Synthetic Image Generation For Enhancing Polyp Classification

Nerea Qing Muñoz Martin

Abstract

This project explores improving polyp classification in endoscopic images by augmenting datasets with synthetic images using diffusion models. Starting from a baseline classifier, a diffusion model trained from scratch was used, generating low-quality images. A pretrained Stable Diffusion model was then finetuned using LoRA. The motivation is to overcome the lack of data in medical imaging. Results indicate that synthetic data augmentation can improve classification performance, although there is still more research to do in this area. This work demonstrates the potential of advanced generative models to support medical AI applications.

Index Terms

Keywords: Polyp Classification, Endoscopic Images, Synthetic Image Generation, Diffusion Models, Classification Models, Stable Diffusion, LoRA, Dreambooth, Data Augmentation, Medical Imaging, Class Imbalance

I Introduction

In recent years, artificial intelligence (AI) has grown rapidly, with many useful applications already helping society and others still being tested. One important area where AI is making progress is healthcare, especially in detecting diseases through medical images. Although there have been many successes, challenges still remain.

This project focuses on improving the classification of polyps in medical images. A major problem in this area is the lack of data and the imbalance between classes. This limits the model's ability to learn effectively and often results in biased predictions.

The motivation behind this study is to develop a model that generates synthetic images of polyps. These images will help balance and augment the dataset, to enhance the performance of classification models, supporting faster and more reliable diagnosis and treatment.

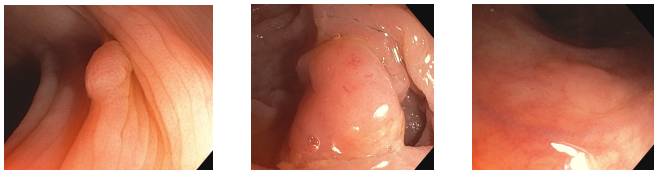
A. Clinical Context

A polyp is a small growth that forms in the inner lining of the colon or rectum. While not all polyps become cancerous, some can develop into colorectal cancer over time. This work focuses on three main types of polyps: adenomatous (AD), sessile serrated adenomatous (SSA), and hyperplastic (HP) polyps (see Figure 1). The clinical descriptions and statistics presented here are based on information from reliable health sources such as the University of Michigan Health website (1).

Adenomatous (AD) polyps are the most common, but only a small percentage of them become cancerous, and it can take many years. However, nearly all malignant polyps begin as adenomatous, which means this type is considered precancerous. When this type of polyp is found, it is tested for cancer.

Sessile serrated adenomatous (SSA) polyps have a sawtooth appearance when viewed under microscope. These are harder to detect and can potentially become cancerous depending on their size and location in the colon. Larger serrated polyps, which are typically flat, are precancerous.

Hyperplastic (HP) polyps, a type of serrated polyps, are small and considered extremely low risk for becoming cancerous.



(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 1: Examples of different polyp types: (a) Adenomatous, (b) SSA, and (c) Hyperplastic

The appearance of these growths is influenced by some factors such as age, genetic factors, inheritance, diet habits... Their detection involves carrying out a procedure called colonoscopy, which consists of inserting a long, thin instrument known as colonoscope into the anus to examine the colon and rectum, as explained in the website Bowel Cancer Australia (2). The colonoscope has a camera at one end that takes video images from inside the patient's body and transmits them to a computer screen, allowing doctors inspect the colon and looking for abnormalities. The images used for this study are taken from this tool during this procedure.

• Contact E-mail: nereaqing@gmail.com
• Supervised by: Jorge Bernal del Nozal (Computer Science Department)
• Academic Year 2024/25

If polyps are found during the procedure, they are usually removed immediately using the same colonoscope. However, in the case of small, low-risk polyps (5 mm or less), the *resect and discard* protocol may be applied, where they are removed and discarded without histological analysis if they can be confidently classified. Conversely, hyperplastic polyps, which pose no malignant risk, may be *left in situ*, meaning they are left in place, thus avoiding unnecessary interventions. If a polyp is too large to be safely removed, the doctor extracts a tissue sample for laboratory analysis to determine the nature of the lesion. If another type of abnormality is found, a biopsy is performed.

B. State of the Art

As part of this work, some research has been conducted to understand what has already been experimented with and which contributions have been made to the generation of synthetic images.

The article proposed by Voetman et al. (3) demonstrated that diffusion models show promising results in replicating images of apples from a real-world scenario by fine-tuning a Stable Diffusion network using DreamBooth. However, it showed some limitations when trying to mimic very large or small images of apples. Nevertheless, it was demonstrated that a pretrained diffusion model could be finetuned to generate a high quality representative dataset of a real-world scenario that could achieve nearly the same results when using an object detector, although not surpassing the results from real images.

Generative Adversarial Networks (GANs) were presented in a study by Guibas et al. (4) as another model used to reproduce synthetic images from real images. This architecture consists of a generator that tries to create synthetic images as close to the real world as possible, and a discriminator that attempts to distinguish which images are not real. A single GAN is unable to determine complex structures; it can only identify simple features such as general color, shape, and lighting. Hence, this lack of detail is unacceptable when referring to medical images, as they need to accurately represent real data.

The architecture proposed in the article was the use of two GANs, breaking the problem into two parts: the first GAN's objective is to produce segmentation masks that represent the variable geometries of the dataset, while the second one's goal is to translate these generated masks to photorealistic images. This approach has shown improvements in the quality of synthetic images.

Furthermore, in the specific domain of endoscopic imaging, Variational Autoencoders (VAEs) have also been explored as generative models. One of the first works to address a problem closely related to the current study is in a paper presented by De La Fuente et al. (5), where VAEs were used to generate synthetic images in order to improve classification performance. VAEs learn a probabilistic mapping from data to a latent space, meaning its latent space is a probability distribution that can be sampled to generate new samples similar to the real data.

The article trained three distinct VAEs on the three types of medical image data and could sample from the latent space. To generate new images that did not resemble the real ones, two samples were taken from the VAE and were averaged. This research demonstrated the possibility of slightly improving the accuracy of a classification model by using this type of architecture.

After conducting this research on state-of-the-art models, it is evident that more contributions need to be made to address the challenge of data scarcity and class imbalance.

C. Objectives

As mentioned in the previous section, some architectures have appeared to be a promising solution to the limited data and class imbalance in the healthcare field. However, results of classification models have not improved significantly.

For that reason, the goal of this study is to contribute to this research by creating a model that generates synthetic medical images to address the mentioned issues. The work is structured across several phases, as outlined in the Gantt Chart (see Appendix, Figure 36).

The initial phase involves an extensive review of related investigations and the development of a baseline classification model for later comparison. This is followed by an experimentation phase, where different generative models will be evaluated and adapted, either through fine-tuning, modifications of the architecture, or training from scratch, to generate synthetic images. These images will then be combined with real data to retrain the classification model.

A final evaluation phase will compare the model's performance against the baseline, with possible refinements based on the results. Throughout the project, biweekly meetings with the supervisor will ensure continuous progress monitoring and planning.

II Dataset

The dataset consists of polyp images provided by Hospital Clínic of Barcelona, captured during colonoscopy procedures, along with their corresponding binary masks, both in TIF format.

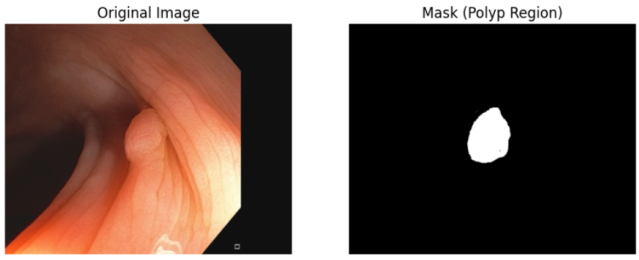


Fig. 2: Original endoscopic image with its corresponding polyp mask.

Figure 2 illustrates a sample pair: the original endoscopic image alongside the associated mask, which indicates the exact polyp region.

The images for the training, validation and test sets are divided into percentages of 70%, 10% and 20% respectively (Figure 3).

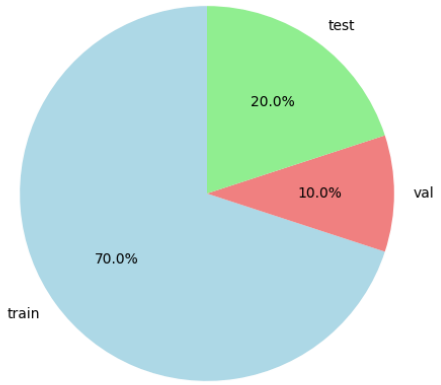


Fig. 3: Distribution proportions of training, validation, and test datasets.

As mentioned before, the focus is on three types of polyps: adenomatous (AD), sessile serrated (SSA), and hyperplastic (HP). Figure 4 illustrates the severe class imbalance present in the dataset, highlighting the need for generating new synthetic data to improve the results of classification models.

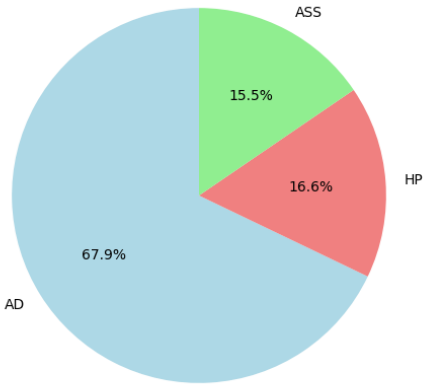


Fig. 4: Class distribution of different polyp types in the dataset.

This pie chart is based on the training set, but the validation set follows the same distribution proportion.

Apart from the images and masks, a CSV file has been provided with a structure of three columns, which can be seen in Table I.

TABLE I: Sample entries from the dataset CSV file showing image ID, histology label, and class abbreviation.

image id	Histologia	cls
0	1	HP
1	1	AD
2	0	ASS

After analyzing the dataset, a preprocessing step has been carried out in order to prepare the data to be fed to the models.

Since the dataset contains masks that indicate the exact region of the polyp, they have been applied to the original image to extract the polyp, as shown in Figure 5.

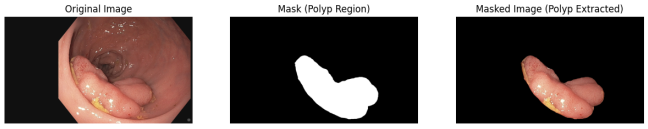


Fig. 5: Example of a polyp region extracted from the original image using the mask.

However, it was ultimately decided to train the models using the original full images rather than the extracted polyp regions. This approach is especially important for the generator model, which benefits from the additional contextual information surrounding the polyp to improve generation quality.

III Methodology

In this section, the different models and architectures implemented throughout the investigation are detailed. As discussed in section I, a classification model will be developed to serve as a baseline. This baseline is essential to assess the quality of the synthetic images by determining whether there has been an improvement or not.

A. Baseline Classifier

The model used as a baseline classifier was a pretrained EfficientNet-B0 architecture. This convolutional neural network (CNN), originally trained on ImageNet, offers a strong balance between performance and computational efficiency, as shown in the article introduced by Marques et al. (6).

For finetuning, the model's final layer was replaced with a custom classification head composed of:

- A **fully connected linear layer** that maps the backbone's output features to a lower-dimensional hidden space.
- A **ReLU** activation function.
- A **dropout** layer to reduce overfitting.
- A final **linear layer** that outputs logits corresponding to the number of polyp classes, which is 3.

Given the dataset's severe class imbalance, some strategies have been explored:

- **Weighted sampling:** during training, samples from minority classes will have a higher probability to be drawn.
- **Weighted loss function:** the loss function will be adjusted to penalize errors on minority classes more heavily.
- **No compensation:** models will be also trained without any explicit class imbalance handling for comparison.

Furthermore, two different input image sizes were used to train the model, 128×128 and 224×224, to explore how image resolution affects learning and model performance.

Finally, an additional setup was tested by combining SSA and HP polyps into a single class, allowing the model to distinguish between potentially malignant (AD) and non-malignant (SSA+HP) polyps.

B. Diffusion-Based Image Generator

The chosen generator is based on a diffusion model architecture, since they are one of the current state-of-the-art approaches that have been proven to be really useful for high-quality image generation.

As explained in the IBM website (7), diffusion models consist of gradually diffusing a data point with random noise, then learning to reverse that diffusion process to reconstruct the original data distribution (Figure 6). Directly transforming random noise into a coherent image is very complex, but transforming a noisy image into a slightly less noisy image is relatively straightforward. Hence, what diffusion models do is to reverse diffusion process as an incremental transformation of a simple distribution (like Gaussian noise) to a more complex distribution (like a coherent image).

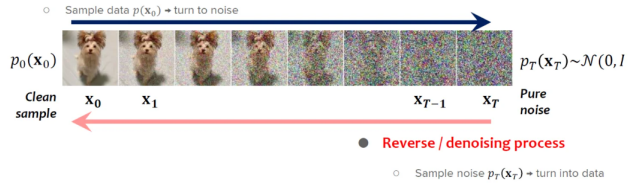


Fig. 6: Illustration of the diffusion process, showing the gradual addition of noise to data and the inverse process of denoising that reconstructs the original data distribution (8).

The diffusion process has three stages:

- **Forward diffusion process:** from training data to pure noise, usually a Gaussian distribution, step-by-step formulated as a Markov chain, in which the outcome at each timestep is influenced only by the timestep immediately preceding it. Thus, at each timestep, a small amount of Gaussian noise is added to the image.
- **Reverse diffusion process:** the model learns to inverse each previous step in the original forward diffusion process. This stage is the core of the diffusion model's learning, where it effectively learns how to generate realistic images by reconstructing data from noisy inputs.
- **Image generation:** the trained model samples a random noise distribution and transforms it into a high quality output by using the reverse diffusion process it has learned.

1) Training from Scratch

As a first approach, a diffusion model was trained from scratch. The implemented diffusion model employs a UNet-style architecture for image-to-image transformation tasks, available on HuggingFace (9). The network consists of multiple downsampling and upsampling blocks with residual layers:

- Sample size: primarily set to 128 or 224 pixels
- In/out channels: 3-channel RGB images
- Layers per block: 2 residual layers per block
- Channel dimensions: progressive channel expansion (128, 128, 256, 256, 512, 512)
- Down and up blocks: a combination of standard convolutional blocks and attention-based blocks

To generate images from different classes, separate models were trained for different classification scenarios:

- **Three-class Setup:** Three independent diffusion models are trained, each focused on a single class. To ensure a more balanced representation but still offering a realistic distribution, images were generated to reach 40% adenomatous, 30% sessile serrated and 30% hyperplastic polyps.
- **AD-vs-Rest Setup:** One model is trained exclusively on AD, a second model is trained on a combination of HP + SSA polyps. The distribution strategy for this setup is 60% for AD and 40% for the Rest class.

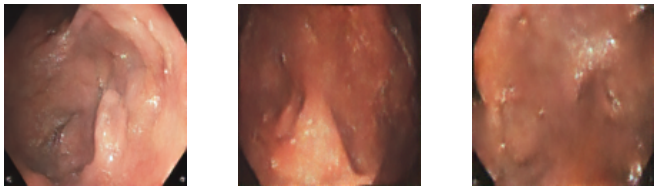
Several experiments were conducted, iterating over different configurations to optimize model performance.

In this approach, all generations were unconditional, meaning the model was trained only based on images without using any text prompts (an empty string was provided as the prompt).

• **Configuration 1:** An initial configuration was tried to assess the feasibility of training a diffusion model from scratch:

- Image size: 128×128
- Batch size: 16
- Epochs: 100
- Diffusion timesteps: 1000

The model successfully generated coherent images across different classes, establishing a strong starting point. In Figure 7, some sample images can be seen:



(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 7: Synthetic polyp images generated with Configuration 1.

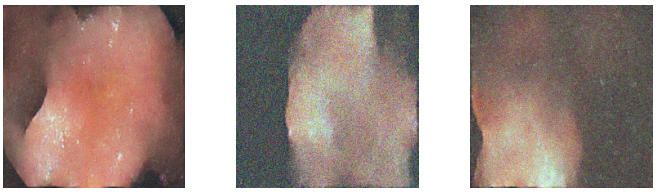
As observed, adenomatous images seem to be the best generated, likely because the model has more samples of this class to learn from.

• **Configuration 2:** To improve image fidelity, the resolution was increased to 224×224. Due to memory constraints, some optimization techniques were applied, including the use of automatic mixed precision package, which reduces memory usage and speeds up training by using lower-precision computations without significantly sacrificing model accuracy, as well as reducing the batch size.

- Image size: 224×224

- Batch size: 4
- Epochs: 100
- Timesteps: 1000

The generated images (Figure 8) preserved the basic structure but were noticeably noisier compared to the 128×128 outputs, likely because the model had to learn more detailed features and required additional training time.



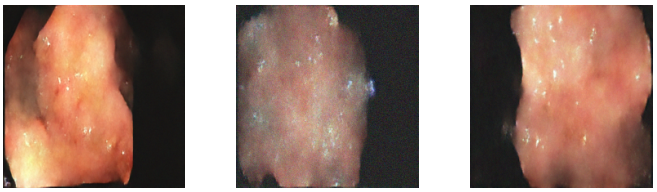
(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 8: Synthetic polyp images generated with Configuration 2.

• **Configuration 3:** To address the noise issue and allow the model more learning capacity, further refinement was done by increasing both training duration and the number of diffusion steps:

- Epochs: Increased to 200
- Timesteps: Increased to 2000

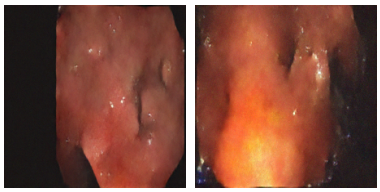
With this configuration, image quality showed some improvement, although noise was still apparent, which are shown in Figure 9.



(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 9: Synthetic polyp images generated with Configuration 3.

• **Configuration 4:** A different configuration was tested using varying channel dimensions (128, 256, 384, 512, 512, 768) along with the AD-vs-Rest setup.



(a) Adenomatous (b) SSA and HP

Fig. 10: Synthetic polyp images generated with Configuration 4.

The resulting images have a noisy appearance as in previous configurations (Figure 10).

2) Fine-Tuning a Pretrained Model

Since the diffusion model trained from scratch showed some limitations with regard to image quality, a pretrained model was used to try to further achieve cleaner images.

The pretrained diffusion model used was the Stable Diffusion v1.4 model, originally released by CompVis, a computer vision and learning research group at Ludwig Maximilian University of Munich, as stated on their GitHub page (10).

This model is publicly available on Hugging Face (11) and performs image generation in a compressed latent space using a combination of different blocks (see Figure 11):

- **Variational Autoencoder (VAE):** to encode and decode images to/from latent space.
- **U-Net architecture:** with cross-attention, for denoising. This is the actual model that will be finetuned and trained.
- **CLIP-based text encoder:** for conditioning image generation on text prompts. If an empty prompt is passed, the image generation is unconditional. Can be trained or not.
- **Noise scheduler:** to manage the diffusion timesteps during training and inference.

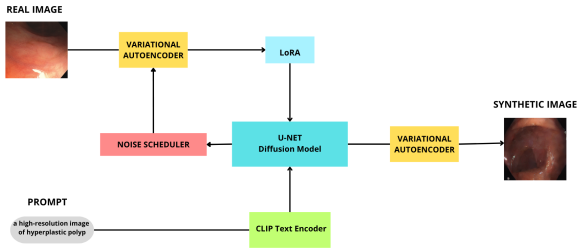


Fig. 11: Architecture diagram of the Stable Diffusion model illustrating its main components and workflow.

These models are huge and do not fit into the GPU, so Low-Rank Adaptation (LoRA) has been applied.

LoRA technique allows to finetune large models by adding small, changeable layers, in this case, to the U-Net, and freezing the rest of the model. It provides a quick way to adapt the model without retraining it (see article by Hu et al. (12)).

As detailed on the IBM website (13), a high-rank matrix can be approximated by the product of two low-rank matrices. LoRA takes advantage of this by adding two low-rank matrices of rank r to the original, frozen model. During finetuning, only these low-rank matrices are updated through gradient descent, while the base model remains unchanged. The updated low-rank matrices represent the learned changes and are combined with the original model weights by multiplication and addition. This results in the final finetuned model used for generating outputs (see the workflow in Figure 35, in the Appendix).

For example, a full finetuning of the U-Net model requires training 866 million parameters. Using LoRA, these

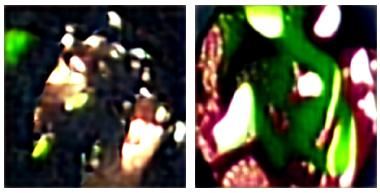
trainable parameters can be reduced, if using a rank of 16, up to 3 million parameters, which means only 0.37% of the model's parameters will be trained.

The first experiment represented the starting point of basic LoRA finetuning on the pretrained model. It set the foundation for understanding how later modifications would affect output.

• **Experiment 1:** Training the model with the prompt *a high-resolution endoscopic image of sessile serrated polyp* and LoRA rank of 8.

(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 12: Synthetic polyp images generated with experiment 1 for the three-class setup.



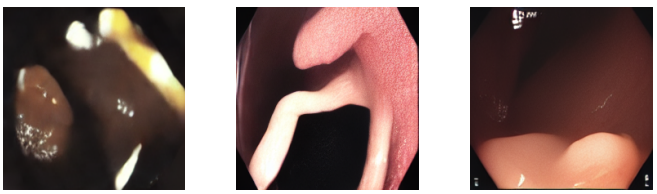
(a) Adenomatous (b) SSA and HP

Fig. 13: Synthetic polyp images generated with experiment 1 for the AD-vs-Rest setup.

This first approach generated very abstract images (Figures 12 and 13), lacking the colors, shapes and textures present in real images.

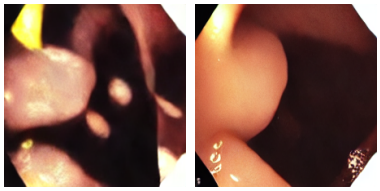
Based on these results, LoRA would have to be adapted in order to learn more from the dataset. In the following experiment, custom attention targets were introduced, achieving deeper module adaptation.

• **Experiment 2:** Training the model with no prompt, which means an unconditional generation, LoRA attention layers were added and rank was increased to 16.



(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 14: Synthetic polyp images generated with experiment 2 for the three-class setup.



(a) Adenomatous (b) SSA and HP

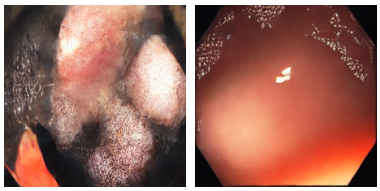
Fig. 15: Synthetic polyp images generated with experiment 2 for the AD-vs-Rest setup.

With this unconditional approach, images started to show some form and colors (Figures 14 and 15). The model seemed to try to represent polyps, although still not close to real images.

• **Experiment 3:** This experiment deviated from pure LoRA by unfreezing some layers of the base model, allowing more expressive tuning. Since all LoRA layers were already active, this setup let the model learn further by also updating some of its original layers.

(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 16: Synthetic polyp images generated with experiment 3 for the three-class setup.



(a) Adenomatous (b) SSA and HP

Fig. 17: Synthetic polyp images generated with experiment 3 for the AD-vs-Rest setup.

In this case, it also seemed to start learning some colors and forms, but still very abstract images (Figures 16 and 17), as in the previous experiment.

Finally, a last experiment was conducted, returning to the conditional training. It was the most advanced strategy, combining Dreambooth with custom weighted loss.

• **Experiment 4:** Apart from training the model U-Net itself, the text encoder was also trained, using Dreambooth with custom weighted loss, and LoRA rank of 16.

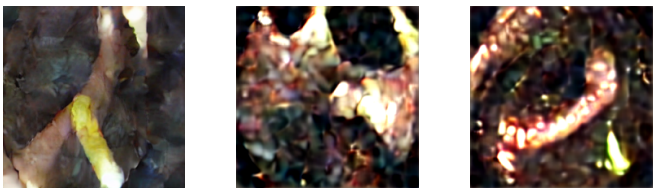
DreamBooth is a fine-tuning technique for diffusion models that allows the model to generate personalized images of a specific concept (see articles Voetman et al. (3) and Ruiz et al. (14)).

It works by teaching the model to associate a unique identifier with the visual features of the subject. The unique

identifier has to be a token that does not exist in our vocabulary (e.g., *sks*).

Hence, the prompt in this experiment is *a high-resolution endoscopic photo of sks adenomatous polyp*. With this prompt, the text encoder is trained and the U-Net pays special attention to the tokens after the unique identifier *sks*, which are *adenomatous polyp*. Each class receives a unique identifier, so the model will have more input help to learn the different classes.

Apart from that, since it generated better images without prompt, a custom weighted loss was set, in which it was given more weight to the input image, and less weight to the prompt, in order to have images closer to the dataset as in the unconditional generation, but still having some sort of signal with the prompt to guide the training.



(a) Adenomatous (b) SSA (c) Hyperplastic

Fig. 18: Synthetic polyp images generated with experiment 4 for the three-class setup.

(a) Adenomatous (b) SSA and HP

Fig. 19: Synthetic polyp images generated with experiment 4 for the AD-vs-Rest setup.

In this last approach, the images generated (Figures 18 and 19) appeared to have a mosaic-like texture with regular structure. The model seemed to be starting to learn the underlying texture and spatial layout of polyp images, but still did not learn it yet.

IV Quantitative Results

In this section, quantitative results are presented. The performance of the classifier trained with augmented dataset (with both real and synthetic data) is compared to the baseline model trained only on real data.

To ensure a fair comparison, synthetic images were added only to the training set, while the test set remained unchanged.

It is important to note that in medical diagnosis, a high recall is especially important, since it indicates that the model correctly identifies most of the actual positive cases. Missing a malignant polyp can be more dangerous than predicting it as malignant by mistake.

A. Classifier Baseline Results

TABLE II: Baseline classifier results trained only with real data.

Image Size: 128×128		
Technique	Accuracy	Macro F1
None	0.60	0.43
Weighted Sampling	0.57	0.56
Weighted Loss	0.42	0.41
AD-vs-Rest	0.64	0.64
Image Size: 224×224		
Technique	Accuracy	Macro F1
None	0.64	0.64
Weighted Sampling	0.67	0.65
Weighted Loss	0.61	0.62
AD-vs-Rest	0.64	0.63

As shown in Table II, in the three-class setup, the macro F1 score is usually lower than the accuracy. For instance, with image size 128×128 and no technique applied, the model achieves an accuracy of 0.60 but only a macro F1 of 0.43. This is expected, as macro F1 gives equal weight to all classes, while accuracy can be high even if the model only performs well on the majority class.

When using image size 128×128, weighted sampling offers a better trade-off between accuracy and macro F1, achieving 0.57 and 0.56, respectively. Although it does not have the highest accuracy, its macro F1 is much closer to the accuracy, indicating more balanced performance across the classes. Therefore, for this resolution, weighted sampling is chosen as it provides a better balance.

For the 224×224 image size, weighted sampling achieves the best results overall, with an accuracy of 0.67 and a macro F1 of 0.65. These are the highest values among all tested techniques, so it is also chosen as a balancing technique.

Regarding the AD-vs-Rest setup, the macro F1 scores remain high: 0.64 for 128×128 and 0.63 for 224×224. This setup groups the two benign classes (HP and SSA) into one class, which helps reduce the imbalance and gives more training examples per class. Because of this, we will use AD-vs-Rest as an additional setup rather than a technique that is discarded, to assess whether models are able to classify between malignant (AD) and non-malignant polyps in general.

B. Impact of Synthetic Images from Diffusion Model Trained from Scratch

- **Configuration 1:** initial experimental setup, using 128×128 images, 100 training epochs, and 1000 diffusion steps.

TABLE III: Comparison of classification metrics per class for Configuration 1.

Baseline Results			
	Precision	Recall	F1-score
AD	0.68	0.79	0.73
SSA	0.52	0.32	0.40
HP	0.20	0.15	0.17
Results with Synthetic Images			
	Precision	Recall	F1-score
AD	0.67	0.82	0.74
SSA	0.36	0.11	0.17
HP	0.24	0.20	0.22

In Table III, the baseline model performs best on AD, with precision 0.68, recall 0.79, and F1-score 0.73. After adding synthetic images, AD remains stable with recall improving slightly to 0.82 and F1 to 0.74, while precision stays at 0.67.

HP shows small gains after augmentation: precision rises from 0.20 to 0.24, recall from 0.15 to 0.20, and F1 from 0.17 to 0.22.

However, SSA performance drops notably, with precision falling from 0.52 to 0.36, recall from 0.32 to 0.11, and F1 from 0.40 to 0.17, suggesting the synthetic SSA images may lack sufficient quality or diversity.

The baseline confusion matrix (Figure 20 in Appendix) shows the model performs best on AD, with clear confusion between SSA, HP, and other classes. This matches the table, where AD has the highest F1-score and SSA and HP perform worse.

In the confusion matrix for configuration 1 (Figure 23 in Appendix), AD prediction improves slightly, reflected in the higher F1-score. However, SSA and HP performance drops sharply, with most examples misclassified as AD, consistent with their F1-scores dropping to zero.

In short, while the synthetic images may help improve AD classification, they still struggle to distinguish SSA and HP, as in the previous configuration. The confusion matrix confirms this by showing heavy misclassifications of SSA and HP as AD.

• **Configurations 2 and 3:** larger image resolution (224×224) and extended training (200 epochs and 2000 timesteps).

TABLE IV: Comparison of classification metrics per class for configurations 2 and 3.

Baseline results			
	Precision	Recall	F1-score
AD	0.85	0.68	0.76
SSA	0.44	0.57	0.49
HP	0.43	0.62	0.51
Results with larger image sizes (configuration 2)			
	Precision	Recall	F1-score
AD	0.72	0.81	0.76
SSA	0.00	0.00	0.00
HP	0.46	0.57	0.51
Results with longer training (configuration 3)			
	Precision	Recall	F1-score
AD	0.74	0.74	0.74
SSA	0.44	0.11	0.17
HP	0.35	0.60	0.44

In Table IV, the baseline model performs best on AD, with precision 0.85, recall 0.68, and F1-score 0.76. SSA and HP show more balanced but lower F1-scores of 0.49 and 0.51, respectively.

With larger images and longer training (Configuration 2), recall for AD increases to 0.81 while F1 remains at 0.76. HP performance stays stable, with recall slightly decreasing from 0.62 to 0.57 and F1 constant at 0.51. However, SSA performance drops, with precision, recall, and F1 all falling to 0.

In Configuration 3, longer training leads to a more balanced outcome: AD precision and recall both reach 0.74, HP shows a slight decline (precision 0.35, recall 0.60, F1 0.44), and SSA improves somewhat from Configuration 2, achieving precision 0.44, recall 0.11, and F1 0.17. This indicates that larger image sizes improve AD recall and maintain HP performance, while longer training helps SSA minimally, though the class remains challenging.

The baseline confusion matrix (Figure 21 in Appendix) shows good performance across all classes, with AD predicted best and some confusion in SSA and HP, matching the balanced F1-scores.

In the confusion matrix of configuration 2 (Figure 24 in Appendix), AD prediction improves, but SSA is completely missed, mostly misclassified as AD, while HP stays stable. This indicates larger image sizes help AD but harm SSA detection.

Configuration 3's matrix (Figure 25 in Appendix) shows AD still detected well, SSA with some correct predictions, and stable HP performance. This aligns with the slight SSA improvement, suggesting longer training helps but only marginally.

After analyzing these results, it can be concluded that larger image sizes help improve the recall for AD and keep performance stable for HP, but do not help at all for SSA,

while longer training improves results slightly for HP and SSA.

• **Configuration 4:** Varying channel dimensions and AD-vs-Rest setup.

TABLE V: Comparison of classification metrics per class for configuration 4.

Baseline results			
	Precision	Recall	F1-score
AD	0.78	0.61	0.68
REST	0.47	0.67	0.56
Results with augmented dataset			
	Precision	Recall	F1-score
AD	0.85	0.58	0.69
REST	0.50	0.80	0.62

Table V shows the results for Configuration 4, where the model uses the AD-vs-Rest setup (grouping SSA and HP together) and tests a different channel size in the diffusion model.

Table V presents results for Configuration 4 using the AD-vs-Rest setup, where SSA and HP are combined. The baseline model performs well on AD, with precision 0.78, recall 0.61, and F1-score 0.68. For the Rest class, recall is higher at 0.67 but precision is lower at 0.47, resulting in an F1-score of 0.56, indicating some false positives labeled as Rest.

After augmenting with synthetic images, the AD class shows improved precision from 0.78 to 0.85, a slight drop in recall from 0.61 to 0.58, and a small F1-score increase to 0.69. The Rest class sees notable improvements in recall from 0.67 to 0.80 and precision from 0.47 to 0.50, raising the F1-score to 0.62. This suggests synthetic augmentation helps overall, particularly enhancing detection of the Rest class.

This configuration appears to help improve performance for both classes, even though AD recall slightly decreases. The changes in precision and recall show a better balance between both classes after augmentation.

The confusion matrices confirm the table results. In the baseline (Figure 22 in Appendix), many AD samples are misclassified as Rest, lowering AD recall, while Rest has many false positives, reducing precision.

After augmentation (Figure 26 in Appendix), fewer Rest false positives improve Rest precision and recall. AD predictions remain similar, with a slight recall drop due to more AD samples misclassified as Rest, but precision increases thanks to fewer false positives.

In summary, augmenting the training dataset with a diffusion model trained from scratch has not resulted in significant improvements, particularly for minority classes such as SSA.

C. Impact of synthetic images from pretrained diffusion

The baseline classifier used in all experiments in this section is the one trained with images of size 224×224.

Its results for the three-class setup are shown in Table IV, and for the AD-vs-Rest setup in Table V.

To make comparisons easier, these baseline tables are included in each experiment. However, the detailed analysis of the results is only provided once, in Section IV-B. The analysis of the confusion matrices is also discussed there, although the matrices themselves are included in the appendix (Figures 21 and 22 for the three-class and AD-vs-Rest setups, respectively).

• **Experiment 1:** basic training with LoRA.

TABLE VI: Comparison of classification metrics for three-class setup in experiment 1.

Baseline results			
	Precision	Recall	F1-score
AD	0.85	0.68	0.76
SSA	0.44	0.57	0.49
HP	0.43	0.62	0.51
Results from experiment 1			
	Precision	Recall	F1-score
AD	0.69	0.91	0.78
SSA	0.20	0.02	0.05
HP	0.48	0.30	0.37

In the three-class setup shown in Table VI, after augmenting the training set, the classifier improves its recall for AD from 0.68 to 0.91 and increases its F1-score slightly from 0.76 to 0.78, while precision lowers moderately from 0.85 to 0.69. However, the performance on SSA drops significantly, with its recall falling from 0.57 to only 0.02, its F1-score from 0.49 to 0.05 and its precision from 0.44 to 0.2. The HP class also sees a decline in its recall going from 0.62 to 0.3 and F1-score going from 0.51 to 0.37, but a little augment in preicison from 0.43 to 0.48.

TABLE VII: Comparison of classification metrics for AD-vs-Rest setup in experiment 1.

Baseline results			
	Precision	Recall	F1-score
AD	0.78	0.61	0.68
REST	0.47	0.67	0.56
Results from experiment 1			
	Precision	Recall	F1-score
AD	0.64	0.89	0.74
REST	0.15	0.04	0.06

In the AD-vs-Rest setup (Table VII), augmentation boosts AD recall from 0.61 to 0.89 and F1-score to 0.74, despite precision dropping from 0.78 to 0.64. However, Rest class performance collapses, with its F1-score falling from 0.56 to 0.06, indicating the synthetic images do not help the model generalize well to HP and SSA.

Comparing the baseline and augmented confusion matrices (Figure 27, see Appendix), the model clearly improves at identifying AD samples, matching the recall and F1-score gains in Table VI. However, SSA and HP detection worsens, with fewer correct predictions and drops in recall and F1-score. HP precision improves slightly, showing more confident but fewer detections.

In the AD-vs-Rest setup (Figure 28 in Appendix), both classes show better separation after augmentation. More AD samples are correctly classified, increasing precision, while Rest also benefits from fewer errors, boosting recall and F1-scores.

From the results obtained from the three-class setup, the synthetic images generated by the LoRA method appears to help the classifier better detect AD but negatively impact its performance on the other two classes, while for the AD-vs-Rest setup, the classifier better separates malignant (AD) from non-malignant (Rest) polyps. The improved balance between precision and recall suggests that the synthetic data enhances the model's ability to generalize across both classes.

• **Experiment 2:** unconditional generation and increase of LoRA layers and rank.

TABLE VIII: Comparison of classification metrics for three-class setup in experiment 2.

Baseline results			
	Precision	Recall	F1-score
AD	0.85	0.68	0.76
SSA	0.44	0.57	0.49
HP	0.43	0.62	0.51
Results from experiment 2			
	Precision	Recall	F1-score
AD	0.69	0.86	0.77
SSA	0.22	0.22	0.22
HP	1.00	0.08	0.14

In the three-class setup (Table VIII), augmentation increases recall from 0.68 to 0.86 but lowers precision from 0.85 to 0.69, leading to a slight F1-score improvement from 0.76 to 0.77. However, SSA and HP performance drops sharply, with F1-scores falling to 0.22 and 0.14, indicating low-quality synthetic images.

TABLE IX: Comparison of classification metrics for AD-vs-Rest setup in experiment 2.

Baseline results			
	Precision	Recall	F1-score
AD	0.78	0.61	0.68
REST	0.47	0.67	0.56
Results from experiment 2			
	Precision	Recall	F1-score
AD	0.84	0.56	0.67
REST	0.48	0.79	0.60

For the AD-vs-Rest setup (Table IX), the AD F1-score stays around 0.67, while the Rest class improves to 0.60. AD precision rises from 0.78 to 0.84, reducing false positives, and Rest recall increases from 0.67 to 0.79, showing better non-AD detection.

The three-class confusion matrix (Figure 29, Appendix) shows improved AD recall with augmentation, matching the table results. However, AD precision drops due to more false positives. SSA and HP classes perform worse, with fewer correct predictions, reflecting their lower F1-scores and suggesting less effective synthetic data for these classes.

In the AD-vs-Rest setup (Figure 30, Appendix), augmentation improves recall and precision for Rest, raising its F1-score, while AD performance remains stable.

In summary, the augmented data enhances detection of AD and Rest in the two-class setup, but the benefit is limited in the three-class setting.

• **Experiment 3:** increase of LoRA layers and unfreezing some layers of the original model.

TABLE X: Comparison of classification metrics for three-class setup in experiment 3.

Baseline results			
	Precision	Recall	F1-score
AD	0.85	0.68	0.76
SSA	0.44	0.57	0.49
HP	0.43	0.62	0.51
Results from experiment 3			
	Precision	Recall	F1-score
AD	0.71	0.72	0.71
SSA	0.00	0.00	0.00
HP	0.41	0.55	0.47

In the three-class setup (Table X), when augmenting the dataset, the F1-score for AD drops slightly to 0.71, with precision going from 0.85 to 0.71 and recall improving a bit from 0.68 to 0.72. However, the SSA class shows a complete loss of performance, with precision, recall, and F1-score all at 0, meaning the model cannot detect this class after augmentation. The HP class also decreases slightly, with F1-score falling from 0.51 to 0.47, precision from 0.43 to 0.41, and recall from 0.62 to 0.55.

TABLE XI: Comparison of classification metrics for AD-vs-Rest setup in experiment 3.

Baseline results			
	Precision	Recall	F1-score
AD	0.78	0.61	0.68
REST	0.47	0.67	0.56
Results from experiment 3			
	Precision	Recall	F1-score
AD	0.78	0.67	0.72
REST	0.50	0.63	0.56

For the two-class setup (Table XI), the AD class improves slightly in recall from 0.61 to 0.67 and in F1-score from 0.68 to 0.72, while precision stays the same at 0.78. For the Rest class, precision improves a little from 0.47 to 0.50, but recall decreases from 0.67 to 0.63, keeping the F1-score stable at 0.56.

The augmented confusion matrix (Figure 31, Appendix) shows a slight decrease in AD precision but improved recall, explaining the mixed metric changes. The model fails to detect SSA entirely after augmentation, consistent with zero scores and no correct SSA predictions. HP performance drops slightly with more confusion and fewer correct predictions.

In the AD-vs-Rest setup (Figure 32, Appendix), AD recall improves slightly while precision remains stable. Rest precision rises a bit but recall drops slightly, keeping the F1-score steady—matching the table's results.

In this experiment, the AD class maintains similar or slightly better recall and F1-score in both setups. However, the SSA class suffers greatly in the three-class setup, with no detections after augmentation, and the HP class also shows a small decline. The two-class setup sees some improvement in AD recall and stability in the Rest class.

• **Experiment 4:** training of U-Net and text encoder, with Dreambooth and a custom weighted loss.

TABLE XII: Comparison of classification metrics for three-class setup.

Baseline results			
	Precision	Recall	F1-score
AD	0.85	0.68	0.76
SSA	0.44	0.57	0.49
HP	0.43	0.62	0.51
Results from experiment 4			
	Precision	Recall	F1-score
AD	0.65	0.98	0.78
SSA	0.00	0.00	0.00
HP	0.00	0.00	0.00

In the three-class setup (Table XII), the F1-score for AD increases slightly to 0.78, with a much higher recall (from 0.68 to 0.98) but lower precision (from 0.85 to 0.65). However, performance for SSA and HP drops completely to 0 in all three metrics: precision, recall, and F1-score. This shows that after training with the augmented data, the model is no longer able to recognize the SSA and HP classes at all.

TABLE XIII: Comparison of classification metrics for AD-vs-Rest setup.

Baseline results			
	Precision	Recall	F1-score
AD	0.78	0.61	0.68
REST	0.47	0.67	0.56
Results from experiment 4			
	Precision	Recall	F1-score
AD	0.77	0.74	0.76
REST	0.54	0.58	0.56

In contrast, in the two-class setup, when looking at Table XIII), it can be seen that the F1-score for AD improves to 0.76, with better recall (from 0.61 to 0.74) and a small drop in precision (from 0.78 to 0.77). For the Rest class, precision increases from 0.47 to 0.54, but recall slightly drops from 0.67 to 0.58, keeping the F1-score stable at 0.56.

In the three-class setup (Figure 33, Appendix), augmentation greatly improves AD recall, but the model fails completely to detect SSA and HP, with all metrics for these classes dropping to zero. This shows augmentation helps AD but harms multi-class performance.

In the two-class setup (Figure 34 in Appendix), augmentation boosts AD recall with a slight precision drop, while Rest precision improves but recall decreases a bit. Overall, F1-scores stay stable or improve slightly, showing a more balanced effect.

These results show that the two-class setup performs more reliably, with small improvements for AD and stable results for Rest. This suggests that the augmentation strategy in this experiment may work better in binary classification, but needs improvement for multi-class tasks.

In summary, the results demonstrate a consistent improvement in the classification of adenomatous polyps across most setups, likely due to their higher representation in the dataset. Hyperplastic polyps showed mixed results, with some improvements, but also some drops in performance. The most challenging class to classify was sessile serrated polyps, which frequently experienced a significant decrease in performance. This could be due to their visual similarity with both AD and HP polyps, making them more difficult for the model to distinguish. Furthermore, as both SSA and HP belong to the minority classes, the imbalance likely contributed to their poor image generation and, therefore, lower classification accuracy. Overall, AD remains the most reliably identified class, while further work is needed to enhance the model's ability to distinguish the more ambiguous and underrepresented classes.

V Conclusions

The experiments carried out in this project explored the use of diffusion models to address data scarcity and class imbalance in polyp classification. Although the results did not show a major improvement in performance, they provided valuable insights and highlighted important limitations that could guide future research.

One of the main challenges was the SSA class. As described in the previous section IV-C, this class is difficult to classify because it shares features with both HP and AD classes. For that reason, the models often got confused.

The diffusion model trained from scratch did not perform very well. The images it generated appeared blurry, still very noisy, although they looked similar to the real ones. This suggests that the model needs more training data to learn properly.

In contrast, the pretrained diffusion model needed less data because of its previous knowledge. However, it did not work very well in this use case. One possible reason is that the original Stable Diffusion model was trained on images of size 512x512, while in this case, the image size had to be reduced to 224x224, significantly affecting the quality of the outputs, because the VAE and U-Net were trained and optimized for higher-resolution images.

Additionally, medical images need a lot of detail, and reducing their size already leads to important information being lost, such as structure and texture, even if the U-Net model had been originally trained at 224x224.

An attempt of training with images of size 512x512 was done, but did not fit the GPU, even when using LoRA with a lower rank and using the package automatic mixed precision.

Even with these problems, experiments have shown that there has been some improvement, especially in the AD-vs-Rest setup.

For future work, it would be a good starting point to use higher-resolution images (512x512) if possible. This could improve the quality of the synthetic images and help classifiers work better. Moreover, it could help to try other quantities or class distributions to avoid strong bias toward synthetic data.

In conclusion, while the results in this work were limited, they show that diffusion models have strong potential in medical imaging tasks. With further research, more data, and proper tuning, these models could become a powerful tool for addressing data imbalance and enhancing classification performance in clinical applications.

References

[1] University of Michigan Health, "Colon and rectal polyps," https://www.uofmhealth.org/conditions-treatments/digestive-and-liver-health/colon-and-rectal-polyps, accessed: 2025-02-17.

[2] Bowel Cancer Australia, "Understanding a colonoscopy procedure," https://www.bowelcanceraustralia.org/colonoscopy/, accessed: 2025-02-26.

[3] R. Voetman, M. Aghaei, and K. Dijkstra, "The big data myth: Using diffusion models for dataset generation to train deep detection models," *arXiv preprint arXiv:2306.09762*, 2023.

[4] J. T. Guibas, T. S. Virdi, and P. S. Li, "Synthetic medical images from dual generative adversarial networks," *arXiv preprint arXiv:1709.01872*, 2017.

[5] N. De La Fuente, M. Majó, I. Luzko, H. Córdova, G. Fernández-Esparrach, and J. Bernal, "Enhancing image classification in small and unbalanced datasets through synthetic data augmentation," in *Workshop on Clinical Image-Based Procedures*. Springer, 2024, pp. 11–21.

[6] G. Marques, D. Agarwal, and I. De la Torre Diez, "Automated medical diagnosis of covid-19 through efficientnet convolutional neural network," *Applied soft computing*, vol. 96, p. 106691, 2020.

[7] D. Bergmann and C. Stryker, "What are diffusion models?" https://www.ibm.com/think/topics/diffusion-models, 2024, accessed: 2025-06-04.

[8] "What are diffusion models?" https://www.iguazio.com/glossary/diffusion-models/, accessed: 2025-06-15.

[9] Hugging Face, "Unet2dmodel," https://huggingface.co/docs/diffusers/main/api/models/unet2d, accessed: 2025-04-03.

[10] CompVis, "Compvis github organization," accessed: 2025-06-19.

[11] "Compvis/stable-diffusion-v1-4," https://huggingface.co/CompVis/stable-diffusion-v1-4, accessed: 2025-06-03.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[13] J. Noble, "What is lora (low-rank adaptation)?" https://www.ibm.com/think/topics/lora, 2025, accessed: 2025-06-03.

[14] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

Appendix

Appendix

The complete implementation of the project, including preprocessing scripts, training pipelines, and evaluation tools, is available at the following GitHub repository: https://github.com/nereaqing/Polyp-Image-Generator

A. Confusion Matrices of Baseline Classifier Models

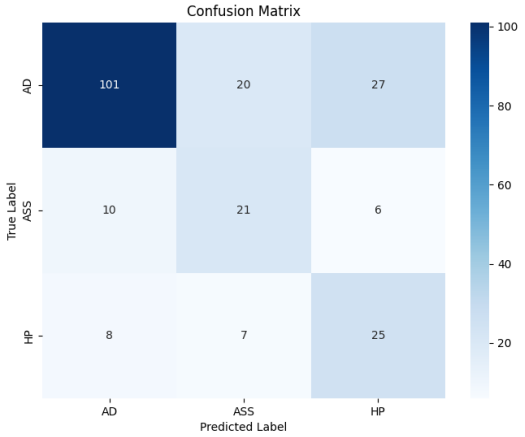


Fig. 20: Confusion matrix showing the classification performance of the baseline model using 128×128 input images.

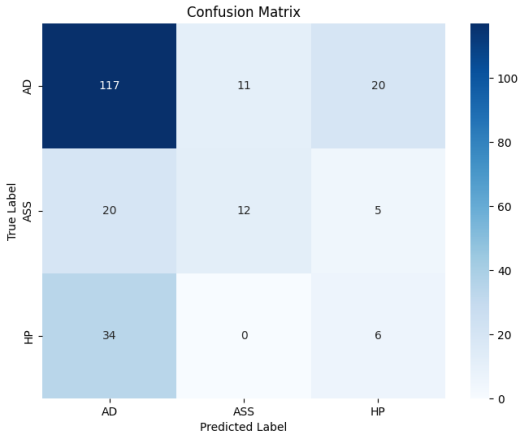


Fig. 21: Confusion matrix of the baseline model trained with images of size 224×224.

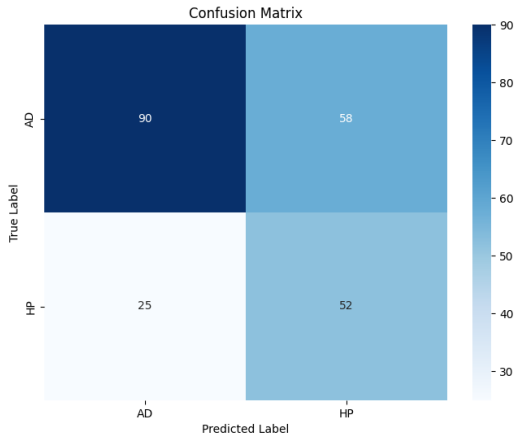


Fig. 22: Confusion matrix of the baseline model trained with images of size 224x224, using AD-vs-Rest setup.

B. Confusion Matrices of Classifier Models Trained with Synthetic Data

1) Diffusion Model Trained from Scratch

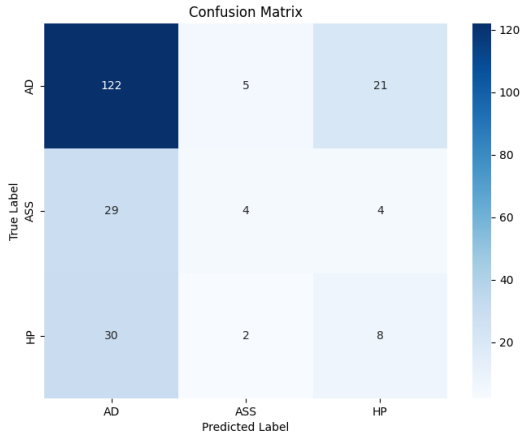


Fig. 23: Confusion matrix of the classifier model trained with synthetic images of size 128x128.

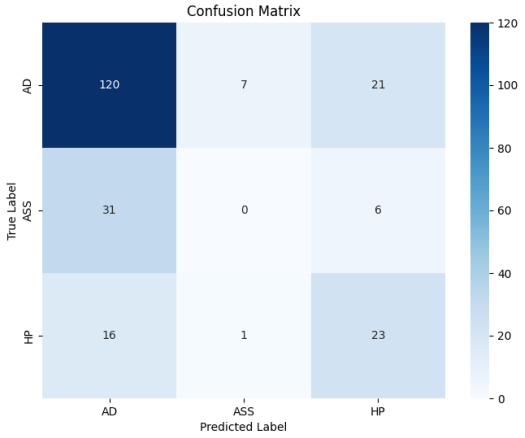


Fig. 24: Confusion matrix of the classifier model trained with synthetic images of size 224x224.

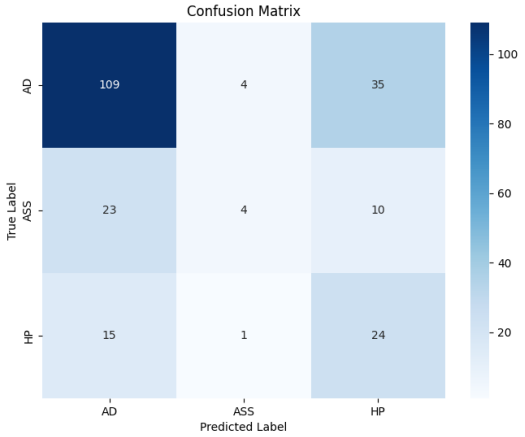


Fig. 25: Confusion matrix of the classifier model trained with synthetic images of size 224x224 and longer training.

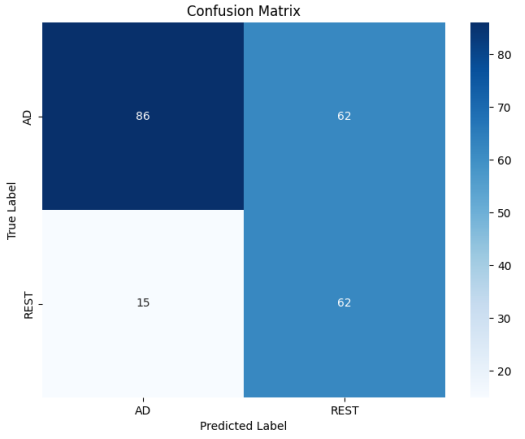


Fig. 26: Confusion matrix of the classifier model trained with synthetic images of size 224x224 with AD-vs-Rest setup.

2) Pretrained Diffusion Model

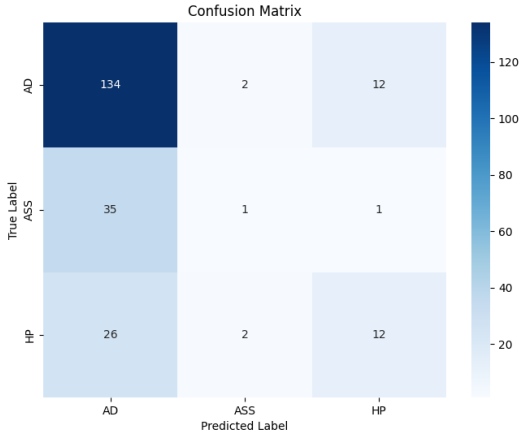


Fig. 27: Confusion matrix using synthetic data from conditional diffusion and LoRA rank of 8.

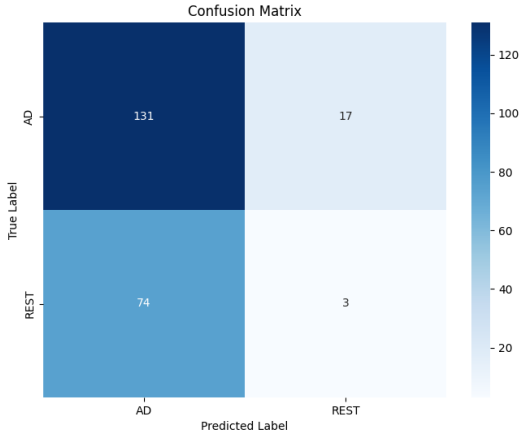


Fig. 28: Confusion matrix using synthetic data from conditional diffusion and LoRA rank of 8, with AD-vs-Rest setup.

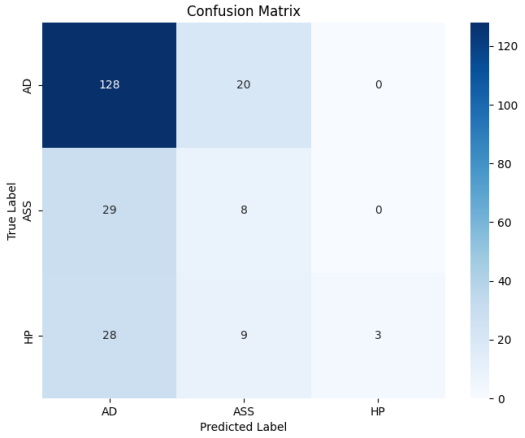


Fig. 29: Confusion matrix using synthetic data from unconditional training, addition of LoRA attention layers and rank of 16.

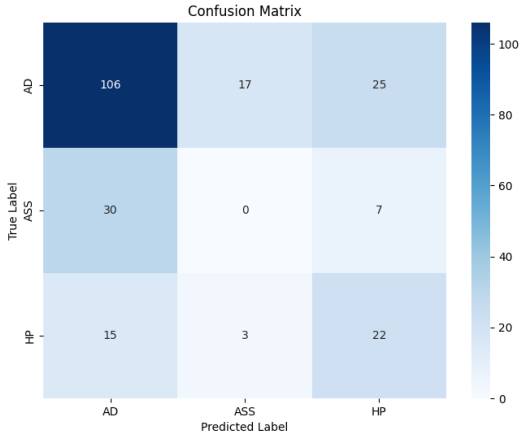


Fig. 31: Confusion matrix using synthetic data from unconditional training and unfreezing original model's layers.

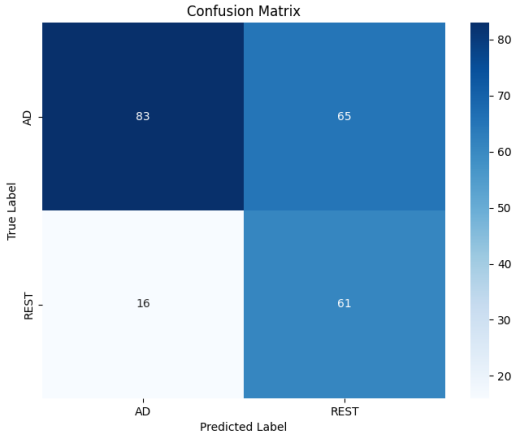


Fig. 30: Confusion matrix using synthetic data from unconditional training, addition of LoRA attention layers and rank of 16, using AD-vs-Rest setup.

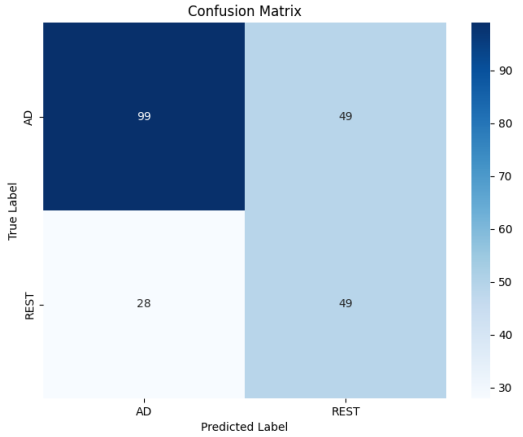


Fig. 32: Confusion matrix using synthetic data from unconditional training and unfreezing original model's layers, using AD-vs-Rest setup.

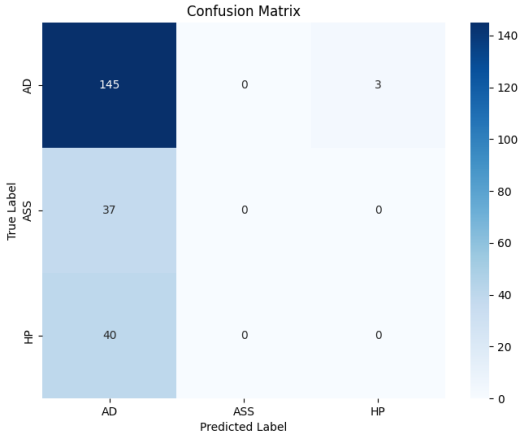


Fig. 33: Confusion matrix using synthetic data from conditional diffusion, training of text encoder, using Dreambooth and a custom weighted loss.

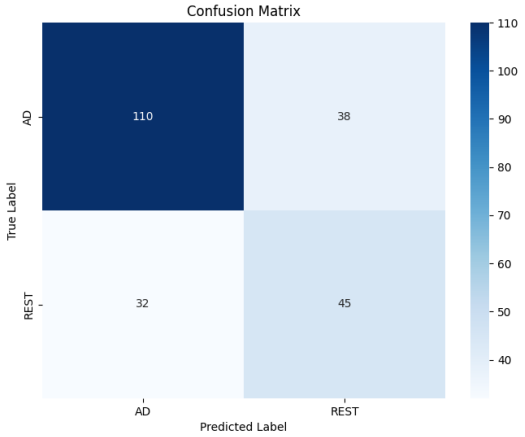


Fig. 34: Confusion matrix using synthetic data from conditional diffusion, training of text encoder, using Dreambooth and a custom weighted loss. Use of AD-vs-Rest setup.

C. Diagrams

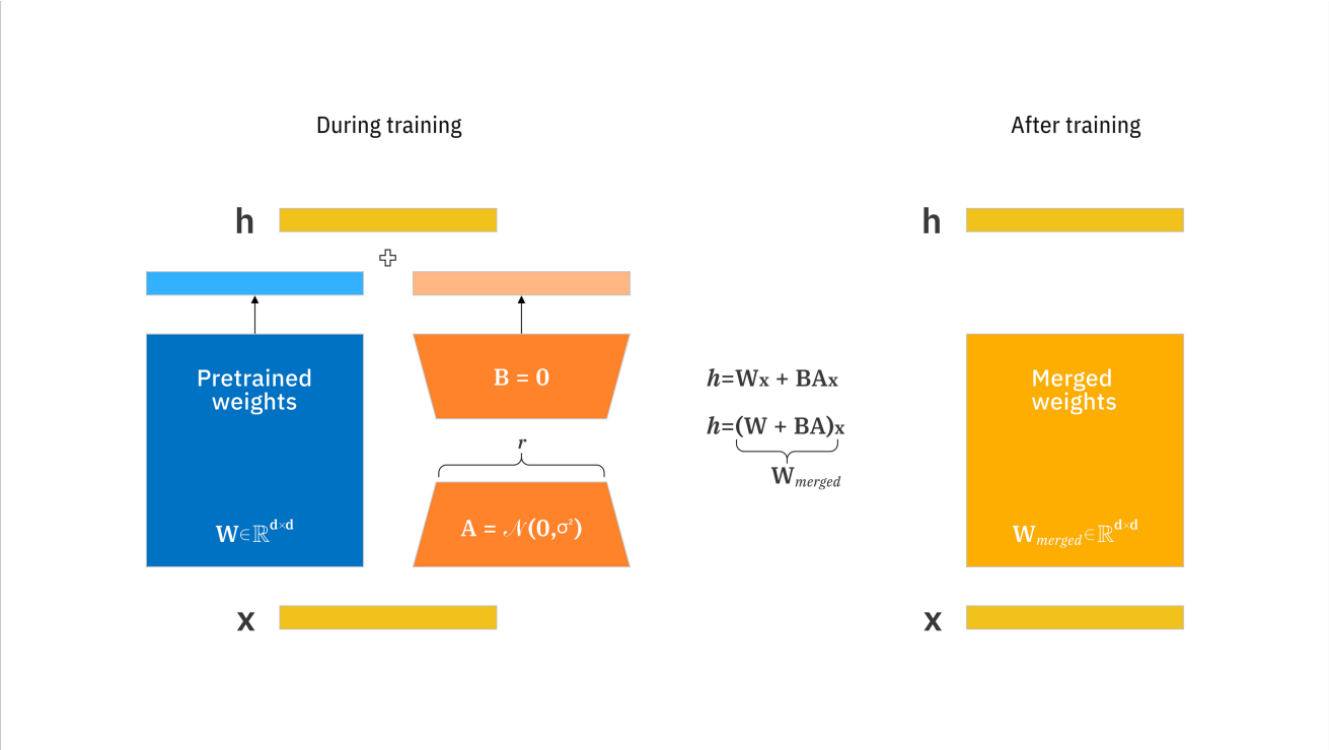
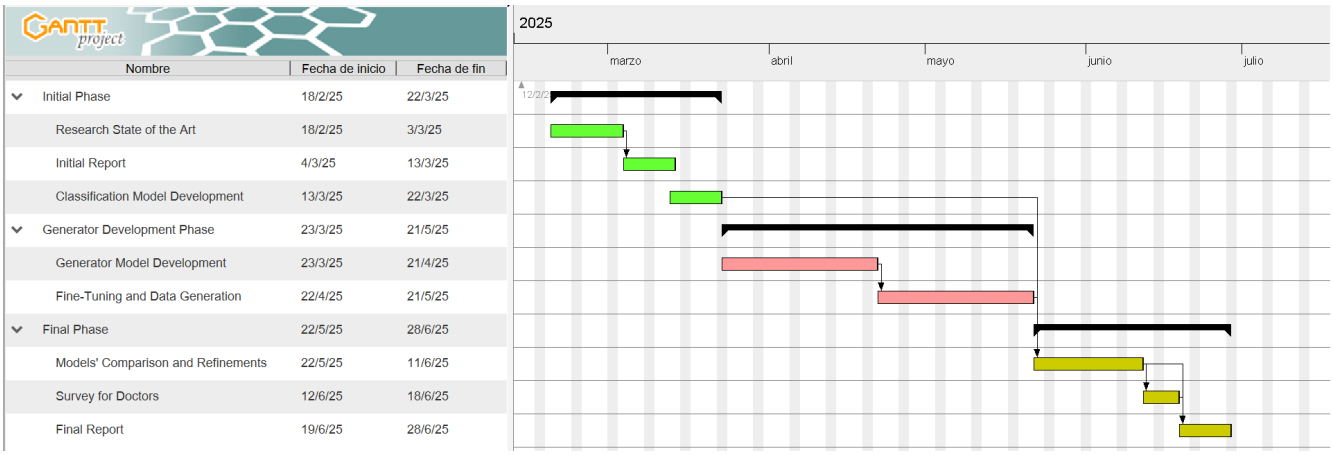


Fig. 35: Diagram illustrating the LoRA fine-tuning process: during training (left), two low-rank matrices are learned and summed up with the frozen original weight matrix; after training (right), these matrices are merged into a single adapted weight matrix for inference (13).



Nombre	Fecha de inicio	Fecha de fin
Initial Phase	18/2/25	22/3/25
Research State of the Art	18/2/25	3/3/25
Initial Report	4/3/25	13/3/25
Classification Model Development	13/3/25	22/3/25
Generator Development Phase	23/3/25	21/5/25
Generator Model Development	23/3/25	21/4/25
Fine-Tuning and Data Generation	22/4/25	21/5/25
Final Phase	22/5/25	28/6/25
Models' Comparison and Refinements	22/5/25	11/6/25
Survey for Doctors	12/6/25	18/6/25
Final Report	19/6/25	28/6/25

Fig. 36: Gantt chart illustrating planned phases and their deadlines.