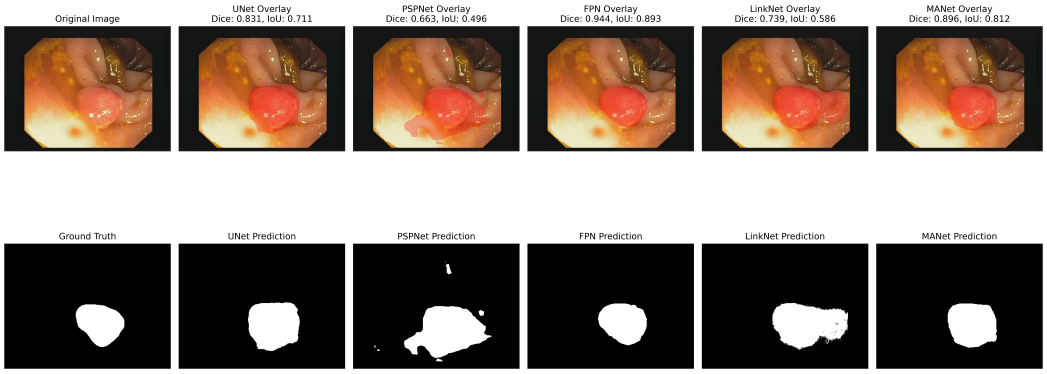

Figure 11. Comparison of polyp segmentation performance across different deep learning architectures. Top row shows the original (real) endoscopic image with model overlay predictions in red (UNet, PSPNet, FPN, LinkNet, and MANet). Bottom row displays the corresponding binary segmentation masks, with ground truth (leftmost) and model predictions, demonstrating the relative accuracy of each architecture in polyp region identification and boundary delineation.

Table 3. Comprehensive performance evaluation of segmentation models with results demonstrating varying capabilities across different architectures with UNet achieving superior performance in most metrics.

Model	IoU	F1 Score	Precision	Recall	PSNR	SSIM	Dice Coef.
UNet [50]	0.65	0.76	0.74	0.85	7.05	0.49	0.78
PSPNet [51]	0.63	0.76	0.71	0.84	6.64	0.44	0.77
FPN [52]	0.64	0.77	0.77	0.78	7.21	0.49	0.77
LinkNet [53]	0.64	0.78	0.73	0.83	7.01	0.47	0.77
MANet [54]	0.64	0.77	0.75	0.80	7.07	0.48	0.77

The models demonstrate consistently strong performance with minimal variation across metrics. UNet achieves a notably higher IoU of 0.65 compared to 0.63–0.64 for other models, indicating superior segmentation accuracy. The Dice coefficients range from 0.77 to 0.78, indicating an excellent overlap between predicted and ground truth segmentations. LinkNet shows the highest F1 score at 0.78, while FPN achieves the best PSNR at 7.21. SSIM values range from 0.44 to 0.49, with UNet and FPN tied for the highest structural similarity. Precision varies from 0.71 to 0.77, and recall ranges from 0.78 to 0.85, demonstrating a good balance between accuracy and completeness across all architectures. The result demonstrates a significant improvement over the results from [55].

4.4. Image Classification Results

The dataset used for classification consists of three sources: the original image dataset contains 1184 test images, 4138 training images, and 592 validation images; the synthetic image dataset has 358 test images, 1260 training images, and 182 validation images; and the augmented dataset comprises 1542 test images, 5398 training images, and 774 validation images. Advanced classification models, including ViT, BiT, FixResNeXt, DeiT, and EfficientNet, were trained using images generated from this investigation. Conventional metrics were used to evaluate these models. This comprehensive strategy aims to improve medical diagnosis and imaging technology. Figure 12 illustrates a 5 × 2 grid of original colonoscopy data randomly selected. Figure 13 illustrates the results of the models based on the original data.

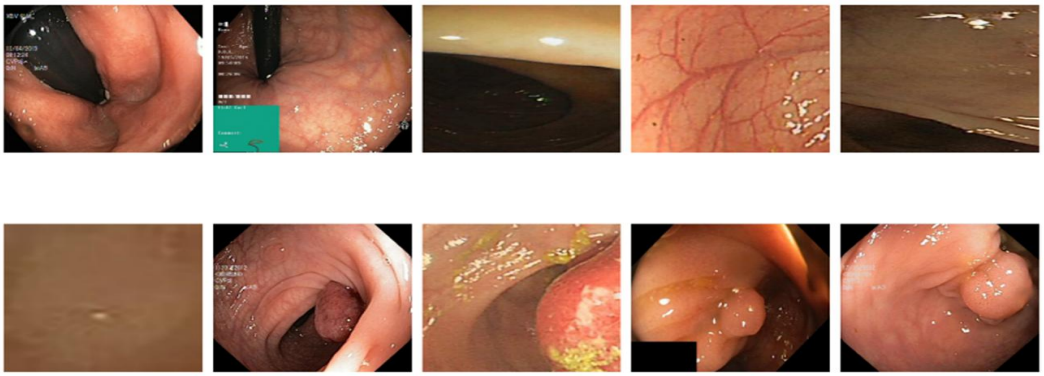

Figure 12. Endoscopic views from a colonoscopy procedure showing various segments and mucosal patterns of the colon interior.

EfficientNet has a pronounced superiority in all performance criteria used. Each metric confirmed that EfficientNet has one of the highest validation accuracy values of 97% and an F1 measures of 96.79% and is superior to all the models used. FixResNeXt was slightly less successful with a maximum validation accuracy of 86% and an F1 score of 85.95%. ViT's results are the second best of all networks, although they are far less impressive than EfficientNet's 96.79% AUC ratings.

Loss measurements shed more light on this discrepancy in performance: FixResNeXt has a validation loss of 0.36, while the remaining models have even higher levels of loss. The validation loss of EfficientNet is 0.11, making it the best performing model. Precision and recall also retain the same traits and are presented as follows. Validation precision and recall of EfficientNet: 96.8%, 96.8%; validation of the second best model, FixResNeXt: 86.2%, 85.9%. Specifically, most of the models demonstrate low discrepancies between the training and validation performance. EfficientNet demonstrates better adaptability to the data of this set, as well as higher efficiency in terms of all scores. The large accuracy margin points to the fact that the EfficientNet architecture is learning features and patterns in the

data that other architectures are unable to, thereby making EfficientNet suitable for tasks such as the ones used in this initial dataset.

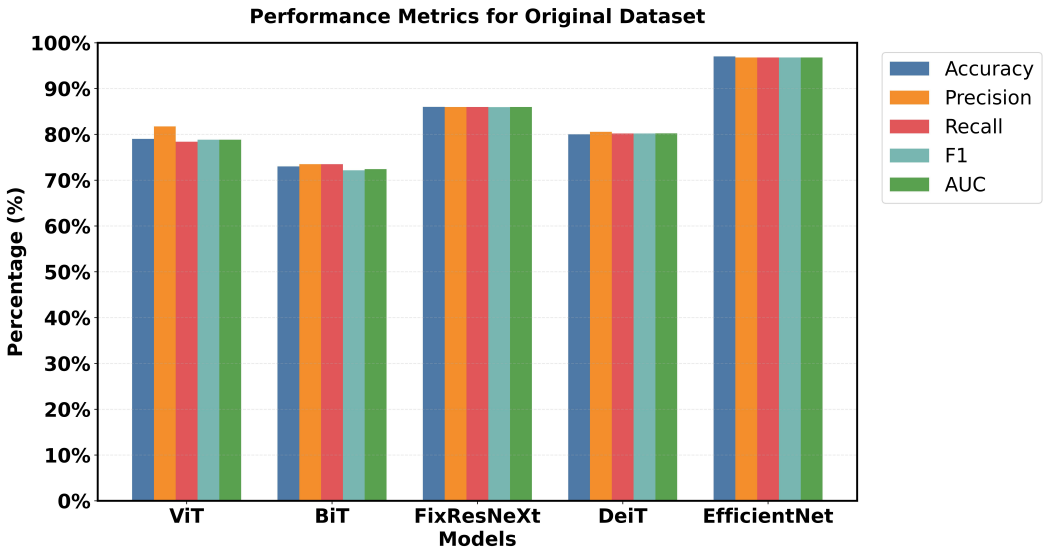

Figure 13. Performance comparison of deep learning and transformer-based models across multiple evaluation metrics, showing EfficientNet achieving superior performance across all measures on the original image dataset.

Figure 14 shows a 5 × 2 grid of randomly sampled synthetic colonoscopy data. Looking at the metrics in Figure 15, EfficientNet is smoother in almost all metrics, although not as dominant as it was with the initial data.

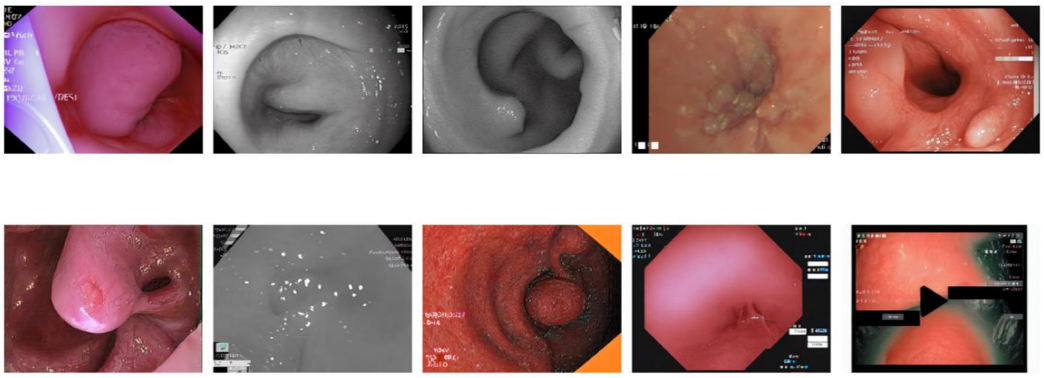

Figure 14. Synthetic endoscopic images generated to simulate various views and conditions encountered during medical endoscopy procedures.

The percentage accuracy of the training was 80 %, the accuracy of the validation was 79 %, and the F1 scores for the training and validation were 79.9% and 79.1%, respectively, which are comparatively lower than the scores obtained on the original data. Other models demonstrate better relative performance, specifically based on the metrics from the validation sets. FixResNeXt takes second place with the best validation accuracy of 71%, and an F1 score of 70.8%. BiT achieved a fantastic validation precision of 75.64% and ViT achieved a validation precision of 75.47%, which is higher than the training precisions of 64.98% and 61.46%, respectively.

Loss metrics further illustrate the narrowed gap: FixResNeXt's 0.58 loss and other models' losses are much closer to the validation loss of 0.5133, which belongs to EfficientNet.

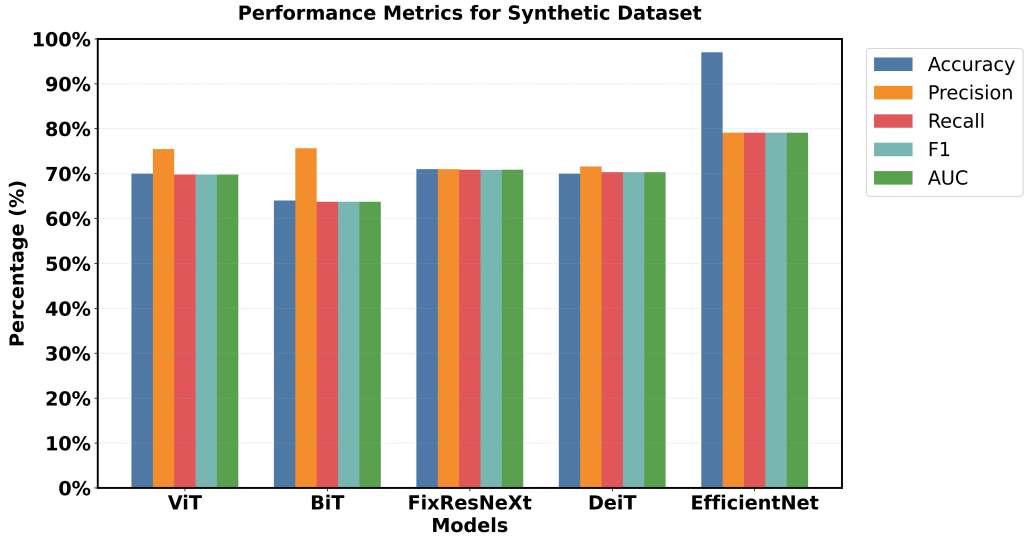

Figure 15. Performance evaluation of five deep learning and transformer-based models on AI-generated colonoscopy data, demonstrating generally lower but more consistent performance across models compared to original data, with EfficientNet maintaining superior accuracy but showing more metric variation.

An evaluation of overall performance highlights the fact that, with most models obtaining percentage scores within the range of 60–70% for the different evaluation criteria, synthetic data seem more difficult. This results from heightened complexity, variety, and noise in the synthetic data production process. The results indicate that the performance discrepancies between the models and the synthetic data stem from possible deficiencies not present in the actual data. Synthetic data may effectively enhance the evaluation of a model's robustness and its ability to generalize to novel occurrences.

Figure 16 shows a 5 × 5 grid of randomly sampled augmented colonoscopy data (combination of original and synthetic data). The augmented dataset causes a significant reversal in the performance of the various models and shows the variation in the data characteristics that affect the model performance. In this case, DeiT and ViT are the best that demonstrate high accuracy and F1 score, as well as the lowest error rate. Training precision of 98% was achieved, with a validation accuracy of 93%, while F1 scores achieved 97.94% in training and 92.89% in validation. The DeiT has shown the following performance, which shows a great improvement compared to the previous models. ViT achieved 93% validation accuracy and a 92.51% F1 score; in contrast, BiT achieved 92% validation accuracy with a 92.38% F1 score. This shows that the augmented dataset is well suited for transformer-based architectures.

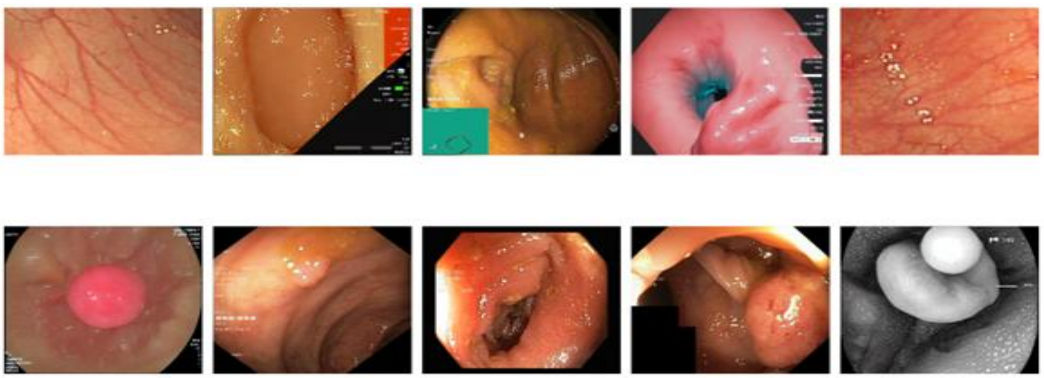

Figure 16. Combined dataset of real and synthetically generated endoscopic images, demonstrating the visual similarity between actual endoscopic findings and artificially generated medical imaging data.

In contrast, EfficientNet outperformed all other models in the previous datasets but has the lowest validation in this case, which is a precision score 86%. This change in the improvement rate is stark and underlines the importance of specific dataset attributes for model effectiveness. Loss metrics also point to these performance changes. Figure 17 shows that DeiT validation loss is 0.20, which is less than Efficient Net's 0.38 and the opposite of what has been seen in previous datasets. Relative performance can also be seen in the AUC scores, where DeiT achieved 92.90%, while EfficientNet achieved 86.05%.

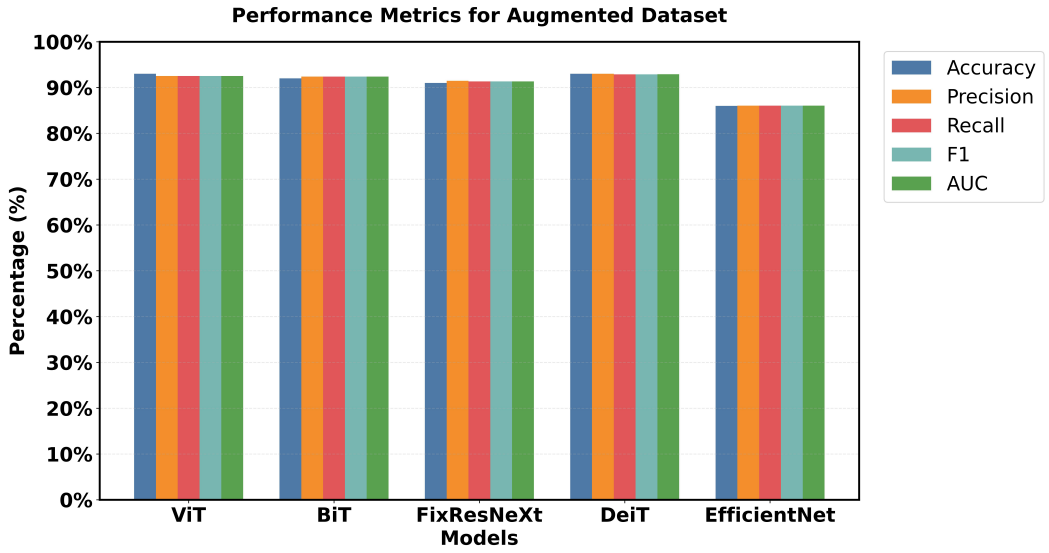

Figure 17. Performance evaluation of deep learning models on an augmented endoscopic dataset, showing substantially improved and more uniform performance across all models and metrics compared to both original and synthetic datasets, with most models achieving approximately 90% across all evaluation measures.

One unique feature of the intelligent data results is that most models had validation scores greater than 90% (EfficientNet not included). Training and validation performance are also closely matched for the top models with relatively little gap. For example, the training accuracy of 98% compared to the validation accuracy of 93% shows that DeiT and ViT have great generalization as seen in Figure 17.

These results indicate that the intelligent dataset contains distinguishable patterns, which makes transformer-based models highly effective. The high overall scores, as well as the proximity to the training validation performance, indicate that this dataset is more 'teachable' in gross terms, but with certain attributes that tend to benefit a given architectural design. This scenario shows that the structure of the model should correspond to the type of data and the qualitative difference that occurs in the work of the model in different sets.

This research captures major variations in the performance of the various models. Looking at the performance of the models in the original dataset, EfficientNet has the highest accuracy score and the highest F1 score (97% accuracy, 96.79% F1). FixResNeXt comes next in a short time, followed by other models that range from at most 75–80% on most of the metrics.

In the case of the synthetic dataset, overall model performances were comparatively lower than those achieved on the original dataset. EfficientNet retains a sample advantage over FixRexNet, although the score boundaries are significantly lower (79% validation accuracy). The performance difference decreases with FixResNeXt and VIT as close competitors, but there are differences in the validation precision: 71% for FixResNeXt and 75.47% for VIT, as seen in Figures 18 and 19.

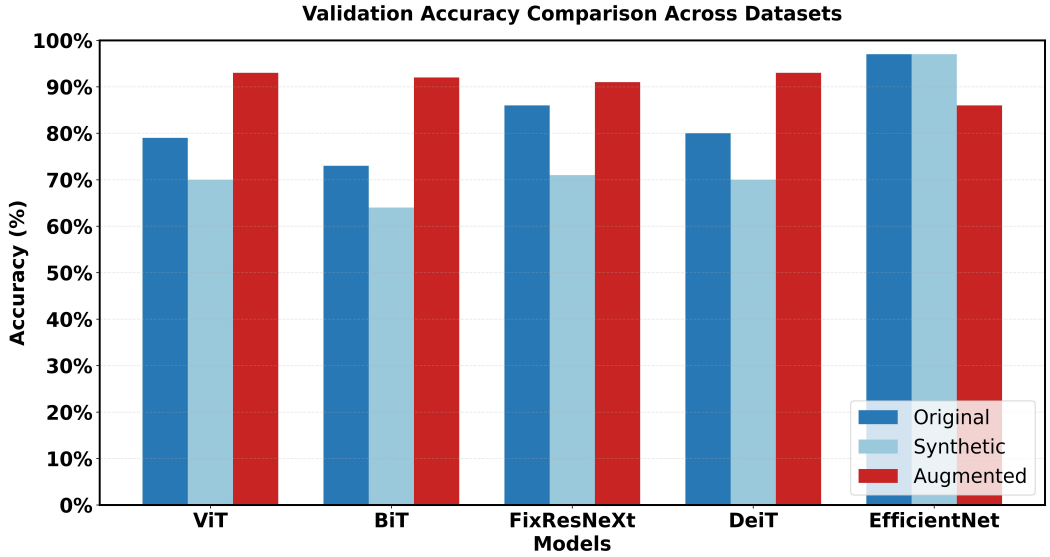

Figure 18. Validated accuracy comparison across the 3 datasets.

The augmented dataset shows a complete transition. The result shows that ViT and DeiT are the most accurate models with a validation accuracy of 93% and having the highest f1 score of 92.89%.

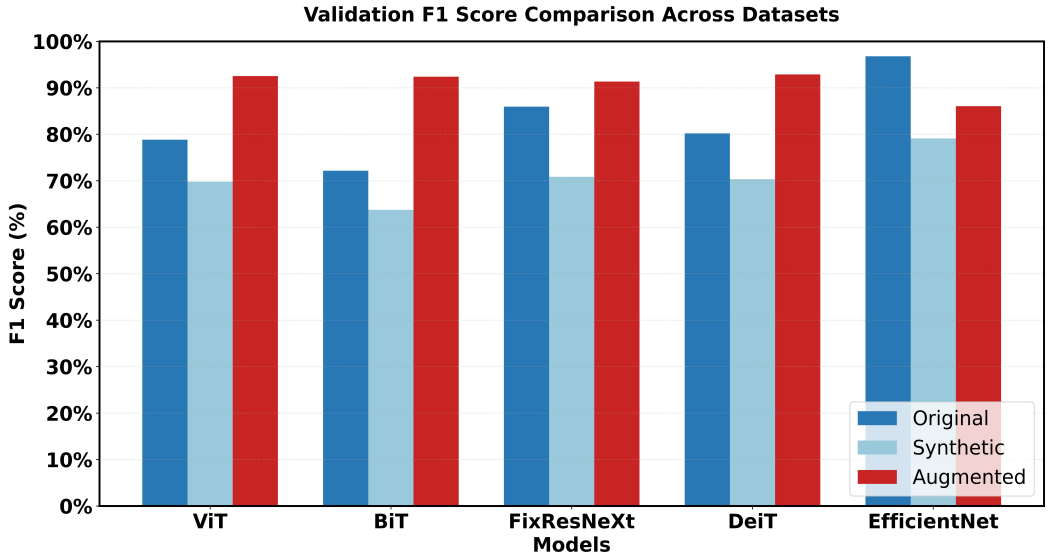

Figure 19. Validated F1 Score comparison across the 3 datasets.

The other two architectures, FixResNeXt and BiT, follow slightly behind ViT and DeiT. In contrast, here the validation metrics look worse for the accuracy of EfficientNet (86%). In particular, most customer variants demonstrate improved accuracy in the augmented dataset compared to the synthetic and original datasets, thus containing more learnable patterns. The Vision Transformer (ViT) demonstrates remarkable proficiency in polyp detection in both real and synthetic colonoscopy images, as shown in the figure. The model, trained on an augmented dataset that combined real colonoscopy images and synthetic data generated through fine-tuned Stable Diffusion, processes the input images by dividing them into fixed-size patches that are linearly embedded and processed through transformer encoder blocks. This training approach enables the ViT to learn robust features common to both real and synthetic polyps, preventing overfitting to specific real-world image characteristics. The model's effectiveness is evident in its high-confidence predictions: accurately identifying a subtle polyp in the real image with 0.95 confidence and a more

prominent polyp in the synthetic image with 0.99 confidence. The ViT generates appropriate bounding boxes in both cases, a smaller and precise box for the real polyp and a larger box that encompasses the full structure of the synthetic polyp, validating the effectiveness of synthetic data enhancement in medical image analysis, as seen in Figure 20.

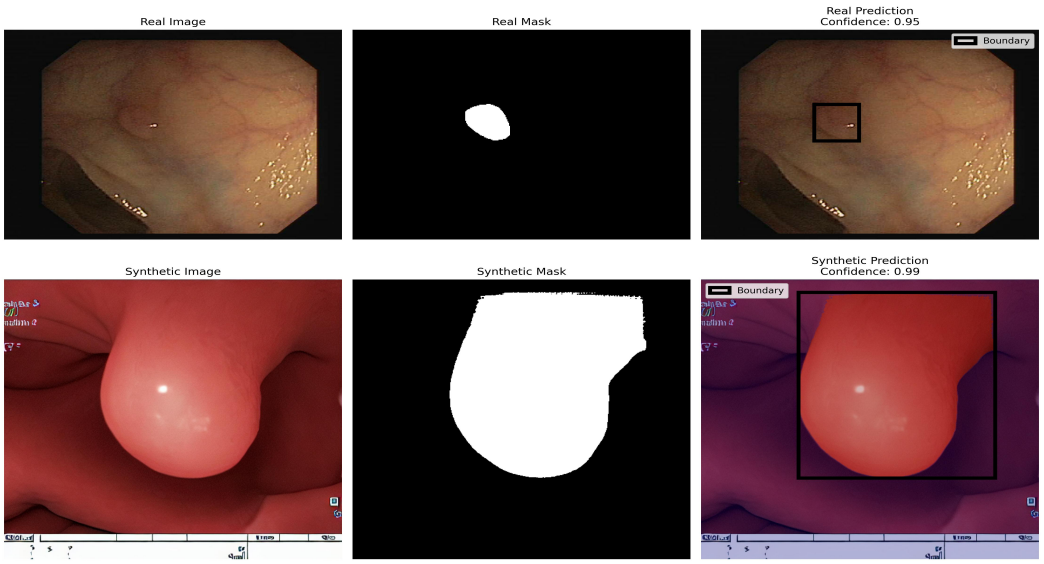

Figure 20. Visual comparison of ViT-based polyp detection performance on real and synthetic colonoscopy images, demonstrating high confidence predictions (0.95 and 0.99, respectively) with accurate boundary box localization.

The difference between training and validation scores, which affect a model's capacity to generalize, is the smallest for the augmented dataset. By comparing the outcomes, the results show that EfficientNet, one of the most variable networks, performs worst in intelligent dataset but better in the original dataset. Conversely, in the augmented dataset, the transformer-based structure (DeiT, ViT) performs much better overall. Intelligent Data Augmentation produced the best overall results for the following reasons:

1. **High Performance**
   The augmented dataset demonstrated exceptional performance, consistently achieving accuracy metrics above 92% in different models and evaluation metrics. This indicates robust and reliable model behavior across various architectural implementations.
2. **Performance Consistency**
   A notable pattern emerged in the comparative analysis. The original dataset showed inconsistent performance, with only one model achieving high accuracy (EfficientNet with 97% validation accuracy). The synthetic dataset consistently showed lower performance across all models. In contrast, the augmented dataset maintained consistently high performance across multiple architectures, indicating improved data quality and representation.
3. **Strong Generalization**
   The minimal gap between training and validation accuracy in the augmented dataset (typically within 2-3 percentage points) indicates effective knowledge transfer, reduced overfitting, and robust model generalization capabilities.
4. **Architecture-Agnostic Performance**
   Unlike the original and synthetic datasets, which showed significant performance variations between different architectures, the augmented dataset demonstrated more balanced performance. This indicates reduced architecture-specific bias in the learning process.

5. **Enhanced Model Applicability**
   The consistent high performance across diverse architectural approaches indicates that the knowledge extracted from the augmented dataset is more universally applicable. The learned features are more generalizable across different model architectures, and the dataset provides robust training signals for various deep learning approaches.

Although the score accrued from the initial dataset was the maximum (97% for EfficientNet), intelligent data offered the highest precision of performance with model consistency and generic precision. This makes it the most valuable from a machine learning perspective because in addition to offering a clear separation for analyzing the data, it can also inform the best choice for model selection and could translate to superior accuracy in real-world applications.

5. Discussion

This study created a novel strategy to address the lack of medical imaging data, especially when it comes to colonoscopy operations for underrepresented groups. The main objective of the study was to improve the detection models of CRC by combining various data augmentation approaches with sophisticated machine learning and deep learning techniques to synthesize colonoscopy images. To produce realistic artificial colonoscopy images, the researchers modified and improved several visual generative AI models, such as CLIP, Stable Diffusion (SD), and DreamBooth (DB) with LoRA. The Fréchet Inception Distance and Inception score were used to thoroughly assess the quality of these produced images. The research used several models for image classification tasks, including EfficientNet, FixResNeXt, Big Transfer, Vision Transformer, and Data-efficient Image Transformers. The study used original, augmented, and synthetic datasets for testing and training. And also for image segmentation tasks, the study implemented U-Net, PSNet, FPN, LinkNet, and MANet. With lower FID scores across all datasets, the results showed that DreamBooth in conjunction with LoRA created the most realistic photos. ViT and DeiT models achieved 93% validation precision and 92.89% F1 scores in classification tasks, according to the expanded dataset. This showed that creating synthetic images from high-quality data can improve model performance and generalizability across various demographic subgroups.

The CLIP, SD, and DB LoRa models were all successfully used for colonoscopy image generation in the study. Fine-tuned SD and DB LoRa produced the lowest FID scores for the datasets in terms of image originality and quality. The adapted models achieved optimal results when generating several images of synthetic colonoscopy. Inception scores were used to prove the diversity and realism of the generated image. DB LoRa received the highest average IS value of 2.36 for all datasets, indicating good image variety and quality. The quality and clinical usefulness of the images generated have therefore been quantitatively assessed using FID and IS metrics. The FID scores remained below 15 while the IS values were above 2 in all different datasets, which is an indication of the high quality and diversification of the images synthesized. The researchers were able to train and test the ViT, BiT, FixResNeXt, DeiT, and EfficientNet models using synthetic, original, and augmented data. The augmented data strengthened the results in addition to obtaining a higher validation precision with ViT and DeiT both having a precision of 93% and an F1 score of 92.89%, which means that synthetic data enhancement is effective. For the image segmentation task, the study adopted and assessed the performance of U-Net, PSNet, FPN, LinkNet, and MANet. FPN produced excellent results, with an IoU of 0.64, an F1 score of 0.78, a recall of 0.75, and a Dice coefficient of 0.77.

This study clearly showed both improved performance and fairness of the model with augmented synthetic data. The augmented dataset performed steadily better than

the original and synthetic datasets for all models and measures, indicating lower bias and better generalization. All research objectives were met systematically, indicating general research contributions in the areas of synthetic image generation, improved model performance, and possible elimination of bias when using AI-enabled analysis of colonoscopy images. By improving AI-based colonoscopy screening with improved training datasets, this study significantly reduces health disparities in CRC therapy outcomes. In addition to demonstrating how artificial data augmentation can improve the effectiveness and equity of AI models for polyp recognition and classification, this study advances the field of generative AI in medical imaging. The innovative approach to medical picture synthesis, extensive evaluation techniques, improved model performance, and important implications for healthcare equity are just a few of its strengths. This study extensively evaluated a number of state-of-the-art models in a variety of fields and showed increased validation accuracy using better datasets. However, this study also has drawbacks, especially when it comes to a thorough clinical validation and moral issues around the use of artificial data in healthcare settings, which need further investigation before practical implementation.

6. Conclusions

In conclusion, this work has shown that generative AI, text-to-image synthesis, and intelligent data augmentation can be applied to overcome limitations arising from a limited number of colonoscopy images and data bias. This study demonstrated a novel way to generate a diverse and realistic number of synthetic colonoscopy images that could reflect a rather wide range of patients by improving state-of-the-art models such as CLIP, DB LoRa and SD. The comparative analysis of these synthetic images computed via FID and IS metrics, as well as the outcome of classification and segmentation models trained to utilize augmented datasets, prove the critical enhancements in AI model effectiveness and non-bias. The proposed augmentation method using original and synthetic images showed better performance than original and purely synthetic datasets for several models and evaluation metrics. This study helps address the problem of minority representation in both the identification and treatment of CRC by improving the training datasets of AI-supported colonoscopy technologies. It also pushes the point of generative AI in analyzing colonoscopy and medical imaging at large, moving with more than colonoscopy applications. However, some limitations must be investigated in subsequent studies, for example, the requirement for clinical evaluation of synthesized images and some ethical concerns about incorporating AI-generated medical data. However, this work sets a robust framework upon which future studies can build to enhance the efficiency of applying artificial intelligence to the analysis of medical images and to overcome disparities affecting health in different countries. This study recommends the use of augmented datasets in AI model training for colonoscopy assistance tools, as synthetic image generation has produced promising results. Augmented data sets improve developer resources, model accuracy, and generalization while also reducing discrimination caused by the underrepresentation of racial minorities in medical imaging data. A notable limitation of this study is the absence of human evaluation through blinded evaluation. Future work would benefit from having experienced medical professionals perform blinded comparisons between real and synthetic colonoscopy images. This expert evaluation would provide valuable information on the clinical viability of the generated images and help validate whether the synthetic data could effectively supplement medical training datasets. Clinical validation studies in real world settings are critical to determining the effectiveness of these methods and addressing potential risks before deployment. Future research should focus on large-scale clinical validation of AI models trained on augmented datasets, the development of ethical guidelines for the use of synthetic medical images in clinical and training settings, and the

application of synthetic data augmentation techniques to other medical imaging modalities besides colonoscopy. General implementation requires careful consideration of practical challenges, approval processes, and integration into existing ones.

Author Contributions: Conceptualization, O.O.E.P., O.T.A. and M.M.R.; methodology, M.M.R., F.K., O.O.E.P. and A.M.J.-O.; software, O.O.E.P.; validation, O.O.E.P., F.K. and M.M.R.; formal analysis, O.O.E.P.; investigation, O.O.E.P.; resources, F.K. and M.M.R.; data curation, O.O.E.P. and M.M.R.; writing—original draft preparation, F.K., M.M.R. and O.O.E.P.; writing—review and editing, O.T.A., A.M.J.-O., F.K. and M.M.R.; visualization, O.O.E.P. and O.T.A.; supervision, F.K. and M.M.R.; project administration, F.K. and M.M.R.; funding acquisition, F.K. and M.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Science Foundation (NSF) under Grant No. 2131307, "CISE-MSI: DP: IIS: III: Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support System", and in part by the Office of the Director, National Institutes of Health (NIH) Common Fund under Award No. 1OT2OD032581-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of AIM-AHEAD, the NIH, or any other funding agencies.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: This work used publicly available data.

Data Availability Statement: The data presented in this study are openly available. The source code and implementation details can be found at https://github.com/Ejigsonpeter/Text-Guided-Synthesis-for-Colon-Cancer-Screening/tree/main (accessed on 20 January 2025). The datasets used in this study are accessible through secure cloud storage. The complete dataset collection is available at https://drive.google.com/drive/folders/12WrV0W1ULOrhfWy8TxAwLHDfqg9iiIna?usp=sharing (accessed on 20 October 2024). The synthetic dataset, original dataset, augmented dataset, and CLEF dataset can be accessed through their respective repositories [https://drive.google.com/file/d/1PXiYVGK6Mv3cLe0xNbf1e5VsFdUbuTq8/view?usp=sharing, https://drive.google.com/file/d/1eqdnWFhXp1afbj62JNc7QGkEkNzVAolF/view?usp=sharing, https://drive.google.com/file/d/1TzA-_La1vub8TM8PHnC5uUDF33wZNw0J/view?usp=sharing, https://drive.google.com/file/d/1_6coUaFtFmMDSITcDiviaESl-5rESfDm/view?usp=sharing (accessed on 20 November 2024)].

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AIGC	AI-Generated Content
ArSDM	Arbitrary-Style Diffusion Model
AUC	Area Under the Curve
BiT	Big Transfer
BLIP	Bootstrapping Language-Image Pre-training
CADe	Computer-Aided Detection
CFM	Cross Fusion Module
cGAN	Conditional Generative Adversarial Network
CIM	Cross Interaction Module
CLIP	Contrastive Language-Image Pre-Training
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
DB	DreamBooth

DeiT	Data-Efficient Image Transformers
DL	Deep Learning
DR	Adenoma Detecting Rate
FCNN	Fully Convolutional Neural Network
FID	Fréchet Inception Distance
FPN	Feature Pyramid Network
GANs	Generative Adversarial Networks
HD	High Definition
IS	Inception Score
LDM	Latent Diffusion Model
LinkNet	Link Network
LoRA	Low-Rank Adaptation
MANet	Multi-Scale Attention Network
Mask R-CNN	Mask Region-based Convolutional Neural Network
mDice	Mean Dice Coefficient
mIoU	Mean Intersection over Union
ML	Machine Learning
Polyp-PVT	Polyp Pyramid Vision Transformer
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSPNet	Pyramid Scene Parsing Network
SAM	Spatial Attention Module
SD	Stable Diffusion
SSPP	Single Sample Per Person
U-Net	U-Shaped Network
VQGAN	Vector-Quantized Generative Adversarial Network

References

1. Wang, P.; Berzin, T.M.; Brown, J.R.G.; Bharadwaj, S.; Becq, A.; Xiao, X.; Liu, P.; Li, L.; Song, Y.; Zhang, D.; et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. *Gut* **2019**, *68*, 1813–1819. [CrossRef] [PubMed]
2. Bernal, J.; Tajkbaksh, N.; Sanchez, F.J.; Matuszewski, B.J.; Chen, H.; Yu, L.; Angermann, Q.; Romain, O.; Rustad, B.; Balasingham, I.; et al. Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Trans. Med. Imaging* **2017**, *36*, 1231–1249. [CrossRef] [PubMed]
3. Kim, J.J.H.; Um, R.S.; Lee, J.W.Y.; Ajilore, O. Generative AI can fabricate advanced scientific visualizations: Ethical implications and strategic mitigation framework. *AI Ethics* **2024**. [CrossRef]
4. Videau, M.; Knizev, N.; Leite, A.; Schoenauer, M.; Teytaud, O. Interactive Latent Diffusion Model. In *Proceedings of the Genetic and Evolutionary Computation Conference*; ACM: New York, NY, USA, 2023; pp. 586–596. [CrossRef]
5. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef]
6. Alhabeeb, S.K.; Al-Shargabi, A.A. Text-to-Image Synthesis with Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction. *IEEE Access* **2024**, *12*, 24412–24427. [CrossRef]
7. Tan, Y.X.; Lee, C.P.; Mai, N.; Lim, K.M.; Lim, J.Y.; Alqahtani, A. Recent Advances in Text-to-Image Synthesis: Approaches, Datasets and Future Research Prospects. *IEEE Access* **2023**, *11*, 88099–88115. [CrossRef]
8. Iglesias, G.; Talavera, E.; Díaz-Álvarez, A. A survey on GANs for computer vision: Recent research, analysis and taxonomy. *Comput. Sci. Rev.* **2023**, *48*, 100553. [CrossRef]
9. Ejiga Peter, O.O.; Rahman, M.M.; Khalifa, F. Advancing AI-Powered Medical Image Synthesis: Insights from MedVQA-GI Challenge Using CLIP, Fine-Tuned Stable Diffusion, and Dream-Booth + LoRA. Conference and Labs of the Evaluation Forum. 2024. Available online: https://ceur-ws.org/Vol-3740/paper-145.pdf (accessed on 12 December 2024).
10. Najjar, R. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics* **2023**, *13*, 2760. [CrossRef]
11. Alnaggar, O.A.M.F.; Jagadale, B.N.; Saif, M.A.N.; Ghaleb, O.A.; Ahmed, A.A.; Aqlan, H.A.A.; Al-Ariki, H.D.E. Efficient artificial intelligence approaches for medical image processing in healthcare: Comprehensive review, taxonomy, and analysis. *Artif. Intell. Rev.* **2024**, *57*, 221. [CrossRef]

12. Arora, A.; Alderman, J.E.; Palmer, J.; Ganapathi, S.; Laws, E.; Mccradden, M.D.; Oakden-Rayner, L.; Pfohl, S.R.; Ghassemi, M.; McKay, F.; et al. The value of standards for health datasets in artificial intelligence-based applications. *Nat. Med.* **2023**, *29*, 2929–2938. [CrossRef]
13. Han, P.; Ye, C.; Zhou, J.; Zhang, J.; Hong, J.; Li, X. Latent-based Diffusion Model for Long-tailed Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 17–18 June 2024.
14. Du, Y.; Jiang, Y.; Tan, S.; Wu, X.; Dou, Q.; Li, Z.; Li, G.; Wan, X. ArSDM: Colonoscopy Images Synthesis with Adaptive Refinement Semantic Diffusion Models. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2023, Vancouver, BC, Canada, 8–12 October 2023. [CrossRef]
15. Ku, H.; Lee, M. TextControlGAN: Text-to-Image Synthesis with Controllable Generative Adversarial Networks. *Appl. Sci.* **2023**, *13*, 5098. [CrossRef]
16. Iqbal, M.A.; Jadoon, W.; Kim, S.K. Synthetic Image Generation Using Conditional GAN-Provided Single-Sample Face Image. *Appl. Sci.* **2024**, *14*, 5049. [CrossRef]
17. Shin, Y.; Qadir, H.A.; Aabakken, L.; Bergsland, J.; Balasingham, I. Automatic Colon Polyp Detection using Region based Deep CNN and Post Learning Approaches. *IEEE Access* **2019**, *6*, 40950–40962. [CrossRef]
18. Qadir, H.A.; Shin, Y.; Solhusvik, J.; Bergsland, J.; Aabakken, L.; Balasingham, I. Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better? In Proceedings of the International Symposium on Medical Information and Communication Technology (ISMICT), Oslo, Norway, 8–10 May 2019; pp. 1–6. [CrossRef]
19. Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *arXiv* **2021**, arXiv:2108.06932. [CrossRef]
20. Repici, A.; Badalamenti, M.; Maselli, R.; Correale, L.; Radaelli, F.; Rondonotti, E.; Ferrara, E.; Spadaccini, M.; Alkandari, A.; Fugazza, A.; et al. Efficacy of Real-Time Computer-Aided Detection of Colorectal Neoplasia in a Randomized Trial. *Gastroenterology* **2020**, *159*, 512–520.e7. [CrossRef]
21. Kudo, S.E.; Misawa, M.; Mori, Y.; Hotta, K.; Ohtsuka, K.; Ikematsu, H.; Saito, Y.; Takeda, K.; Nakamura, H.; Ichimasa, K.; et al. Artificial Intelligence-assisted System Improves Endoscopic Identification of Colorectal Neoplasms. *Clin. Gastroenterol. Hepatol.* **2020**, *18*, 1874–1881.e2. [CrossRef]
22. Zhou, J.; Wu, L.; Wan, X.; Shen, L.; Liu, J.; Zhang, J.; Jiang, X.; Wang, Z.; Yu, S.; Kang, J.; et al. A novel artificial intelligence system for the assessment of bowel preparation (with video). *Gastrointest Endosc* **2020**, *91*, 428–435.e2. [CrossRef]
23. Mahmood, F.; Chen, R.; Durr, N.J. Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training. *IEEE Trans. Med. Imaging* **2018**, *37*, 2572–2581. [CrossRef] [PubMed]
24. Goceri, E. Medical image data augmentation: Techniques, comparisons and interpretations. *Artif. Intell. Rev.* **2023**, *56*, 12561–12605. [CrossRef]
25. Yang, Z.; Zhan, F.; Liu, K.; Xu, M.; Lu, S. AI-Generated Images as Data Source: The Dawn of Synthetic Era. *arXiv* **2023**. [CrossRef]
26. Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P.S.; Sun, L. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *arXiv* **2023**. [CrossRef]
27. Bandi, A.; Adapa, P.V.S.R.; Kuchi, Y.E.V.P.K. The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet* **2023**, *15*, 260. [CrossRef]
28. Bendel, O. Image synthesis from an ethical perspective. *AI Soc.* **2023**. [CrossRef]
29. Derevyanko, N.; Zalevska, O. Comparative analysis of neural networks Midjourney, Stable Diffusion, and DALL-E and ways of their implementation in the educational process of students of design specialities. *Pedagog. Psychol.* **2023**, *9*, 36–44. [CrossRef]
30. Sánchez-Peralta, L.F.; Bote-Curiel, L.; Picón, A.; Sánchez-Margallo, F.M.; Pagador, J.B. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artif. Intell. Med.* **2020**, *108*, 101923. [CrossRef] [PubMed]
31. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016.
32. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
33. Wang, P.; Xiao, X.; Glissen Brown, J.R.; Berzin, T.M.; Tu, M.; Xiong, F.; Hu, X.; Liu, P.; Song, Y.; Zhang, D.; et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat. Biomed. Eng.* **2018**, *2*, 741–748. [CrossRef]
34. Misawa, M.; Kudo, S.E.; Mori, Y.; Cho, T.; Kataoka, S.; Yamauchi, A.; Ogawa, Y.; Maeda, Y.; Takeda, K.; Ichimasa, K.; et al. Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience. *Gastroenterology* **2018**, *154*, 2027–2029.e3. [CrossRef]
35. Guo, Y.; Bernal, J.; Matuszewski, B.J. Polyp Segmentation with Fully Convolutional Deep Neural Networks—Extended Evaluation Study. *J. Imaging* **2020**, *6*, 69. [CrossRef]

36. Borgli, H.; Thambawita, V.; Smedsrud, P.H.; Hicks, S.; Jha, D.; Eskeland, S.L.; Randel, K.R.; Pogorelov, K.; Lux, M.; Nguyen, D.T.D.; et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data* **2020**, *7*, 283. [CrossRef]
37. Beaumont, R. LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets. 2022. Available online: https://laion.ai/blog/laion-5b/ (accessed on 15 December 2024).
38. Hicks, S.; Storås, A.; Halvorsen, P.; De Lange, T.; Riegler, M.; Thambawita, V. Overview of ImageCLEFmedical 2023—Medical Visual Question Answering for Gastrointestinal Tract. 2023. Available online: https://ceur-ws.org/Vol-3497/paper-107.pdf (accessed on 15 December 2024).
39. Wang, W.; Tian, J. CP-CHILD Records the Colonoscopy Data. figshare 2020. Available online: https://figshare.com/articles/dataset/CP-CHILD_zip/12554042?file=23383508 (accessed on 15 December 2024).
40. Rahman, M.S. Binary Polyps Classification. 2024. Available online: https://www.kaggle.com/datasets/mdsahilurrahman71/binary-polyps-classification?resource=download (accessed on 15 December 2024).
41. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
42. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
43. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and A Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. [CrossRef]
44. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
45. Hore, A.; Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In Proceedings of the International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369. [CrossRef]
46. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
47. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef] [PubMed]
48. Ejiga, P.O.; Oluwafemi, O. Text-Guided Synthesis for Colon Cancer Screening. GitHub Repository. 2024. Available online: https://github.com/Ejigsonpeter/Text-Guided-Synthesis-for-Colon-Cancer-Screening (accessed on 15 December 2024).
49. HuggingFace. Mask Generation. 2024. Available online: https://huggingface.co/docs/transformers/tasks/mask_generation (accessed on 15 December 2024).
50. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *arXiv* **2015**. [CrossRef]
51. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *arXiv* **2017**. [CrossRef]
52. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2016**. [CrossRef]
53. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017. Available online: https://arxiv.org/abs/1707.03718 (accessed on 15 December 2024).
54. Safari, F.; Savić, I.; Kunze, H.; Ernst, J.; Gillis, D. A Review of AI-based MANET Routing Protocols. In Proceedings of the 2023 19th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Montreal, QC, Canada, 21–23 June 2023; pp. 43–50. [CrossRef]
55. Ejiga Peter, O.O. Advancing Colonoscopy Analysis Through Text-to-Image Synthesis Using Generative AI for Intelligent Data Augmentation, Image Classification, and Segmentation. 2024. Available online: https://www.proquest.com/openview/9a3add722e60af686957df5383de11f5/1?pq-origsite=gscholar&cbl=18750&diss=y (accessed on 8 January 2025).