

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Improved Synthetic Generation of Medical Image Data with Controllable Diffusion Models

Francisca Lopes Lúcio



Mestrado em Engenharia Biomédica

Supervisor: Luís Filipe Caeiro Margalho Guerra Rosado

Co-Supervisor: Ana Maria Mendonça

October 29, 2025

Resumo

A retinopatia diabética (RD) é uma das principais causas de cegueira a nível mundial, e o diagnóstico automatizado depende de bases de dados extensas, equilibradas e devidamente anotadas. Contudo, os dados clínicos reais apresentam frequentemente limitações relacionadas com escassez, desbalanceamento das classes e preocupações de privacidade. Esta dissertação explora a geração sintética de imagens de fundo de retina recorrendo a um modelo de difusão controlável, com o objetivo de ultrapassar estas limitações sem comprometer a relevância clínica.

O modelo Stable Diffusion foi combinado com a ControlNet, para permitir síntese específica por classe, utilização de máscaras de segmentação e descrições associadas à doença. Este processo de geração controlada possibilitou a criação de imagens que mantêm a estrutura da retina, ao mesmo tempo que representam níveis de gravidade da RD menos representados nos dados originais. A avaliação das imagens sintéticas seguiu uma abordagem multidimensional, considerando fidelidade, diversidade e utilidade. Em particular, a utilidade foi medida através do fine-tuning do modelo fundacional RETFound utilizando os datasets APTOS 2019, IDRiD, e Messidor-2, avaliando de forma prática o impacto da augmentação sintética no desempenho da classificação da RD.

Os resultados indicaram que a utilização de imagens sintéticas trouxe melhorias no APTOS 2019, sobretudo na capacidade de reconhecer classes minoritárias, e proporcionou ganhos moderados mas relevantes no IDRiD. Já no caso do Messidor-2, a integração de imagens sintéticas não resultou em benefícios e, por vezes, gerou confusão entre níveis de severidade próximos, reforçando que o impacto da geração sintética depende fortemente das características do dataset.

A análise de qualidade evidenciou uma elevada diversidade e boa perceção visual em todos os conjuntos de dados. Contudo, a fidelidade apresentou variações, esta foi mais elevada no IDRiD, intermédia no APTOS e mais reduzida no Messidor-2. Estes resultados mostram que a utilização de mecanismos de controlo pode potenciar a qualidade da geração, mas também sublinham que a eficácia da síntese depende de um ajuste cuidadoso dos parâmetros e das especificidades de cada dataset.

Em conclusão, este trabalho demonstra que a geração sintética baseada em modelos de difusão, potenciada por mecanismos de controlo, constitui uma abordagem promissora para a augmentação de imagens médicas.

Palavras-chave: Dados Sintéticos, Imagens de Fundo de Retina, Stable Diffusion, ControlNet, RETFound, Retinopatia Diabética

Abstract

Diabetic retinopathy (DR) is a leading cause of blindness worldwide, with automated diagnosis relying on large, balanced, and well-annotated datasets. However, real clinical data are often limited by scarcity, class imbalance, and privacy concerns. This thesis investigates the synthetic generation of retinal fundus images using a controllable diffusion model, aiming to address these challenges while preserving clinical relevance.

A Stable Diffusion model was combined with ControlNet conditioning to enable class-specific synthesis guided by anatomical masks and disease-related prompts. This controllable generation framework allowed the production of images that preserved retinal structure while targeting under-represented DR severity levels. The evaluation of synthetic images followed a multidimensional approach, incorporating fidelity, diversity, and utility. In particular, utility was assessed through fine-tuning the RETFound foundation model on APTOS 2019, IDRiD, and Messidor-2 datasets, providing a practical measure of the contribution of synthetic augmentation to DR classification.

Results showed that synthetic augmentation improved performance in APTOS 2019, particularly for minority classes, and achieved modest but meaningful gains in IDRiD. For Messidor-2, however, augmentation did not yield benefits and occasionally led to confusion between adjacent severity levels, highlighting the dataset-specific nature of outcomes.

Quality assessment confirmed high diversity and perceptual quality across all datasets, though fidelity varied. Fidelity was higher in IDRiD, moderate in APTOS, and lower in Messidor-2. These findings demonstrate that controllability enhances generative performance, but also emphasize that the effectiveness of synthetic augmentation depends on careful parameter tuning and dataset characteristics.

In conclusion, this work shows that diffusion-based synthetic generation, enhanced by controllability mechanisms, represents a promising direction for medical image augmentation.

Keywords: Synthetic Data Generation, Retinal Imaging, Stable Diffusion, ControlNet, RETFound, Diabetic Retinopathy

Acknowledgements

I would like to express my deepest gratitude to all those who made this dissertation possible.

First and foremost, I am grateful to my supervisor, Luís Rosado, for the opportunity, trust, and continuous guidance throughout this journey. I also want to thank him for his support and patience, which were essential to the development of this work. I extend my sincere thanks to my co-supervisor, Maria Mendonça, for her invaluable feedback and expertise, which ensured that the generated images achieved the highest possible quality. My gratitude also goes to Ivo Façoco for his assistance and for the foundational work that served as the backbone of my project, without his contributions, this work would not have been possible.

To my family, I owe my deepest thanks; to my parents for their unconditional love and support, and to my brother for his constant encouragement.

Finally, I would like to thank my friends, especially those who took me dancing, who made this journey all the more beautiful.

Francisca Lúcio

Images see with the eyes of those who see them.
José Saramago

Contents

Abbreviations	xiii
1 Introduction	1
1.1 Context and Motivation	1
1.2 Objectives	2
1.3 Overview	2
2 Diabetic Retinopathy	4
2.1 Disease Characterization	4
2.2 Diagnosis and levels of Diabetic Retinopathy	4
2.2.1 Diabetic Retinopathy Scale	5
2.3 Retinal Lesions Associated with Diabetic Retinopathy	6
2.4 Summary and Conclusions	10
3 Synthetic Medical Imaging: Generative Model Approaches	12
3.1 Synthetic Medical Images	12
3.2 Generative models	13
3.2.1 Variational Autoencoders	13
3.2.2 Generative Adversarial Networks	14
3.2.3 Diffusion Models	14
3.3 Strategies to Control the Diffusion Generation Process	16
3.3.1 Training Models from Scratch	16
3.3.2 Fine-Tuning Pre-Trained Models	16
3.3.3 Modifying the Sampling Process	17
3.3.4 Input-Based Conditioning	17
3.3.5 Auxiliary Networks for Conditional Control	18
3.4 Summary and Conclusions	18
4 Evaluation of Synthetic Retinal Fundus Images	20
4.1 Quality Evaluation of Synthetic Retinal Fundus Images	20
4.1.1 Fréchet Inception Distance (FID)	21
4.1.2 GMM-GIQA: Gaussian Mixture Model-based Quality Assessment	21
4.1.3 IP&IR: Improved Precision and Recall	22
4.2 Utility of Synthetic Retinal Fundus Images	22
4.2.1 Evaluating the Utility of Synthetic Retinal Fundus Images Using RETFound	23
4.2.2 Classification Evaluation Metrics	24
4.3 Summary and Conclusions	25

5	Generating Synthetic Retinal Fundus Images	26
5.1	Datasets Processing	26
5.1.1	Datasets Description	27
5.2	Stable Diffusion for Retinal Fundus Image Generation	31
5.3	ControlNet for DR Retinal Fundus Image Generation	31
5.3.1	ControlNet Inference Parameters	32
5.4	Summary and Conclusions	32
6	RETFound for Utility Assessment of Synthetic Data	36
6.1	Validation via RETFound Fine-Tuning	36
6.1.1	Image Preprocessing for RETFound Fine-Tuning	37
6.2	Summary and Conclusions	38
7	Experiments	39
7.1	ControlNet Fine-Tuning Experiments	39
7.1.1	Negative Prompts	39
7.1.2	Parameter Variations	40
7.2	RETFound Fine-Tuning Experiments	44
7.2.1	Image Preprocessing	44
7.2.2	Fine-Tuning and Evaluation	45
7.2.3	Complementary Experiments	45
7.3	Summary and Conclusions	45
8	Results and Discussion	48
8.1	APTOS 2019 Dataset	48
8.1.1	Confusion Matrices Analysis	48
8.2	IDRiD Dataset	50
8.2.1	Confusion Matrices Analysis	50
8.3	Messidor-2 Dataset	52
8.3.1	Confusion Matrices Analysis	52
8.4	Summary and Conclusions	53
9	Conclusion	59
9.1	Future Work	60
	References	61

List of Figures

2.1	Retina Fundus photograph with no evidence of disease or abnormality. While taking the photograph, the gaze is directed straight into the camera, positioning the macula centrally in each image, while the optic disc is oriented toward the nasal side. Image from Messidor dataset [15].	5
2.2	Representation of Diabetic Retinopathy Severity Levels and Associated Retinal Lesions. This figure illustrates the progression of diabetic retinopathy, highlighting the characteristic lesions present at each severity level, from mild NPDR to PDR, including microaneurysms, hemorrhages, hard exudates and neovascularization. Images adapted from FGADR dataset [98]	7
	(a) No DR	7
	(b) Mild NPDR	7
	(c) Moderate NPDR	7
	(d) Severe NPDR	7
	(e) PDR	7
2.3	Visualization of Retinal Lesion Distribution and Quadrant-Based Analysis. This figure illustrates the division of the retina into quadrants and demonstrates the spatial distribution of key retinal lesions associated with diabetic retinopathy, demonstrating their localization patterns in Severe NPDR and PDR levels. Image (a): Adapted from [72]	9
	(a) Division of the retina into quadrants.	9
	(b) Example of distribution and Localization of Microaneurysms in Retinal Fundus photograph with a severity level of Severe NPDR.	9
	(c) Example of distribution and Localization of Hemorrhages in Retinal Fundus photograph with a severity level of Severe NPDR.	9
	(d) Example of distribution and Localization of Hard Exudates in Retinal Fundus photograph with a severity level of PDR.	9
2.4	Retina Fundus photograph with a severity level of PDR. The image highlights characteristic lesions associated with this advanced stage of the disease, including neovascularization at the optic disc, and the other type of lesions described. Image adapted from FGADR dataset [98].	10
3.1	Stable Diffusion Architecture. The encoder (E) compresses the input image (x) into a latent representation (z), which undergoes a diffusion process in the latent space to create (z_T). A U-Net is used to iteratively denoise and refine the latent representation, reconstructing intermediate states (z_{T-1}) until the final image ($\hat{x}$) is generated. The decoder (D) then converts the refined latent representation back into the image space, completing the generation process. Image adapted from [68]	16

3.2	Architecture of Stable Diffusion with ControlNet. The process begins with a preprocessor that extracts conditioning information (c_i) from an input image, which is then transformed into feature representations (c_f) by the encoder (ε). A user-defined text prompt and time step are provided as inputs to the Stable Diffusion U-Net, along with the latent input (z_t). ControlNet refines the diffusion process by leveraging the conditioning features (c_f), enabling precise control over the generated image. The output latent representation (z_{t-1}) is progressively denoised to produce the final synthetic image. Image adapted from [80]	19
4.1	Performance of the RETFound model in DR classification. The model was fine-tuned on the Kaggle APTOS-2019, IDRID, and MESSIDOR-2 datasets, using AUROC as the evaluation metric. The results presented compare RETFound with other models, including SL-ImageNet, SSL-ImageNet, and SSL-Retinal. RETFound consistently outperforms the other models, demonstrating its superior performance and generalizability in DR classification from retinal images. Image from [99]	24
5.1	Preprocessing pipeline of retinal fundus images. This figure illustrates the sequence of operations applied to standardize the images, extract anatomical masks, and generate conditioning inputs for the generative model.	28
	(a) Original Image	28
	(b) Retinal Contour Extraction	28
	(c) CenterCrop and Resize	28
	(d) Optic Disc Segmentation	28
	(e) Conditioning Image	28
5.2	Example of synthetic data generation for class balancing. Additional images are generated from the original retinal fundus image to represent the other diabetic retinopathy severity classes, ensuring a more uniform class distribution across the dataset.	29
	(a) Original Image, Moderate Class	29
	(b) Synthetic Image, No DR	29
	(c) Synthetic Image, Mild DR	29
	(d) Synthetic Image, Severe DR	29
	(e) Synthetic Image, Proliferative DR	29
5.3	Class distribution before and after synthetic augmentation for each dataset. These bar plots illustrate the effect of synthetic image generation in balancing diabetic retinopathy severity classes across the three datasets, mitigating the original class imbalance present in the real data.	30
	(a) APTOS2019: 3662 real images; 14336 synthetic images.	30
	(b) IDRiD: 745 real images; 1862 synthetic images.	30
	(c) Messidor-2: 1937 real images; 6812 synthetic images	30
5.4	Pipeline of the Stable Diffusion model for retinal fundus image generation. Pretrained on the EyePACS and AIROGS datasets, the model receives as inputs latent representations of retinal images and is guided by textual prompts specifying disease labels (DR or glaucoma) and image quality attributes (low/medium/high brightness and contrast). These conditioning signals direct the generation process, producing synthetic images with controlled fidelity and diversity that reflect both the pathological characteristics and the desired quality parameters. Image adapted from [81]	34

5.5	Pipeline of the ControlNet-based synthetic image generation process. Preprocessed retinal fundus images serve as the inputs, while masks (retinal contour and optic disc segmentation) are incorporated as structural conditioning. Text prompts specifying the target DR severity class guide. In the img2img framework, ControlNet integrates these conditioning signals with the diffusion model to generate anatomically consistent synthetic images that reflect the assigned DR severity grade. Image adapted from [80]	35
6.1	Pipeline overview of RETFound fine-tuning for diabetic retinopathy classification. The figure illustrates the input steps, into the RETFound masked autoencoder with an MLP classification head." Image adapted from [94]	37
7.1	Examples of visual anomalies observed in synthetic images, compared with their counterparts obtained after reducing strength. Subfigures (a) and (b) show samples generated from the IDRiD dataset when synthesizing the No DR class from the same real Moderate DR image, while subfigures (c) and (d) illustrate the synthesis of No DR images from the same real Severe DR image. The reduction of the strength parameter mitigates the presence of artifacts, resulting in more realistic retinal fundus images.	41
	(a) Synthetic image with anomaly, generated from Moderate DR to No DR	41
	(b) Synthetic image after lowering strength, from same input as (a)	41
	(c) Synthetic image with anomaly, generated from Severe DR to No DR.	41
	(d) synthetic image after lowering strength, from same input as (c)	41
7.2	Examples from the APTOS dataset illustrating the effect of increasing the number of inference steps with the goal of enhancing perceptual quality and fidelity of synthetic retinal fundus images. All images were generated by conditioning from Moderate DR to Proliferative DR. Images (a) and (b) derived from the same real moderate DR image (a = lower steps, b = higher steps); Images (c) and (d) derived from the same real moderate DR image (c = lower steps, d = higher steps). For each subfigure the reported values (Precision, Recall, GIQA) allow quantitative comparison of image quality across configurations.	42
	(a) Synthetic image generated with lower inference steps. Precision = 1, Recall = 0, GIQA = 0,7951.	42
	(b) Synthetic image obtained from the same original input as (a), using higher inference steps. Precision = 1, Recall = 1, GIQA = 0.8154	42
	(c) Synthetic image generated with lower inference steps. Precision = 1, Recall = 0, GIQA = 0.9306.	42
	(d) Synthetic image obtained from the same original input as (c), using higher inference steps. Precision = 1, Recall = 1, GIQA = 0.9468	42
7.3	Examples of synthetic retinal fundus images for quality-filtered augmentation. Images generated from the Messidor dataset. These examples illustrate the differences in image quality considered during the augmentation process, highlighting the visual criteria behind the filtering strategy.	46
	(a) Synthetic image with IP = 0	46
	(b) Synthetic image with IP = 0	46
	(c) Synthetic image with IP = 0	46
	(d) Synthetic image with IP = 1	46
	(e) Synthetic image with IP = 1	46
	(f) Synthetic image with IP = 1	46

8.1	Confusion matrices for the APTOS 2019 dataset showing the best results. Class indices are defined as follows:0 = No DR; 1 = Mild DR; 2 = Moderate DR; 3 = Severe DR; 4 = Proliferative DR.	49
	(a) Real data	49
	(b) Real data + Synthetic images	49
8.2	Confusion matrices for the IDRiD dataset showing the best results. Class indices are defined as follows:0 = No DR; 1 = Mild DR; 2 = Moderate DR; 3 = Severe DR; 4 = Proliferative DR.	51
	(a) Real data	51
	(b) Real data + Synthetic images	51
8.3	Confusion matrices for the Messidor-2 dataset showing the best results. Class indices are defined as follows:0 = No DR; 1 = Mild DR; 2 = Moderate DR; 3 = Severe DR; 4 = Proliferative DR.	52
	(a) Real data	52
	(b) Real data + Synthetic images	52
8.4	UMAP Visualization of Real vs Synthetic Features. This figure demonstrates the UMAP visualization of feature embeddings for real and synthetic retinal fundus images. Real samples are represented in the color blue and synthetic samples in orange, highlighting the overlap and alignment between datasets. This analysis supports the evaluation of fidelity and diversity in synthetic data generation and aids in identifying discrepancies or gaps in feature representation.	54
	(a) APTOS 2019	54
	(b) IDRiD	54
	(c) Messidor-2	54
8.5	Illustrative example of synthetic retinal fundus generation from APTOS 2019 dataset. Real fundus image with its respective class, together with the synthetic images generated from it. Each image is annotated with its class label and associated evaluation metrics (precision, recall, and GIQA), enabling a direct assessment of visual quality.	55
	(a) Original Image (Proliferative DR)	55
	(b) Synthetic (No DR) Precision=1, Recall=1, and GIQA=0.9639	55
	(c) Synthetic (Mild) Precision=1, Recall=0, and GIQA=0.9602	55
	(d) Synthetic (Moderate DR) Precision=0, Recall=0, and GIQA=0.9188	55
	(e) Synthetic (Severe DR) Precision=0, Recall=0, and GIQA=0.9457	55
8.6	Illustrative example of synthetic retinal fundus generation from IDRiD dataset. Real fundus image with its respective class, together with the synthetic images generated from it. Each image is annotated with its class label and associated evaluation metrics (precision, recall, and GIQA), enabling a direct assessment of visual quality.	56
	(a) Original Image (Moderate DR)	56
	(b) Synthetic (No DR) Precision=1, Recall=1, and GIQA=0.7639	56
	(c) Synthetic (Mild DR) Precision=1, Recall=1, and GIQA=0.7359	56
	(d) Synthetic (Severe DR) Precision=1, Recall=1, and GIQA=0.7656	56
	(e) Synthetic (Proliferative DR) Precision=1, Recall=1, and GIQA=0.6548	56

8.7	Illustrative example of synthetic retinal fundus generation from Messidor-2 dataset. Real fundus image with its respective class, together with the synthetic images generated from it. Each image is annotated with its class label and associated evaluation metrics (precision, recall, and GIQA), enabling a direct assessment of visual quality.	57
(a)	Original Image (No DR)	57
(b)	Synthetic (Mild) Precision=0, Recall=0, and GIQA=0.8753	57
(c)	Synthetic (Moderate DR) Precision=0, Recall=0, and GIQA=0.73258	57
(d)	Synthetic (Severe DR) Precision=0, Recall=0, and GIQA=0.7217	57
(e)	Synthetic (Proliferative DR) Precision=0, Recall=0, and GIQA=0.6818	57

List of Tables

2.1	Characterization of Lesions by DR Levels and Retinal Region	8
5.1	Distribution of DR Classes Across Datasets	31
6.1	Fine-tuning parameters used for RETFound in this study.	37
7.1	Initial inference parameters for ControlNet fine-tuning.	39
7.2	Modified inference parameters.(Lowering the strength parameter)	40
7.3	Modified inference parameters. (Increasing the number of inference steps)	40
7.4	Modified inference parameters. Increasing the guidance scale.	43
7.5	Modified inference parameters for IDRiD	43
7.6	Modified inference parameters for Messidor-2	43
7.7	Distribution of samples across training, validation and test splits for APTOS 2019, IDRiD and Messidor-2.	44
8.1	Classification performance of RETFound fine-tuning on the APTOS 2019 dataset using real data only versus real plus synthetic data. The best results are highlighted in bold.	50
8.2	Quality evaluation of synthetic images generated for the APTOS 2019 dataset us- ing pyMDMA metrics.	50
8.3	Classification performance of RETFound fine-tuning on the IDRiD dataset using real data only versus real plus synthetic data. The best results are highlighted in bold.	51
8.4	Quality evaluation of synthetic images generated for the IDRiD dataset using pyMDMA metrics.	51
8.5	Classification performance of RETFound fine-tuning on the Messidor-2 dataset using real data only versus real plus synthetic data. The best results are highlighted in bold.	52
8.6	Quality evaluation of synthetic images generated for the Messidor-2 dataset using pyMDMA metrics.	53

Abbreviations and Symbols

CVAE	Conditional Variational Autoencoder
DDPM	Denoising Diffusion Probabilistic Model
DM	Diffusion Model
DMO	Diabetic Macular Odema
DR	Diabetes Retinopathy
GAN	Generative Adversarial Network
ML	Machine Learning
NPDR	Non-proliferative Diabetic Retinopathy
NVD	Neovascularisation at the Optic Disc
NVE	Neovascularisation of the retina elsewhere
PDR	Proliferative Diabetic Retinopathy
SD	Stable Diffusion
VAE	Variational Autoencoder
VLM	Visual Language Model

Chapter 1

Introduction

This thesis explores the synthetic generation of retinal fundus images using a controllable diffusion model, specifically targeting diabetic retinopathy. The primary aim is to develop high-quality synthetic retinal images that not only mirror the visual and statistical properties of real datasets but also contribute to enhancing predictive ML models for Diabetic Retinopathy detection and classification. Achieving high fidelity and diversity in synthetic images is essential for their effective use. Fidelity ensures that synthetic images accurately preserve structural, textural, and pathological characteristics of real data, while diversity promotes robustness by representing a wide range of variations. This balance is critical, as excessive focus on fidelity may limit representativeness, whereas overly diverse datasets risk compromising realism. It is important to explore these trade-offs, aiming to optimize both of these parameters. Furthermore, this work evaluates the utility of these synthetic images in clinical applications, such as training deep learning models for disease detection.

1.1 Context and Motivation

Diabetic retinopathy (DR) is a leading cause of adult blindness worldwide [23], among individuals with diabetes. The early detection and accurate grading of DR are crucial for preventing vision loss and improving patient outcomes. Retinal fundus images, essential for diagnosing and monitoring DR, present significant challenges in terms of both acquisition and analysis. As the demand for large-scale, annotated datasets rises, the need for synthetic data generation, particularly through controllable methods, has emerged as a promising solution. Synthetic retinal fundus images can augment existing datasets, address class imbalances, and provide data diversity necessary for robust model training. Furthermore, these synthetic images can be generated to mimic the specific pathological lesions found in real retinal fundus images, such as microaneurysms, hemorrhages, and exudates. This approach holds the potential to overcome the limitations of traditional data collection, enabling the creation of large-scale, high-quality datasets for the development of automated DR detection systems.

1.2 Objectives

The main contributions of this dissertation are the following:

- **Controllable Generation of Synthetic Retinal Fundus Images:** This work develops a methodology to generate synthetic retinal fundus images for diabetic retinopathy (DR) using a controllable diffusion model. By integrating ControlNet [95], the generation process can be conditioned on anatomical structures and DR-specific prompts. This allows precise control over image characteristics, ensuring that synthetic images preserve the structural, textural, and pathological features of real data while capturing diverse variations observed across different DR severity levels. The approach is particularly aimed at producing images that help balance underrepresented classes in existing datasets.
- **Evaluation of Synthetic Image Quality and Utility:** The fidelity, diversity, and clinical relevance of the synthetic images are rigorously assessed. Fidelity ensures that synthetic images realistically reflect real retinal structures and lesions, while diversity provides a broad coverage of variations necessary for robust model training. Utility is evaluated by fine-tuning the RETFound [99] foundation model, quantifying the impact of synthetic augmentation on DR classification performance. This step demonstrates whether the inclusion of controlled synthetic data improves model robustness, particularly for minority DR classes, and supports the development of more reliable automated diagnostic tools.

The ultimate aim is to optimize the fidelity, diversity, and clinical utility of synthetic retinal fundus images for diabetic retinopathy diagnosis, contributing to the development of more accurate and scalable diagnostic tools.

1.3 Overview

In addition to the introduction, this dissertation contains 5 chapters.

Chapter 2, provides an overview of diabetic retinopathy, focusing on its clinical characteristics, diagnostic criteria, classification, and the retinal lesions associated with the disease. It establishes a foundational understanding of DR.

Chapter 3, explores the critical role of synthetic medical imaging and examines various generative models that enable the controlled synthesis of medical images.

Chapter 4 focuses on the evaluation of the synthetic images. It includes an analysis of the fidelity and diversity and discusses the clinical utility of the synthetic data, particularly in relation to their impact on the performance of models for DR classification.

Chapter 5 details the methodology for generating synthetic retinal fundus images. It describes the preprocessing steps, the use of Stable Diffusion as the generative backbone, and the integration of ControlNet to enable class-specific synthesis, addressing the problem of class imbalance in real-world datasets.

Chapter 6 focuses on the validation of synthetic images through their integration into diabetic retinopathy classification tasks. By fine-tuning RETFound, the chapter evaluates the clinical utility of synthetic augmentation and quantifies its impact on model generalization.

Chapter 7 outlines the experimental setup, covering both the fine-tuning of the ControlNet pipeline for synthetic image generation and the fine-tuning of RETFound for classification. The design ensures a systematic assessment of image quality and downstream performance.

Chapter 8 presents the results and discussion of the experiments. It reports quantitative findings for APTOS 2019, IDRiD, and Messidor-2, complemented by visual examples of synthetic images.

Finally, Chapter 9 provides the conclusions of the thesis, summarizing the key findings and contributions. It also discusses outlines directions for future work.

Chapter 2

Diabetic Retinopathy

This chapter provides an overview of diabetic retinopathy (DR), focusing on its diagnostic criteria, disease classification, and characteristic retinal lesions. The aim is to establish a foundational understanding of DR's clinical aspects, which is essential to guide the synthetic generation of realistic retinal images with diabetic retinopathy, which will be explored in later chapters.

2.1 Disease Characterization

Diabetic Retinopathy (DR) is a severe complication of diabetes mellitus that results in damage to the retina due to microvascular alterations. As a microangiopathy, DR involves the deterioration of small blood vessels in the retina, leading to various forms of retinal lesions. In its early stages, DR is often asymptomatic, where affected individuals may not experience any noticeable changes in vision. However, as the disease progresses to more advanced stages, symptoms such as blurred or fluctuating vision become evident, at which point the condition may be irreversible and could lead to blindness. Consequently, DR is a significant public health concern, being one of the leading causes of vision loss among adults [23]. Regular screening of individuals with diabetes is essential to detect and manage DR in its earlier stages, thereby reducing the risk of sight-threatening complications and improving patient outcomes.

2.2 Diagnosis and levels of Diabetic Retinopathy

The diagnosis of DR is primarily performed through fundus photography using a fundus camera, which captures detailed images of the retina, including the optic disc, fovea, and retinal blood vessels (Figure 2.1). These images can be obtained by a non-specialist, however, the clinical diagnosis and characterization of DR is subsequently conducted by a specialist, who assesses retinal changes, identifying potential abnormalities based on a fundoscopic examination. Comprehensive and systematic recording of these diagnostic findings allows clinicians to compare retinal images across time points, enabling effective monitoring of disease progression in individual patients.

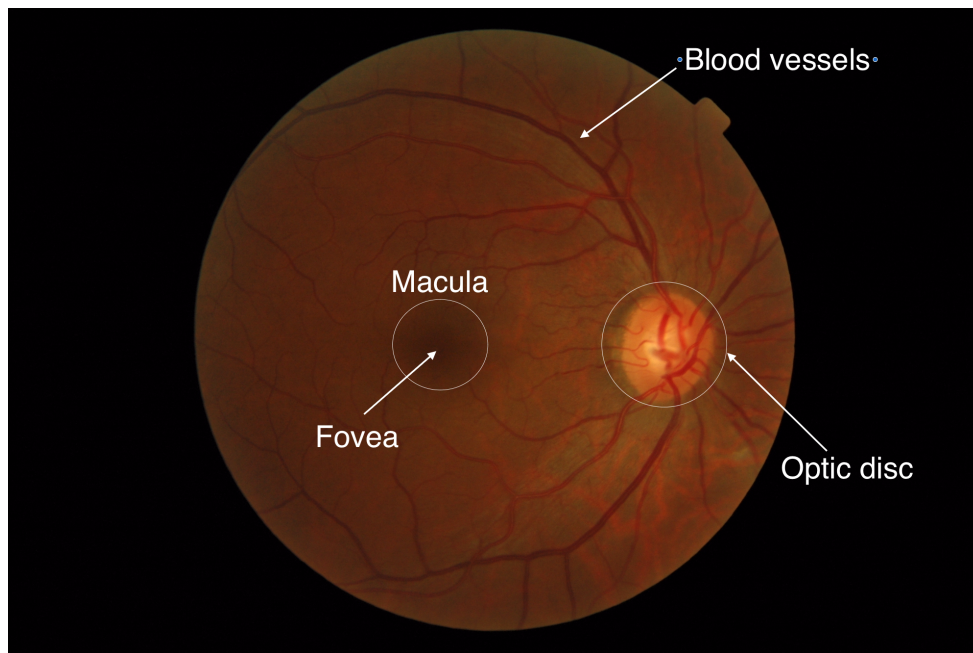


Figure 2.1: **Retina Fundus photograph with no evidence of disease or abnormality.** While taking the photograph, the gaze is directed straight into the camera, positioning the macula centrally in each image, while the optic disc is oriented toward the nasal side. Image from Messidor dataset [15].

Retinal images are then graded according to findings that correspond to specific levels of retinopathy, following internationally recognized disease severity scales [87] for diabetic retinopathy and diabetic macular oedema (DMO) [25].

2.2.1 Diabetic Retinopathy Scale

DR is classified into two main stages: non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR). **NPDR** is further subdivided into four levels of severity:

1. **No Retinopathy:** No observable abnormalities.
2. **Mild NPDR:** Presence of microaneurysms only.
3. **Moderate NPDR:** More than just microaneurysms but without severe characteristics.
4. **Severe NPDR:** Presence of one or more of the following in the absence of proliferative features:
 - Over 20 intraretinal hemorrhages in each of the four retinal quadrants;
 - Definite venous beading in two or more quadrants;
 - Prominent intraretinal microvascular abnormalities in at least one quadrant.

The **proliferative stage (PDR)** is characterized by one or more of the following:

- Extraretinal neovascularization (either neovascularisation at the optic disc (NVD) at or neovascularisation of the retina elsewhere (NVE));
- Vitreous or preretinal hemorrhage.

Diabetic macular oedema (DMO) can occur at any stage of DR and is characterized by retinal thickening in the macula. DMO is classified into:

1. **Mild DMO:** Retinal thickening or hard exudates in the posterior pole, distant from the macula's center.
2. **Moderate DMO:** Retinal thickening or hard exudates near the macula's center but not involving it.
3. **Severe DMO:** Retinal thickening or hard exudates involving the center of the macula.

In Figure 2.2, is evident that distinguishing between the different stages of diabetic retinopathy is challenging without a thorough understanding of the specific lesions that characterize each stage. Therefore, the following section provides a detailed description of these lesion types.

2.3 Retinal Lesions Associated with Diabetic Retinopathy

In NPDR, characteristic retinal lesions include microaneurysms, dot and blot intraretinal hemorrhages, hard exudates, cotton wool spots, and venous beading. In contrast, PDR is marked by the development of new blood vessels on the optic disc and other areas of the retina, and vitreous and preretinal hemorrhages [25].

Microaneurysms are small, circular dilations of retinal capillaries caused by the weakening of blood vessel walls due to diabetes [76]. They appear as tiny red dots [6] with diameters between 10 and 100 μm [84] and are often the earliest detectable sign of DR. The number of microaneurysms correlates with the severity of NPDR, ranging from 0 to 5 in mild cases, 5 to 15 in moderate cases, and over 15 in severe cases [3]. Across all stages of DR, most microaneurysms are concentrated in the temporal and inferior areas of the retina [72].

Dot and blot intraretinal hemorrhages occur when microaneurysms rupture within the deeper retinal layers [1], appearing as dense, dark red spots [40]. These hemorrhages are typically larger than microaneurysms [24], often equal to or larger than a blood vessel at the disk margin (from 228.93 μm to 243.18 μm) [49], with uneven or indistinct edges. In mild NPDR, hemorrhages are typically absent, whereas moderate NPDR cases present between 0 and 5 hemorrhages, and severe NPDR cases exhibit more than 5 [3]. The distribution of hemorrhages does not vary significantly by severity. For moderate, and severe DR, the temporal area shows the highest concentration of hemorrhages, followed by the inferior area [72].

Hard exudates consist of lipids and proteins that leak from damaged blood vessels, creating bright, reflective yellowish-white lesions in the retina [4, 37]. These lesions are absent in mild DR cases, while moderate cases typically present with 3 to 5 exudates, and severe cases exhibit 6 to

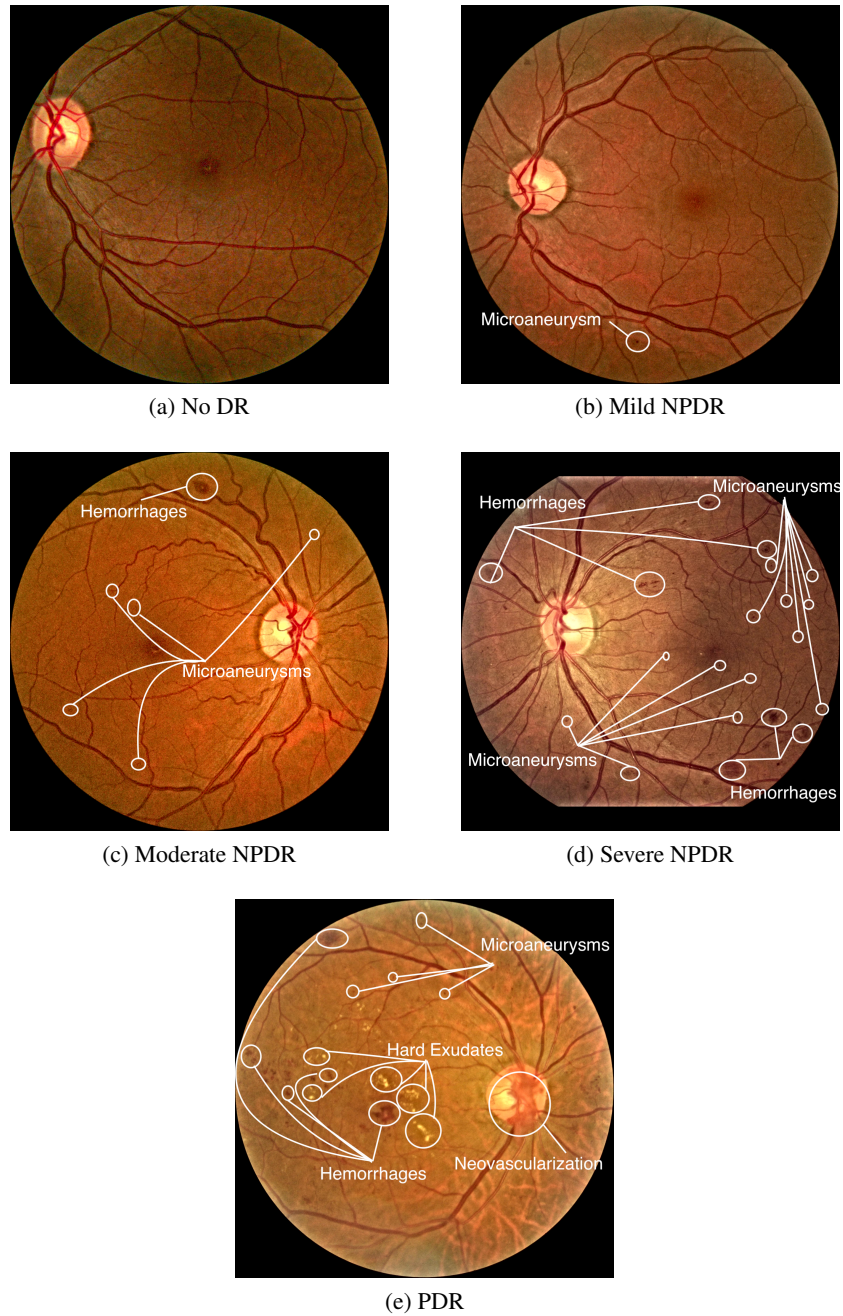


Figure 2.2: Representation of Diabetic Retinopathy Severity Levels and Associated Retinal Lesions. This figure illustrates the progression of diabetic retinopathy, highlighting the characteristic lesions present at each severity level, from mild NPDR to PDR, including microaneurysms, hemorrhages, hard exudates and neovascularization. Images adapted from FGADR dataset [98]

Table 2.1: Characterization of Lesions by DR Levels and Retinal Region

Lesion	DR Level	Average Number	Average Size	Spatial Distribution (by Retinal Region)
Microaneurysms	Mild	From 0 to 5	10–100 μm	Predominantly in the temporal region
	Moderate	From 5 to 15	10–100 μm	Predominantly in the temporal region
	Severe	Over 15	10–100 μm	Predominantly in the temporal and inferior regions
Hemorrhages	Mild	Absent	-	-
	Moderate	From 0 to 5	100 - 243 μm	Predominantly in the temporal region
	Severe	Over 5	100 - 243 μm	Predominantly in the temporal and inferior regions
Hard Exudates	Mild	Absent	-	-
	Moderate	From 0 to 5	Total area covered by hard exudates: 0.089 mm ²	Within 500 to 1500 μm from the foveal center (approaching the macula center)
	Severe	From 6 to 9	Total area covered by hard exudates: 0.501 mm ²	Within 500 μm from the foveal center (in the macula center)

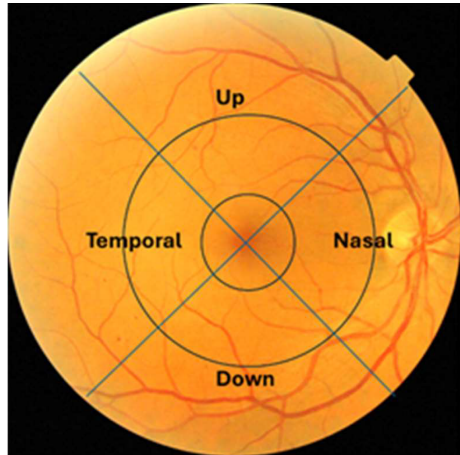
9 [22]. However, due to the challenges in detecting these lesions, there is considerable variability among patients. Potentially, for a more accurate characterization, the area occupied by these lesions should be assessed rather than their count [22]. In moderate DR, the total area occupied by hard exudates is, on average, 0.089 mm², with their distribution approaching the macula center, within 500 to 1500 μm from the foveal center. In severe DR, the total area occupied by hard exudates is, on average, 0.501 mm², with their distribution primarily concentrated in the macula center, within 500 μm from the foveal center [75].

For a better understanding of the regions where retinal lesions are concentrated, the division of the retina into quadrants is illustrated in Figure 2.3, highlighting the temporal area. For a comprehensive overview, Table 2.1, maps the average number, size, and distribution of lesions across different stages of DR, detailing their presence in specific retinal regions for each DR level.

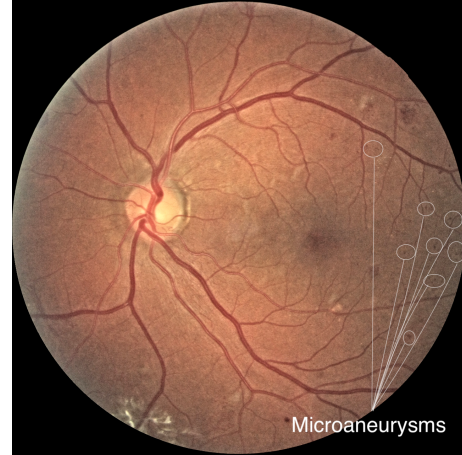
Cotton wool spots are pale, fluffy lesions caused by localized retinal ischemia due to arteriolar blockage, usually appearing as small gray-white patches with fimbriated edges [25].

Venous beading is characterized by segmental dilations of retinal veins, reflecting areas of venous endothelial proliferation and indicating a more advanced stage of NPDR [25].

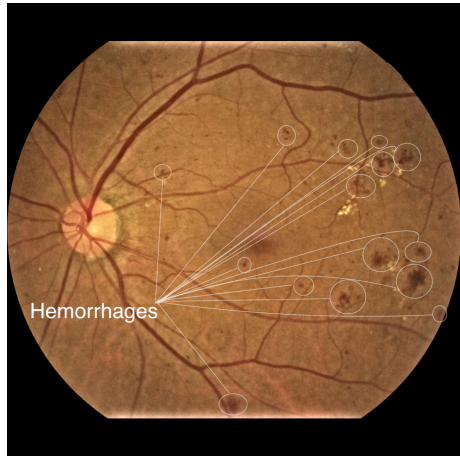
In PDR, **neovascularization** represents a severe retinal abnormality. This process involves the growth of new, fragile blood vessels as a compensatory response to severe hypoxia in retinal capillaries. These vessels are highly prone to rupture, posing a significant risk for vision loss. Detecting neovascularization is challenging, as it can resemble normal retinal vessels in imaging [32].



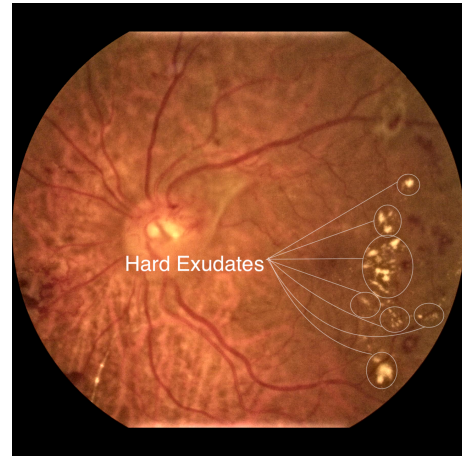
(a) Division of the retina into quadrants.



(b) Example of distribution and Localization of Microaneurysms in Retinal Fundus photograph with a severity level of Severe NPDR.



(c) Example of distribution and Localization of Hemorrhages in Retinal Fundus photograph with a severity level of Severe NPDR.



(d) Example of distribution and Localization of Hard Exudates in Retinal Fundus photograph with a severity level of PDR.

Figure 2.3: Visualization of Retinal Lesion Distribution and Quadrant-Based Analysis. This figure illustrates the division of the retina into quadrants and demonstrates the spatial distribution of key retinal lesions associated with diabetic retinopathy, demonstrating their localization patterns in Severe NPDR and PDR levels. Image (a): Adapted from [72]

; Images (b), (c), (d): Adapted from FGADR dataset [98].

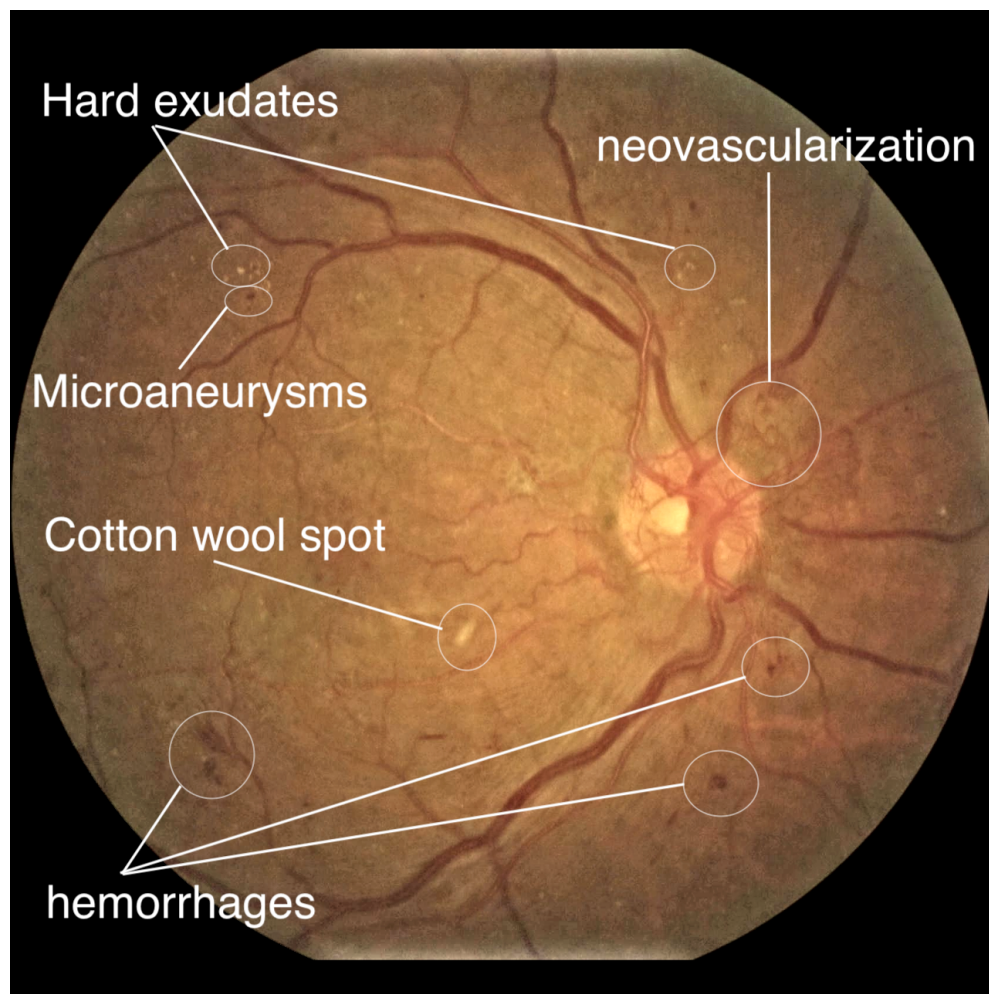


Figure 2.4: **Retina Fundus photograph with a severity level of PDR.** The image highlights characteristic lesions associated with this advanced stage of the disease, including neovascularization at the optic disc, and the other type of lesions described. Image adapted from FGADR dataset [98].

Vitreous hemorrhage occurs when blood from ruptured neovascular vessels enters the vitreous cavity, partially obscuring retinal visualization, often with a boat-shaped appearance in sub-hyaloid hemorrhages [39].

Preretinal hemorrhages occur when blood collects between the retina and vitreous cavity, often forming boat-shaped or D-shaped hemorrhages due to the force of the leakage against the retina [40].

In Figure 2.4 can be observed, the described types of lesions associated with Diabetic Retinopathy.

2.4 Summary and Conclusions

In summary, DR is a significant microvascular complication of diabetes mellitus that leads to progressive retinal damage, affecting vision. This chapter has outlined the clinical characteristics of

DR, emphasizing the importance of early diagnosis and regular screening to prevent severe visual impairment. The classification of DR into non-proliferative and proliferative stages, along with the identification of associated retinal lesions, provides a framework for understanding the disease's characteristics. Additionally, detailed insights into the average number, size, and distribution of lesions across different stages of DR, along with their presence in specific retinal regions, enhance our understanding of how these lesions manifest in the retina.

In conclusion, a thorough understanding of typical fundus images and retinal lesions associated with various stages of DR is crucial for creating this type of synthetic images. By integrating this knowledge, it becomes possible to ensure that the generated images closely mimic the characteristics of real-world data.

Chapter 3

Synthetic Medical Imaging: Generative Model Approaches

This chapter highlights the importance of synthetic images in advancing medical applications, particularly in supporting diagnostic processes where large and diverse datasets are essential. This chapter further explores key generative models—including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models (DMs)—that enable the controlled synthesis of medical images. This chapter also focuses on strategies to control the diffusion generation process. The aim is to understand how these generative techniques can be effectively applied to produce synthetic retinal fundus images, enhancing diagnostic support for diseases such as diabetic retinopathy.

3.1 Synthetic Medical Images

One of the most critical aspects of automating medical applications is the availability of high-quality, large-scale data. In the medical field, the success of machine learning applications largely depends on the availability of such training data [7]. This is further complicated due to data privacy regulations. Medical data is scarce because of its sensitive nature, where precision and accuracy in diagnosis are paramount, the collection and annotation of specific medical datasets is both time-consuming and expensive, requiring trained experts to annotate these images.

One approach to address the limited availability of medical image datasets is through data augmentation techniques. Traditional methods include transformations such as flipping, rotating or cropping. A more recent, state-of-the-art approach involves generating new data samples by synthesizing new images [48].

By creating synthetic medical images, it is possible to improve model training by adding diversity to the dataset, enabling the representation of varied scenarios and handling complex conditions. These synthetic images can also be controlled to simulate specific, relevant conditions. The primary advantage of this approach over traditional techniques lies in preserving patient privacy.

Synthetic datasets also facilitate the training process and reduce annotation costs. [66] demonstrated that their method reduced the annotation load by creating synthetic images leading to performance improvements with sinus surgery and intra-abdominal surgery datasets.

In the medical field, synthetic images have proven capable of enhancing classifier performance and generalization in pathology detection. [43] demonstrated that integrating advanced generative models with synthetic data could be a pivotal step in overcoming limitations in medical imaging datasets. Supplementing real datasets with synthetic data significantly enhanced model performance.

However, medical diagnosis often depends on subtle changes in the appearance of complex organs, making it challenging to generate meaningful medical data [59]. The same challenge arises in the diagnosis of diabetic retinopathy. The difference between a state of no disease and one that leads to blindness may be as subtle as neovascularization, a detail that synthesized images may fail to capture. One way to approach this problem is by training conditional models to enable controlled image synthesis, focusing on the fine details that characterize the disease.

3.2 Generative models

Generative models have been at the forefront of deep unsupervised learning for the last decade [62]. These models aim to capture the true data distribution and can generate unseen data from it. They have been employed to create high-quality images, and one of their key applications is in medical imaging [74, 35]. These algorithms produce detailed and accurate images based on descriptions in natural language [8].

One way to ensure high-quality and realistic images for specific pathological stages is by controlling the output towards the desired characteristics. This can be achieved by using conditional models for image synthesis. These models accept additional inputs that influence the generated content, allowing for the control of the features in the output. This input can vary widely, ranging from textual descriptions to segmentation masks [93]. Model architectures can be designed or adapted to incorporate this additional input, enabling content control during image generation.

These models can be classified into three main categories:

- **Variational Autoencoders (VAEs)**
- **Generative Adversarial Networks (GANs)**
- **Diffusion Models (DMs)**

3.2.1 Variational Autoencoders

VAEs [44] consist of an encoder that attributes probabilistic terms to an image, followed by a decoder that can randomly sample from the probability distributions of each attribute to generate a new image [18].

Kaplan et al. [41] generated retinal fundus images using VAEs, and their results showed that while the generated images preserved the global structure of the retina, they failed to generate the vessels. Additionally, the images appeared blurry.

Conditional Variational Autoencoders (CVAEs) are an extension of VAEs in which a conditional input is added. In the medical field, this allows for the differentiation of images from different classes of a pathology [85].

Pesteie et al. [67] demonstrated improved classification results by using CVAEs for data augmentation on ultrasound images of the spine and magnetic resonance images of the brain. This model enabled the expansion of the existing dataset, introducing more intra-class variance.

These types of models offer an advantage over Generative Adversarial Networks (GANs) due to their diverse outputs; however, the resolution of the generated images is not high-quality [83, 67], which can be problematic in the medical field, where small details can be critical and the ability to capture fine nuances is essential.

3.2.2 Generative Adversarial Networks

GANs [27] consist of a generator that synthesises data similar to the input and a discriminator that evaluates this data. Based on feedback from the discriminator, the generator minimizes the loss between the two distributions, progressively generating samples that resemble the input data [7].

A way to control the output of GANs is by adding conditional inputs to the model, creating a conditional GAN (cGAN) [38]. This allows for greater selectivity in which parts of the learned distribution are used to generate data, which can be especially useful in medical imaging [33].

Conditional GANs have been applied in various cases for generating retinal fundus images [92, 13, 78, 50]. Andreini et al. [5] proposed a method using GANs conditioned on the vascularization of retinal images, generating images with unprecedented high resolution and realism.

While GANs can generate high-resolution images, they often fail to capture the entire data distribution (mode-collapse), resulting in a lack of diversity in the generated images [43, 33].

3.2.3 Diffusion Models

A novel approach has emerged with diffusion models (DM), which are likelihood-based and capable of estimating data distributions and sampling quality more effectively than other methods [60]. These models have demonstrated superior performance over GANs due to their stability [16], attracting considerable attention for their potential to generate realistic and diverse high-resolution images [59].

Diffusion models operate by gradually adding noise to an image with increasing intensity until it is completely noise. A deep neural network, usually a U-Net, learns this diffusion process and is later employed to reverse it, progressively removing the noise step by step to generate a new image.

The synthesis of retinal fundus images through diffusion models has become a particularly active research area, motivated by both privacy concerns and the demand for large, diverse, and

high-quality datasets. One of the most significant advances is the SynFundus-1M dataset [77], comprising over one million synthetic fundus images generated with a Denoising Diffusion Probabilistic Model (DDPM) trained on 1.3 million authentic fundus images. SynFundus-1M achieved competitive FID scores and expert evaluations confirmed that the synthetic images were often indistinguishable from real ones.

Other works have demonstrated how conditioning mechanisms can further enhance structural and pathological fidelity. Ilanchezian et al. [36] developed a counterfactual generation framework, where conditioning on disease-specific attributes allowed the synthesis of plausible alternative scenarios, such as adding or removing pathological findings.

Several studies have also assessed the value of diffusion-augmented data for diabetic retinopathy (DR) detection and grading [10, 48]. Catapang et al. [10] introduced a framework, AffiDAVIT, which integrates diffusion-based data augmentation with vision transformers (ViT), for DR detection. Their study explored the effectiveness of Stable Diffusion for generating synthetic fundus images. They evaluated the utility of these synthetic images through qualitative assessments, showing that diffusion-augmented data improved test set performance of their ViT model, particularly in reducing false negatives and improving accuracy. However, when analyzed by retina specialists, the synthetic images were distinguishable from real ones.

Similarly, Lakas et al. [48] investigated the impact of diffusion-based augmentation on deep learning classifiers for DR grading using the IDRiD dataset. By comparing the performance of classifiers trained on datasets augmented with diffusion models versus traditional techniques, they demonstrated that diffusion augmentation led to superior accuracy in both ResNet-50 and CNN models. Their results highlight the potential of diffusion models to surpass conventional augmentation methods, particularly in enhancing the generalization capabilities and overall performance of machine learning models in medical image analysis.

3.2.3.1 Stable Diffusion

Stable Diffusion (SD) models represent a further evolution in diffusion-based generative techniques, combining the high-resolution image generation capabilities of diffusion models with advanced language comprehension features.

SD builds upon the principles of Latent Diffusion modeling [71]. In latent diffusion, the model is trained to generate compressed representations (latent data) of images. The generation process includes a Forward Diffusion Process, which progressively injects noise into the latent data over t steps, and a Reverse Diffusion Process, which systematically denoises the latent representations across the same t steps.

Figure 3.1 illustrates the process of generating synthetic retinal fundus images using the Stable Diffusion model.

Unlike earlier diffusion models primarily focused on visual data alone, Stable Diffusion utilizes Visual Language Models (VLMs) to interpret and generate images with nuanced understanding of semantic context [79]. The integration of VLMs has enabled SD to produce diverse and detailed images that align closely with descriptive inputs.

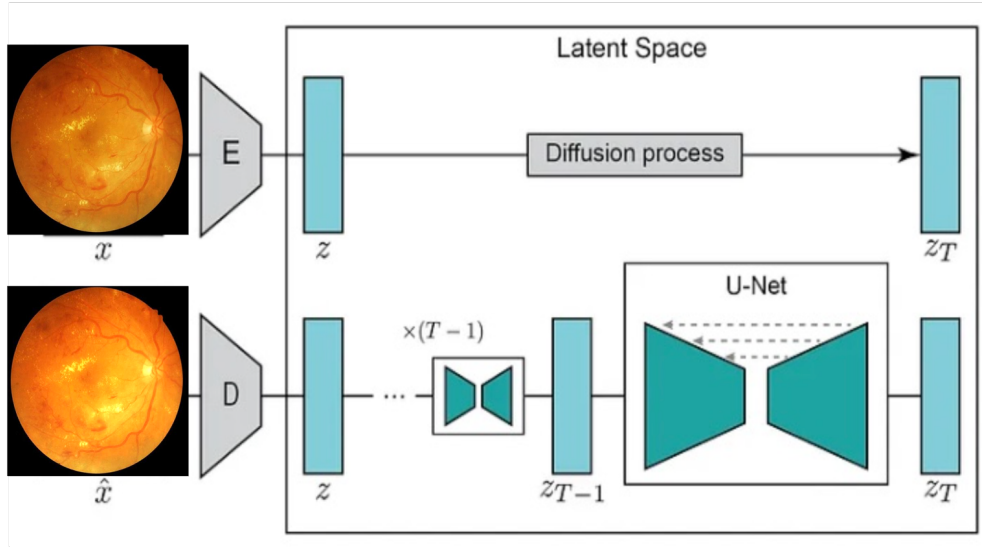


Figure 3.1: **Stable Diffusion Architecture.** The encoder (E) compresses the input image (x) into a latent representation (z), which undergoes a diffusion process in the latent space to create (z_T). A U-Net is used to iteratively denoise and refine the latent representation, reconstructing intermediate states (z_{T-1}) until the final image ($\hat{x}$) is generated. The decoder (D) then converts the refined latent representation back into the image space, completing the generation process. Image adapted from [68]

3.3 Strategies to Control the Diffusion Generation Process

Controlling the diffusion generation process is a critical aspect of tailoring the outputs of generative models to specific requirements. In recent years, significant advancements have been made in developing methods to influence the characteristics of generated images, such as realism, semantic consistency, and adherence to desired styles or attributes [86]. These strategies can be broadly categorized into approaches such as fine-tuning existing models, introducing auxiliary networks, leveraging input conditioning, and modifying sampling directions.

3.3.1 Training Models from Scratch

In specific scenarios, training a model from scratch, such as in the **Composer** [57] framework, may be necessary to ensure compatibility with custom controllable elements [52]. Composer trains models conditioned on distinct elements such as captions, sketches, or color information, allowing for bespoke architectures tailored to specific controllable elements. However, the high computational cost and time required make this strategy less commonly employed compared to leveraging pre-trained models.

3.3.2 Fine-Tuning Pre-Trained Models

Fine-tuning involves modifying the weights of pre-trained diffusion models to adapt them to specific tasks or datasets. Techniques like **DreamBooth** [73], **Textual Inversion** [26] and **Custom**

Diffusion [46] exemplify this approach.

DreamBooth enables models to learn new subjects, such as generating personalized images based on a small set of photos. Textual Inversion, on the other hand, teaches models about new concepts, enabling the generation of images in specific styles or with unique features. Custom Diffusion focuses on modifying cross-attention layers to refine semantic or stylistic outputs.

The advantage of fine-tuning lies in its ability to refine models for high-fidelity outputs. However, it requires additional computational resources and training data, which can limit its accessibility.

3.3.3 Modifying the Sampling Process

Instead of altering model weights, sampling-based methods guide the generation process by influencing the direction of the denoising steps. These methods include:

- **Pix2Pix Zero** [64]: Performs zero-shot image editing by leveraging reference embeddings and attention maps to guide the generation process toward desired concepts.
- **SEGA (Semantic Guidance)** [9]: Controls conceptual elements, allowing for the addition or removal of features.
- **Attend and Excite** [11]: Ensures specific subjects in the prompt are prominently represented by enforcing attention thresholds during generation.

These techniques offer memory-efficient alternatives to fine-tuning, making them more accessible.

3.3.4 Input-Based Conditioning

Input conditioning utilizes prompts, images, or combinations thereof to steer the generation process. There are techniques that allow users to influence outputs without modifying the model itself. This include:

- **Image-to-Image Generation (img2img)** [56]: Starts with an existing image and modifies it to align with desired attributes or styles.
- **Prompt Weighting** [65]: Adjusts the influence of specific parts of a textual prompt to emphasize certain characteristics in the generated image.
- **Hybrid Conditioning** [69]: Combines text prompts, images, and auxiliary data for unparalleled flexibility in controlling outputs.

These methods are particularly effective for applications requiring nuanced adjustments, such as targeted image enhancements.

3.3.5 Auxiliary Networks for Conditional Control

Auxiliary networks such as **ControlNet** [95] and **T2I-Adapter** [58] extend the capabilities of pre-trained diffusion models by adding new trainable parameters. These networks introduce additional conditioning mechanisms, such as edge detection, semantic segmentation, or depth maps, which provide precise control over the generation process. This approach preserves the integrity of the original model while enhancing its adaptability to new tasks.

3.3.5.1 ControlNet for Conditional Diffusion

ControlNet [95] introduces a novel mechanism to enhance conditional control in diffusion models by adding specific constraints to the image generation process. Built as an extension to pre-trained stable diffusion models [45], ControlNet allows for precise control over the spatial, structural, and contextual aspects of generated images through the inclusion of auxiliary conditions such as edge maps, depth maps, segmentation masks, or encoded image representations [21]. This approach is particularly advantageous in medical imaging, where it is crucial to maintain anatomical accuracy in generated images.

The architecture of ControlNet is structured around two primary components: a fixed (locked) diffusion model that retains the original generation capabilities, and a trainable control network that adjusts the output based on the provided conditions [68]. By separating these components, ControlNet maintains the underlying generative power of the diffusion model while allowing the control network to learn specific conditional features. Figure 3.2 demonstrates the integration of ControlNet into the Stable Diffusion pipeline for condition-based image synthesis.

One of the main advantages of ControlNet is its flexibility in specifying a variety of spatial and semantic conditions. For instance, in applications requiring strict anatomical detail, segmentation masks of specific structures or organs can be provided as input conditions [30, 21, 45]. These masks serve to guide the model in generating images that preserve the shape, location, and boundaries of key anatomical features. This characteristic allows for the creation of images that not only display the desired pathology but also maintain realistic visual coherence.

3.4 Summary and Conclusions

In summary, synthetic medical imaging represents a transformative approach within the medical field, providing effective solutions to challenges such as data scarcity, privacy preservation, and cost reduction in image annotation. This chapter has reviewed the advantages of synthetic images for medical applications, emphasizing their potential to diversify datasets and simulate various pathological conditions. This is particularly relevant for retinal fundus images in the context of diabetic retinopathy, where different stages of the disease can be effectively represented.

A detailed exploration of generative models has highlighted the benefits of utilizing a Stable Diffusion Model for creating diverse and high-quality images, as well as the significance of controlled synthesis. Additionally, the chapter has introduced strategies to control the diffusion

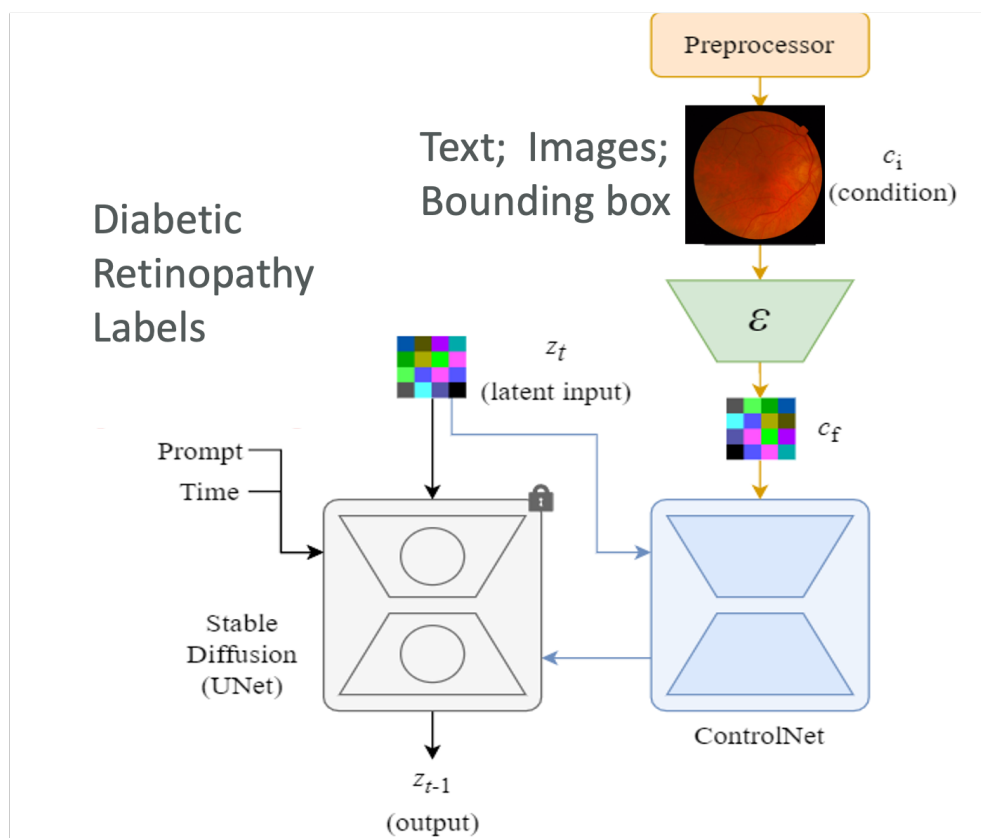


Figure 3.2: **Architecture of Stable Diffusion with ControlNet.** The process begins with a preprocessor that extracts conditioning information (c_i) from an input image, which is then transformed into feature representations (c_f) by the encoder ($\mathcal{E}$). A user-defined text prompt and time step are provided as inputs to the Stable Diffusion U-Net, along with the latent input (z_t). ControlNet refines the diffusion process by leveraging the conditioning features (c_f), enabling precise control over the generated image. The output latent representation (z_{t-1}) is progressively denoised to produce the final synthetic image. Image adapted from [80]

generation process providing insights into techniques that enable the synthesis of retinal fundus images with specific attributes. These strategies, such as the implementation of ControlNet, allow for precise control and accuracy over the generated images. This is especially important for retinal fundus images, where controlled synthesis is crucial for accurately depicting the characteristic lesions associated with diabetic retinopathy, depending on the stage of the disease.

Chapter 4

Evaluation of Synthetic Retinal Fundus Images

In medical imaging, the evaluation of synthetic datasets is particularly critical, as subtle inaccuracies in the data can lead to diagnostic reliability or misclassification of pathological conditions. This chapter, reviews key approaches for evaluating synthetic retinal fundus images across three complementary dimensions: fidelity, diversity, and utility. Fidelity and diversity are addressed through the implementation of metrics that capture structural, textural, and pathological characteristics, ensuring that synthetic data not only resemble real-world distributions but also exhibit sufficient variability to enhance representativeness. Utility is examined in terms of the capacity of synthetic images to support clinically relevant tasks, such as diabetic retinopathy classification, where the performance of models trained with augmented datasets is assessed. Importantly, this dimension also provides a means to validate newly generated synthetic classes, ensuring that data augmentation through synthetic images contributes meaningfully to model robustness and clinical applicability. Evaluation of utility includes the assessment of RETFound [99], using established classification metrics such as AUROC and AUPR to ensure comparability with prior studies. The overall objective is to provide a comprehensive framework for validating synthetic retinal datasets, ensuring their reliability for clinical and research applications.

4.1 Quality Evaluation of Synthetic Retinal Fundus Images

The evaluation of synthetic retinal fundus images relies on metrics that compare generated data with real clinical images. These metrics assess whether the synthetic data are realistic and representative. Two fundamental contexts underpinning this evaluation are fidelity and diversity, which together ensure that synthetic datasets are clinically meaningful.

Fidelity in the context of synthetic data generation refers to the degree to which synthetic images accurately replicate the visual and statistical properties of real-world data [89]. High fidelity implies that synthetic data maintains the essential structural, textural, and distributional characteristics present in real datasets.

Diversity, on the other hand, relates to the ability of synthetic datasets to capture a wide range of variations observed in real-world data, including different pathological features and demographic variations. A diverse dataset is essential to train robust machine learning models that generalize well across diverse clinical scenarios [53].

Ensuring fidelity and diversity of synthetic retinal fundus images is essential for generating datasets that are both realistic and representative. High-fidelity images accurately replicate the appearance of retinal structures and pathological lesions, while diversity ensures the inclusion of various disease stages and lesion patterns. To evaluate these aspects, a combination of advanced metrics can be applied, each focusing on different aspects of image quality and diversity.

In this study, the metrics considered include the Fréchet Inception Distance (FID), the Gaussian Mixture Model-based Quality Assessment (GMM-GIQA), and Improved Precision and Recall (IP&IR).

For the computation of these quality metrics, it was used the pyMDMA library [20].

4.1.1 Fréchet Inception Distance (FID)

FID [34] is one of the most widely used metrics for evaluating generative models. It measures the similarity between the feature distributions of real and synthetic images by modeling them as multivariate Gaussian distributions and computing the Fréchet distance between the two. A lower FID score indicates a closer alignment between real and synthetic data distributions, suggesting higher quality.

FID typically uses features extracted from an InceptionV3 network pretrained on ImageNet, which consists primarily of natural images. As a result, these features may not capture clinically relevant characteristics in medical images [61, 88], leading to questionable clinical utility. Moreover, FID provides a single aggregated score that reflects the overall similarity between real and generated image distributions, but it does not disentangle fidelity from diversity. To overcome these limitations, newer metrics such as (IP&IR) [47] have been proposed, as they separately assess fidelity and diversity, offering more clinically meaningful insights.

4.1.2 GMM-GIQA: Gaussian Mixture Model-based Quality Assessment

The GMM-GIQA metric is tailored specifically for image quality assessment (IQA) in generated image datasets [28]. By employing a Gaussian Mixture Model (GMM), this metric learns the distribution of real images, including characteristics of healthy and pathological patterns. It then calculates the probability of a synthetic image belonging to this distribution, yielding a quality score.

This metric utilizes meaningful features extracted from real images, such as color histograms, texture descriptors, or deep features obtained from a pretrained neural network [28], to encapsulate essential structural and textural information. The next step involves modeling with a GMM, which fits the probability distribution of the data as a weighted sum of multiple Gaussian components,

each with its own mean and covariance [54]. This allows the model to capture the multimodal nature of retinal image features.

For the quality assessment, from the synthetic image, it is extracted its feature vector and computed the likelihood of this feature under the GMM [29]:

$$Q_{GMM-GIQA} = \log p(X_{\text{synthetic}})$$

A higher likelihood indicates that the synthetic image's features are well-represented by the GMM, suggesting high quality and realism.

When applied to synthetic retinal images, this metric can help confirm that generated images retain the distinct characteristics associated with diabetic retinopathy, such as microaneurysms, hemorrhages, and exudates, contributing to their diagnostic relevance.

4.1.3 IP&IR: Improved Precision and Recall

The Improved Precision and Recall (IP&IR) metric [47] evaluates the fidelity and diversity of generative models by comparing the feature distributions of real and synthetic datasets. This metric provides a nuanced assessment by separately measuring:

- **Precision:** The proportion of synthetic images whose features lie within the distribution of real images. A higher precision indicates better fidelity, meaning synthetic images closely replicate real-world features.
- **Recall:** The proportion of real images whose features are covered by the synthetic image distribution. A higher recall reflects greater diversity, ensuring the synthetic dataset encompasses the variability observed in real data.

This metric utilizes a pretrained neural network to extract feature embeddings for both real and synthetic images [47]. These embeddings represent images in a high-dimensional space where semantic similarities correspond to proximity. Then methods like k-Nearest Neighbors (k-NN), estimate the boundaries or regions that encapsulate the characteristics of each dataset [82], helping to determine how closely the synthetic data aligns with the real data and whether it adequately covers the variability observed in real-world samples.

This metric helps evaluate both fidelity and diversity, ensuring that synthetic retinal images not only look realistic but also capture the diversity found in clinical settings. In synthetic datasets for diabetic retinopathy, high precision and recall indicate that the synthetic images represent a wide range of observable features.

4.2 Utility of Synthetic Retinal Fundus Images

Beyond visual resemblance, synthetic images must provide practical utility by supporting critical clinical applications. For instance, models trained on synthetic data should demonstrate consistent

performance when applied to real clinical tasks, such as grading the severity of diabetic retinopathy or identifying early disease lesions. This step is particularly important for validating newly generated synthetic data classes, as their inclusion aims to augment existing datasets and improve the robustness of downstream classification models. Utility evaluation, therefore, involves assessing whether models trained on synthetic images perform equivalently to those trained on real data.

4.2.1 Evaluating the Utility of Synthetic Retinal Fundus Images Using RETFound

Foundation models, such as RETFound [99], represent a powerful approach to this evaluation by providing a robust, generalizable framework for disease detection and classification. This particular model, for fundus photographs, has demonstrated superior performance compared with traditional deep learning models. [99, 79] RETFound leverages self-supervised learning to extract meaningful patterns from vast quantities of unlabelled retinal images, making it especially suited for assessing synthetic images generated to mimic real clinical data [79].

4.2.1.1 Foundation models

Foundation models have revolutionized AI by shifting from narrowly trained, task-specific models to versatile architectures that perform consistently across a wide range of applications [12]. Developed by learning general representations from extensive datasets, these models require significantly less labelled data to adapt to new tasks and are particularly effective in zero-shot and few-shot learning contexts. This adaptability and high label efficiency allow RETFound to serve as a benchmark for evaluating the quality and clinical relevance of synthetic retinal images.

4.2.1.2 RETFound

RETFound was trained on a combination of 1.3 million natural images and 1.6 million unlabelled retinal images through a masked autoencoder approach, where it learned to reconstruct missing parts of an image to derive clinically relevant features [99]. This pretraining enables RETFound to accurately classify and predict retinal diseases from retinal photography. In RETFound study, the authors conducted a evaluation of the model's performance in diabetic retinopathy classification using three publicly available datasets: Kaggle APTOS-2019 [42], IDRiD [70], and MESSIDOR-2 [15, 2]. For instance, when fine-tuned on Kaggle APTOS-2019, RETFound achieved an AU-ROC of 0.822 on IDRiD and 0.738 on MESSIDOR-2, outperforming comparison models such as SL-ImageNet and SSL-ImageNet. The remaining results, as reported by [99], of the RETFound model's performance across the different datasets are represented in Figure 4.1. These results highlight RETFound's superior performance and generalizability in DR classification tasks.

To assess the utility of synthetic images, the objective is to enhance RETFound's performance through the use of synthetic retinal fundus images as a data augmentation strategy. By augmenting the training datasets with synthetic images, we aim to improve RETFound's performance in diabetic retinopathy classification. This will be evaluated by comparing the model's performance on an augmented dataset containing synthetic images versus a dataset without them. The model's

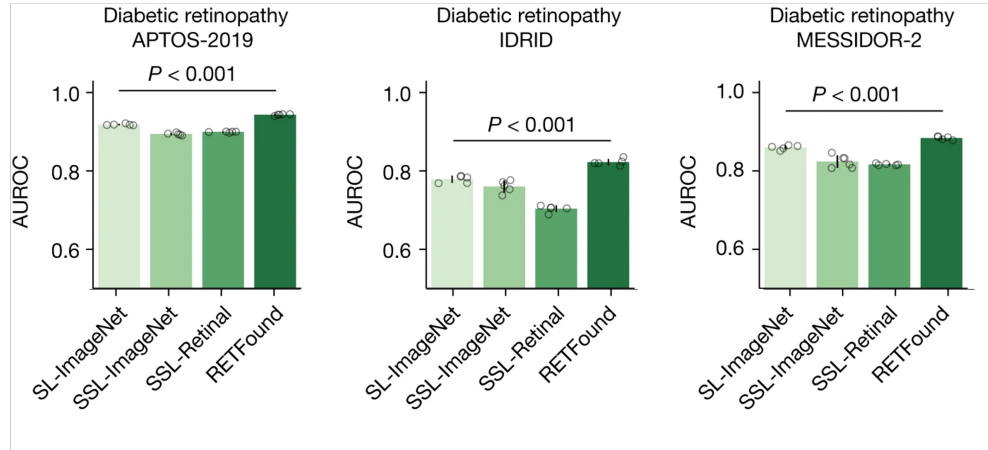


Figure 4.1: **Performance of the RETFound model in DR classification.** The model was fine-tuned on the Kaggle APTOS-2019, IDRID, and MESSIDOR-2 datasets, using AUROC as the evaluation metric. The results presented compare RETFound with other models, including SL-ImageNet, SSL-ImageNet, and SSL-Retinal. RETFound consistently outperforms the other models, demonstrating its superior performance and generalizability in DR classification from retinal images. Image from [99]

performance will be assessed using the same cross-dataset experimental setup employed in the original study, enabling a direct comparison to evaluate the impact of synthetic data augmentation and, consequently, its utility.

4.2.2 Classification Evaluation Metrics

To assess the classification performance of RETFound when trained with datasets augmented using synthetic retinal fundus images, we employed the same evaluation metrics reported in the RETFound paper: the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision–Recall Curve (AUPR). These metrics provide threshold-independent measures of classification performance and allow for a consistent comparison between results obtained in this study and those reported in the RETFound paper.

4.2.2.1 Area Under the Receiver Operating Characteristic Curve (AUROC)

AUROC is one of the most widely adopted metrics for evaluating binary classification tasks [51]. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1–specificity) across a range of classification thresholds [91]. AUROC is defined as the probability that a randomly chosen positive instance is assigned a higher score by the classifier than a randomly chosen negative instance. Values range from 0.5, representing random chance, to 1.0, indicating perfect discrimination.

However, AUROC has limitations, it gives equal weight to all regions of the score distribution and may be misleading when comparing models across datasets with differing levels of class imbalance [55]. For this reason, complementary metrics such as AUPR are often reported alongside AUROC to provide a more nuanced view of model performance.

4.2.2.2 Area Under the Precision–Recall Curve (AUPR)

AUPR complements AUROC by focusing on the relationship between precision and recall across varying thresholds [96]. The PR curve emphasizes performance on the positive class, making it especially informative in applications where early retrieval of true positives is critical, such as in disease detection. AUPR has gained significant traction in biomedical and clinical research, where class imbalance and the importance of correctly identifying positive cases are common.

Like AUROC, AUPR is threshold-independent and condenses classifier performance into a single value ranging from 0 to 1, with higher values reflecting stronger discrimination. By capturing the trade-off between avoiding false positives (precision) and recovering as many true positives as possible (recall), AUPR provides an interpretable and clinically relevant measure of classifier performance [31].

4.3 Summary and Conclusions

In the context of synthetic retinal fundus images, fidelity and diversity metrics are essential for validating both the resemblance to real images and the representativeness of the generated datasets. Fidelity metrics ensure that synthetic data preserve the structural, textural, and pathological properties of real retinal images, while diversity metrics guarantee sufficient variability to promote model robustness across different clinical scenarios. Together, these dimensions provide a rigorous foundation for assessing the overall quality of synthetic datasets.

Beyond quality, utility plays a central role in determining whether synthetic data can support meaningful clinical applications. Utility evaluation not only measures the effectiveness of synthetic images in tasks such as diabetic retinopathy classification but also provides a mechanism for validating newly generated synthetic classes, ensuring that data augmentation contributes to improved representativeness and model robustness. In this context, RETFound serves as a benchmark foundation model, with performance assessed using established classification metrics such as AUROC and AUPR, thereby ensuring comparability with prior studies.

In conclusion, combining fidelity, diversity, and utility establishes a comprehensive framework for the evaluation of synthetic retinal fundus images. This multidimensional validation ensures that synthetic data are not only visually and statistically aligned with real-world distributions but also clinically meaningful and effective for data augmentation strategies, ultimately reinforcing their role in advancing medical imaging research and practice.

Chapter 5

Generating Synthetic Retinal Fundus Images

The generation of synthetic images is a critical step towards developing balanced and representative datasets. In this chapter, it is presented the methodology employed to produce synthetic images that reflect specific disease severity classes. The primary objective is to augment existing real-world datasets, addressing the substantial class imbalance commonly observed in publicly available fundus image datasets.

It was adopted a controlled image synthesis approach, leveraging both pretrained generative models and additional conditioning strategies.

This chapter details the complete workflow, beginning with standardized preprocessing of the original datasets, followed by the use of Stable Diffusion as a pretrained generative backbone, and concluding with the fine-tuning and application of ControlNet for class-specific image synthesis.

5.1 Datasets Processing

To ensure standardized inputs and preserve the anatomical integrity of the retinal fundus images, a dedicated preprocessing pipeline was implemented. The procedure included cropping and resizing and anatomical mask generation.

Each image was automatically cropped around the retinal border to remove irrelevant background regions. The cropping coordinates were determined by analyzing the intensity distribution in the value channel of the HSV color space, which provides a more robust separation between retina and background compared to grayscale intensity. The largest detected contour corresponding to the retinal boundary was used to define the region of interest. When the resulting crop was not square, zero-padding was applied to balance the dimensions, and all images were subsequently resized to a fixed resolution (512x512) for consistency and model compatibility.

In parallel, anatomical masks were generated to serve as conditioning inputs. First, a retinal contour mask was created by delineating the detected boundary.

Additionally, optic disc segmentation was performed automatically using a pretrained SegFormer-based semantic segmentation model [90]. The segmentation network predicted a binary mask of the optic disc, obtained by upsampling the output logits and selecting the most probable class for each pixel.

These masks were subsequently integrated into the preprocessing pipeline as conditioning inputs for the generative experiments.

This pipeline ensured standardized and anatomically meaningful preprocessing, while also providing auxiliary conditioning, essential for the subsequent generative modeling tasks.

Figure 5.1 illustrates the preprocessing pipeline applied to the retinal fundus images,

5.1.1 Datasets Description

Three publicly available retinal fundus image datasets were used in this study, the APTOS 2019 Blindness Detection [42], IDRiD [70], and Messidor-2 [15, 2].

The APTOS 2019 dataset contains fundus images collected from subjects in rural regions of India, acquired under varying imaging conditions as part of the Aravind Eye Hospital initiative for the Asia Pacific Tele-Ophthalmology Society (APTOS). Each image is annotated with a DR severity grade ranging from 0 (No DR) to 4 (Proliferative DR), following standard clinical classifications.

The IDRiD dataset comprises high-resolution macula-centered fundus images captured at an eye clinic in Nanded, India. In addition to DR severity grading, the dataset includes detailed annotations of common DR lesions—including microaneurysms, hemorrhages, and hard and soft exudates—as well as normal retinal structures. It also provides grading for diabetic macular edema (DME).

The Messidor-2 dataset consists of macula-centered fundus images collected at Brest University Hospital, France. Each examination contains two images (one per eye), capturing real-world variability in fundus photography. Although Messidor-2 does not include native ground-truth DR grades, third-party labeling initiatives have made diagnostic annotations available for research purposes.

All three datasets exhibit a high class imbalance, as summarized in Table 5.1.

To address this, synthetic image generation will be employed to balance the classes. Specifically, for each original image of a given class, artificial images will be generated corresponding to all other DR severity classes, as illustrated in Figure 5.2.

The resulting class distributions after augmentation are depicted in Figure 5.3, illustrating how the inclusion of synthetic images balances all DR severity classes across the three datasets. This strategy ensures a more uniform distribution of samples across classes, supporting more robust model training and evaluation.

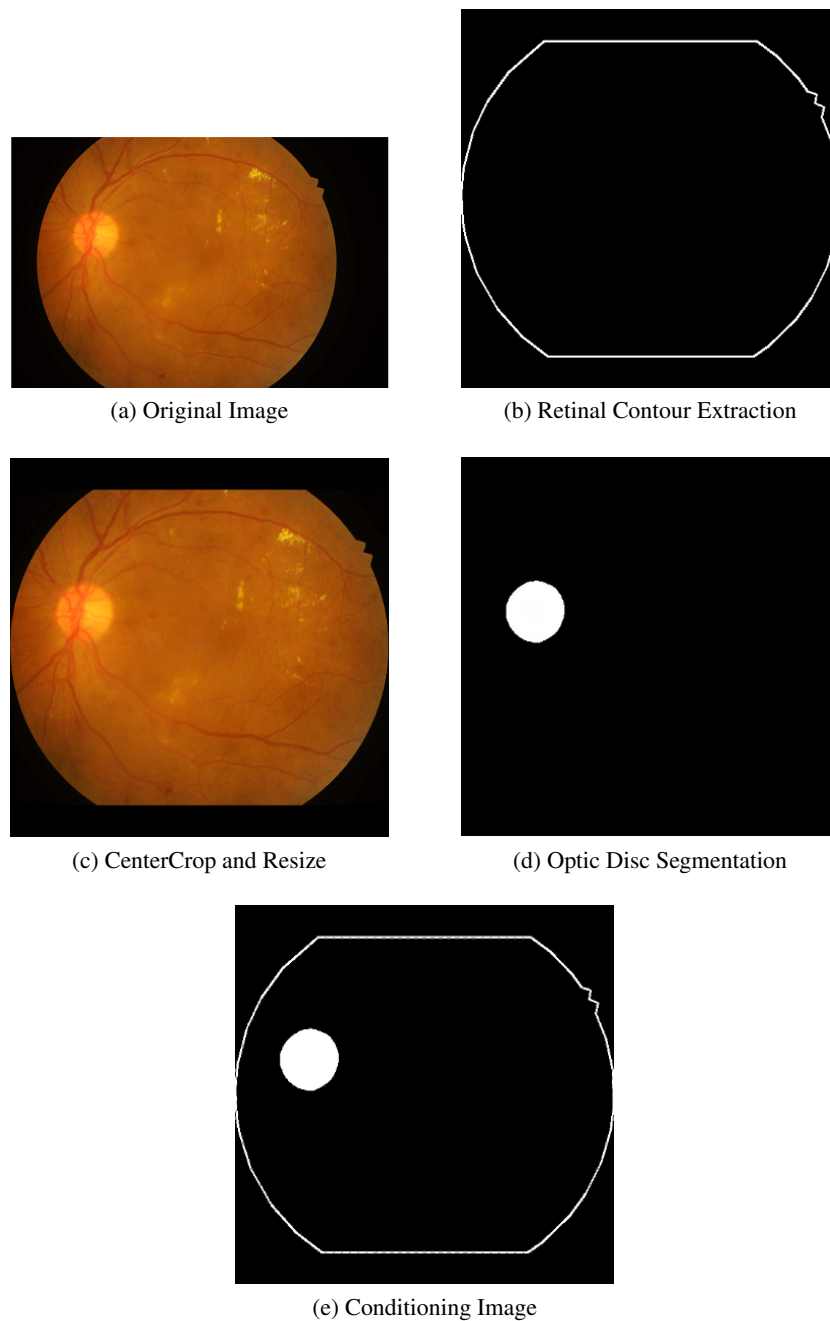


Figure 5.1: **Preprocessing pipeline of retinal fundus images.** This figure illustrates the sequence of operations applied to standardize the images, extract anatomical masks, and generate conditioning inputs for the generative model.

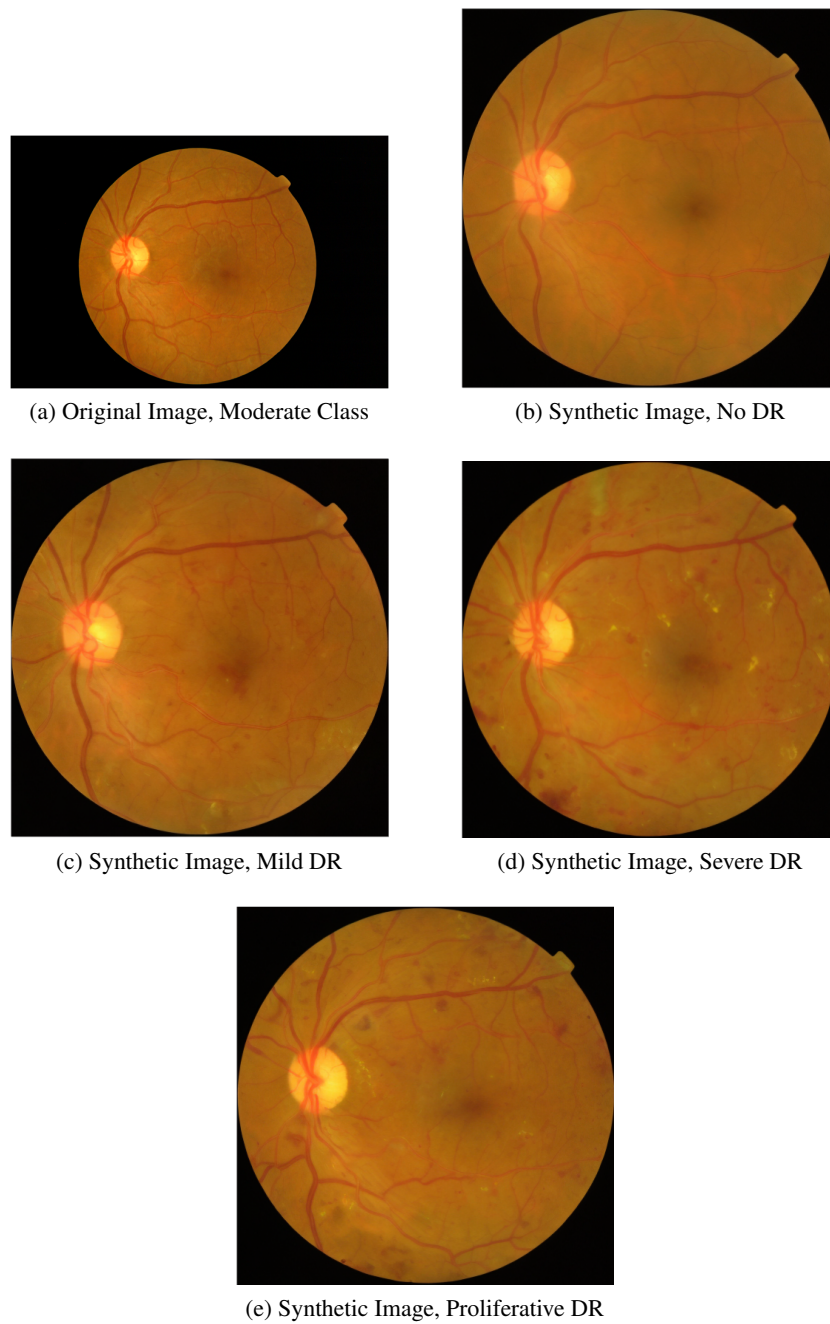
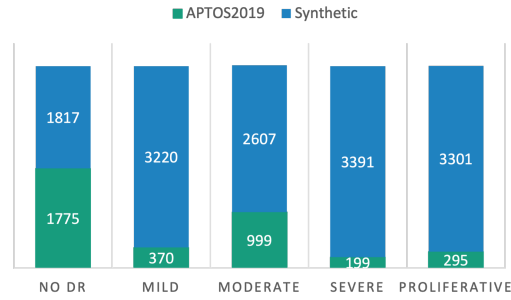
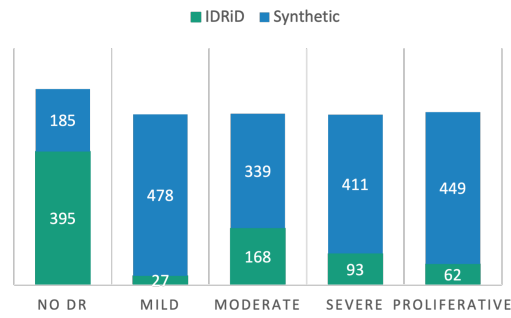


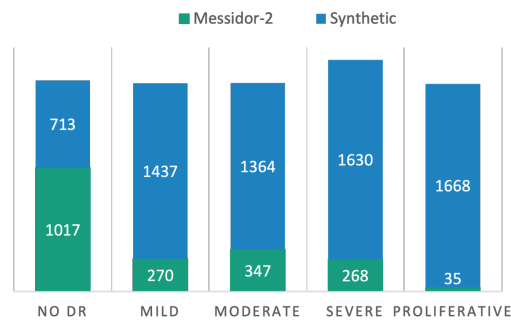
Figure 5.2: **Example of synthetic data generation for class balancing.** Additional images are generated from the original retinal fundus image to represent the other diabetic retinopathy severity classes, ensuring a more uniform class distribution across the dataset.



(a) **APTOS2019**: 3662 real images; 14336 synthetic images.



(b) **IDRiD**: 745 real images; 1862 synthetic images.



(c) **Messidor-2**: 1937 real images; 6812 synthetic images

Figure 5.3: Class distribution before and after synthetic augmentation for each dataset. These bar plots illustrate the effect of synthetic image generation in balancing diabetic retinopathy severity classes across the three datasets, mitigating the original class imbalance present in the real data.

Table 5.1: Distribution of DR Classes Across Datasets

Class	Datasets		
	APTOS 2019	IDRiD	Messidor-2
No DR	1775	395	1017
Mild DR	370	27	270
Moderate DR	999	168	347
Severe DR	199	93	268
Proliferative DR	295	62	35
Total	3632	745	1937

5.2 Stable Diffusion for Retinal Fundus Image Generation

The generative backbone used in this study is based on Stable Diffusion, a latent diffusion model that operates by progressively denoising random latent variables under the guidance of conditioning inputs, as illustrated in Figure 5.4.

The pretrained model adopted in this work was developed using large-scale retinal fundus datasets, specifically EyePACS [17] and AIROGS [14], which provide extensive coverage of diabetic retinopathy and glaucoma cases [19]. During its training, text prompts were employed not only to specify disease-related class labels, for Diabetic Retinopathy and Glaucoma, but also to encode image quality attributes. These attributes included categorical descriptors of brightness (low/medium/high), contrast, and other relevant quality factors, thereby enabling the generation of synthetic images with controlled fidelity and variability.

In this study, the Stable Diffusion model wasn't retrained from scratch. Instead, this pretrained backbone was employed as the basis for fine-tuning ControlNet, which was subsequently adapted to our datasets and experimental setup for the synthesis of diabetic retinopathy images.

5.3 ControlNet for DR Retinal Fundus Image Generation

To generate synthetic retinal fundus images, it was adopted a ControlNet-based approach, leveraging a pretrained model originally trained on the FGADR [97] and DDR [63] datasets. Instead of training the network from scratch, it was performed fine-tuning to adapt it to the characteristics of the selected datasets individually (Aptos 2019, IDRiD and Messidor-2) and the specific requirements of diabetic retinopathy synthesis. ControlNet extends diffusion models by introducing additional conditioning inputs, allowing for more controllable image generation.

In this study, generation was exclusively performed in img2img mode, which adapts an existing image into a new version guided by conditioning prompts. Unlike txt2img, which synthesizes images solely from text prompts, img2img starts from a real preprocessed fundus image and progressively transforms it during the diffusion process. This approach ensures that essential retinal

structures (e.g., the circular shape, and overall spatial arrangement) are preserved, while new features consistent with the target DR severity are incorporated. Within ControlNet, img2img leverages both the anatomical masks and the textual labels to constrain the generative process, ensuring anatomical plausibility and disease-specific characteristics in the synthetic outputs.

As input images, were employed the preprocessed fundus photographs described above (Figure 5.1c). The conditioning signals consisted of two anatomical masks: the retinal contour mask and the optic disc segmentation mask (Figure 5.1e), both integrated into the pipeline to preserve anatomical plausibility and provide structural guidance during generation. Additionally, text prompts were used to specify the target DR severity class. The following prompts were defined: “retinal fundus image”, “mild diabetic retinopathy retinal fundus image”, “moderate diabetic retinopathy retinal fundus image”, “severe diabetic retinopathy retinal fundus image”, and “proliferative diabetic retinopathy retinal fundus image”. When necessary, negative prompts were introduced to explicitly suppress undesired features or attributes, enhancing the specificity of the generated outputs.

The output of this process consisted of synthetic retinal fundus images that maintained the anatomical structure of the original inputs, while reflecting the newly assigned DR severity labels. This strategy enabled the generation of balanced datasets across all severity grades, thereby mitigating class imbalance present in the original data.

5.3.1 ControlNet Inference Parameters

The ControlNet inference process depends on several hyperparameters:

- **Guidance Scale:** controls the strength of alignment between the generated image and the conditioning inputs. Higher values enforce stricter adherence to the prompts.
- **Inference Steps:** defines the number of denoising iterations during the diffusion process. More steps generally improve image quality and detail but increase computational cost.
- **Noise Scheduler:** determines how noise is added and removed during the diffusion process. The choice of scheduler affects both the sharpness and fidelity of the generated images.
- **Strength:** specifies the degree of transformation applied to the input image in img2img mode. Lower values retain more features of the original image, while higher values allow for greater modifications toward the target prompt.

These parameters were varied systematically across experiments, with the objective of identifying optimal configurations for retinal fundus image synthesis.

5.4 Summary and Conclusions

In this chapter, was presented a comprehensive methodology for generating synthetic retinal fundus images, aimed at augmenting real datasets and addressing class imbalance across diabetic retinopathy (DR) severity levels.

Analysis of the datasets revealed substantial discrepancies in the number of images per class, with the mild and proliferative DR classes consistently underrepresented compared to other severity levels across all three datasets. To mitigate this imbalance, synthetic image generation was employed to augment these underrepresented classes, supporting more uniform class distributions.

The use of a pretrained Stable Diffusion model as a generative backbone, combined with ControlNet fine-tuning, enabled the production of synthetic images. Conditioning through anatomical masks and disease-specific text prompts ensured that the generated images maintained anatomical plausibility while reflecting the desired DR severity classes. Negative prompts were optionally employed to suppress undesired features, further improving the specificity of the outputs.

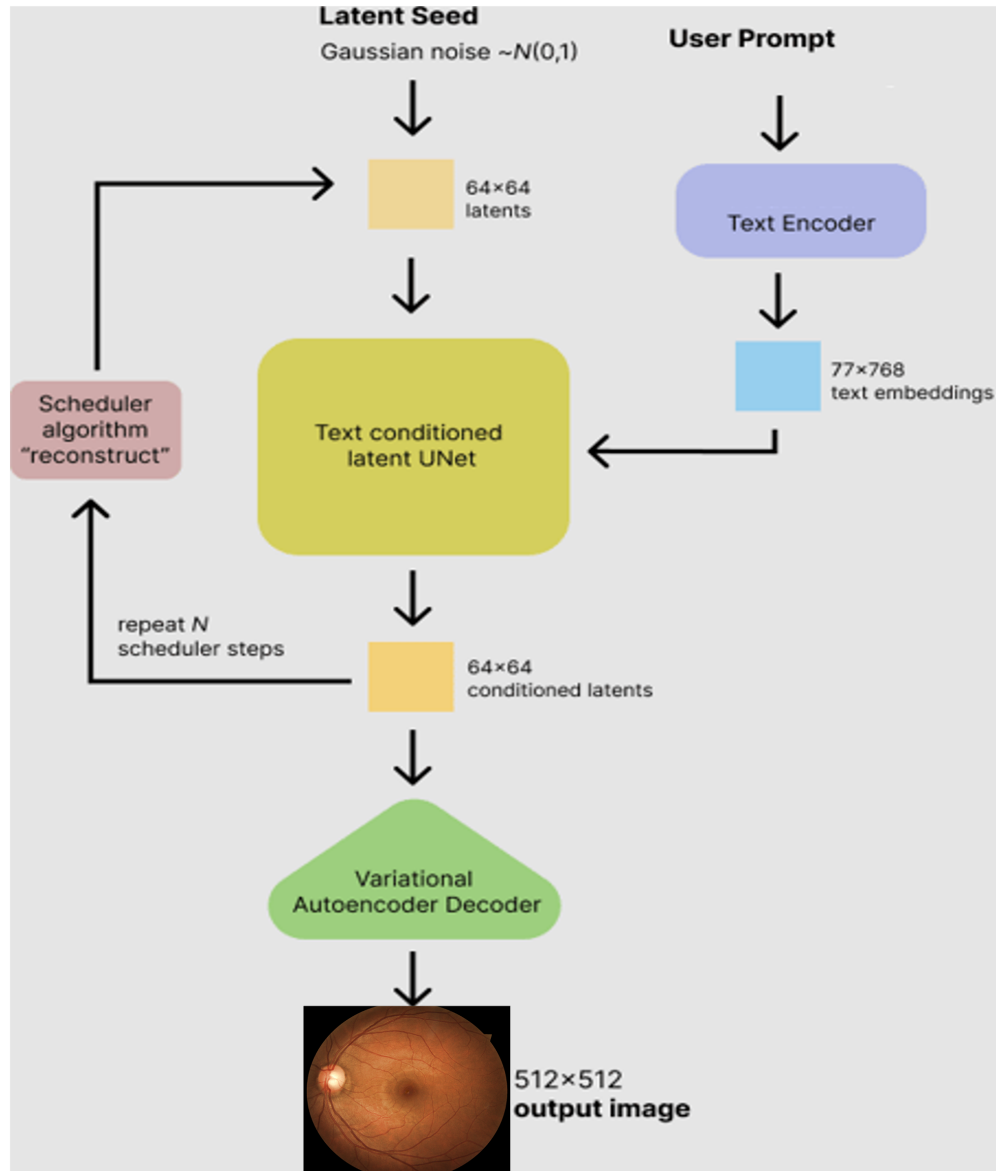


Figure 5.4: **Pipeline of the Stable Diffusion model for retinal fundus image generation.** Pre-trained on the EyePACS and AIROGS datasets, the model receives as inputs representations of retinal images and is guided by textual prompts specifying disease labels (DR or glaucoma) and image quality attributes (low/medium/high brightness and contrast). These conditioning signals direct the generation process, producing synthetic images with controlled fidelity and diversity that reflect both the pathological characteristics and the desired quality parameters. Image adapted from [81]

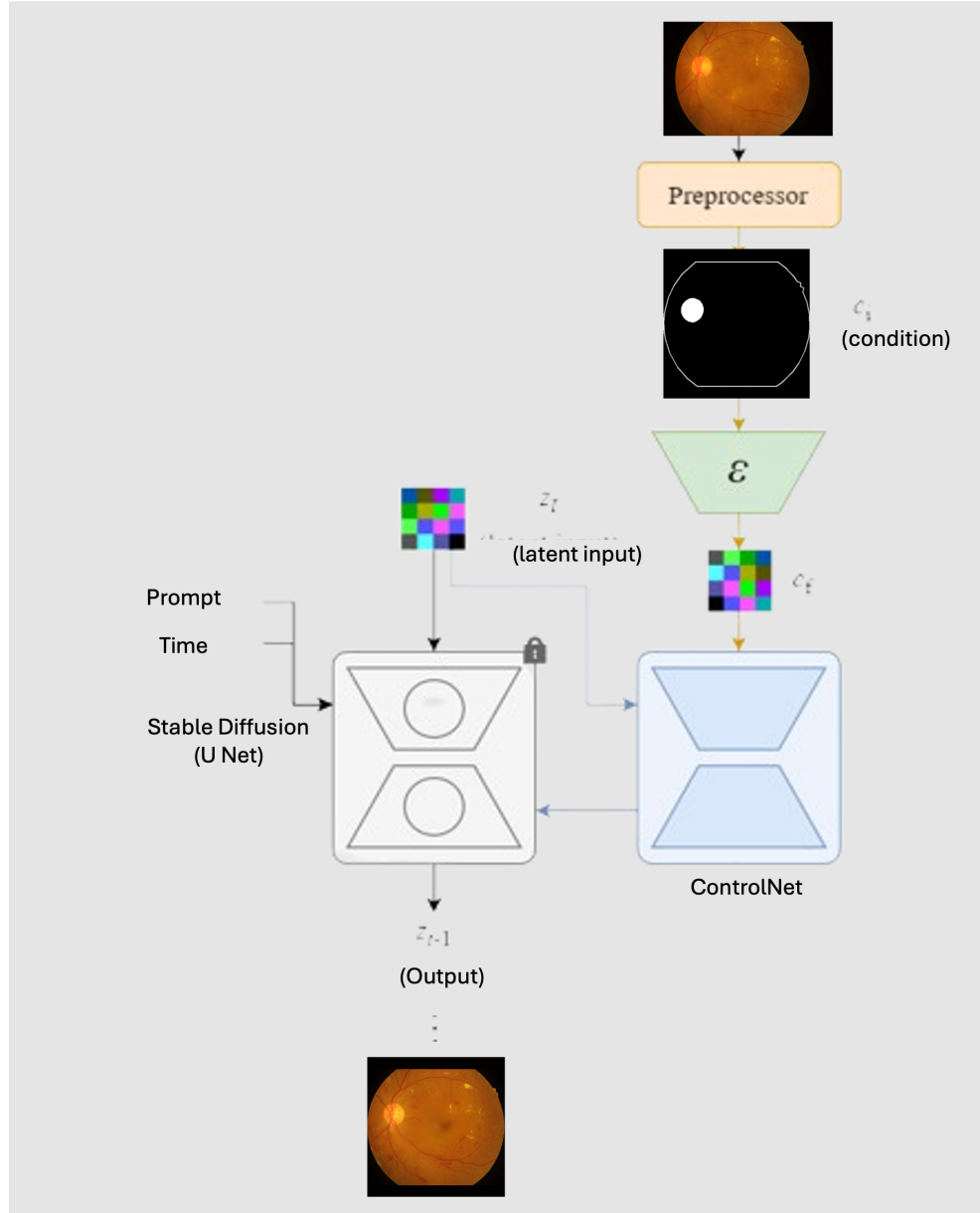


Figure 5.5: **Pipeline of the ControlNet-based synthetic image generation process.** Preprocessed retinal fundus images serve as the inputs, while masks (retinal contour and optic disc segmentation) are incorporated as structural conditioning. Text prompts specifying the target DR severity class guide. In the img2img framework, ControlNet integrates these conditioning signals with the diffusion model to generate anatomically consistent synthetic images that reflect the assigned DR severity grade. Image adapted from [80]

Chapter 6

RETFound for Utility Assessment of Synthetic Data

The validation of synthetic retinal fundus images is a critical step to ensure their practical utility in clinical and research applications. In this chapter, the evaluation is performed through fine-tuning the RETFound foundation model for diabetic retinopathy (DR) classification. The RETFound-based evaluation provides a direct measure of the usefulness of synthetic images for downstream diagnostic tasks, allowing the impact of synthetic data on model generalization and classification performance to be quantified.

To explore clinical utility, synthetic images were employed to augment real DR datasets, addressing class imbalance across severity levels.

6.1 Validation via RETFound Fine-Tuning

To assess the utility of synthetic retinal images, the RETFound foundation model [99] was employed. RETFound is a large-scale, self-supervised model trained on millions of unlabelled retinal fundus images and natural images, capable of extracting clinically relevant features for various downstream tasks, including disease classification. Its architecture incorporates a masked autoencoder with a multilayer perceptron (MLP) head for classification tasks, enabling it to adapt efficiently to different retinal image analysis challenges. The overall RETFound fine-tuning pipeline, including model architecture, is illustrated in Figure 6.1.

In the original study [99], RETFound was fine-tuned on publicly available datasets, namely APTOS 2019, IDRiD, and Messidor-2, for diabetic retinopathy (DR) classification. The authors applied standardized data splits for training, validation, and testing to ensure comparability across experiments.

In this work, the same datasets and data splits were employed to replicate the baseline fine-tuning process, allowing direct comparison with the reported performance of RETFound. Subsequently, the fine-tuning was repeated using an augmented dataset composed of both real and synthetically generated images, under identical training conditions. This procedure was applied

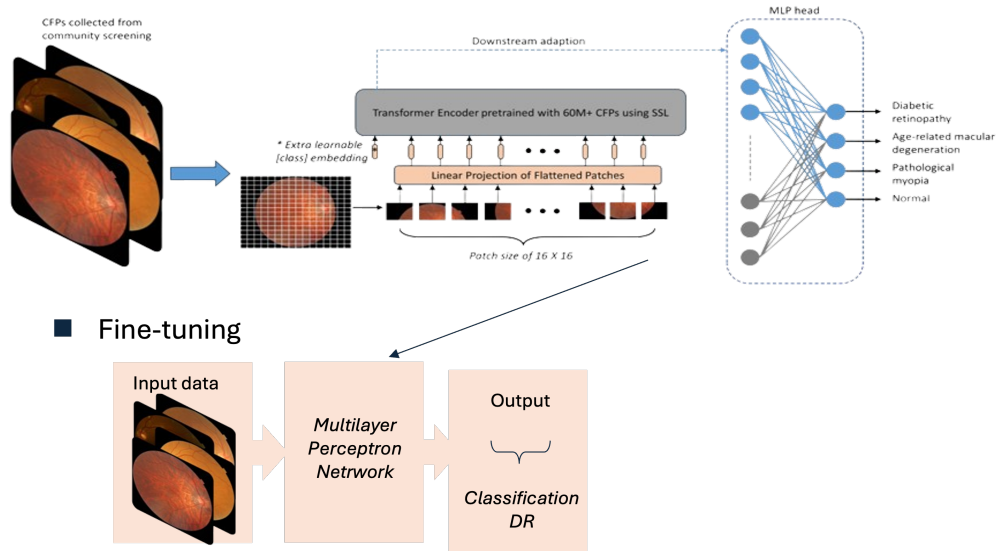


Figure 6.1: **Pipeline overview of RETFound fine-tuning for diabetic retinopathy classification.** The figure illustrates the input steps, into the RETFound masked autoencoder with an MLP classification head." Image adapted from [94]

independently for each dataset, enabling evaluation of the impact of synthetic data on model performance relative to the baseline trained solely on real images.

By comparing the classification performance of RETFound trained with real-only data versus real plus synthetic data, this methodology provides an empirical measure of the utility of the synthetic images in enhancing downstream predictive tasks.

Table 6.1 shows the fine-tuning parameters used in this study.

Table 6.1: Fine-tuning parameters used for RETFound in this study.

Parameter	Value
Model	RETFound_MAE
Batch size	16
Epochs	100
Learning rate	5e-3
Layer decay	0.65
Input size	224 × 224

6.1.1 Image Preprocessing for RETFound Fine-Tuning

For the fine-tuning of RETFound, the original preprocessing pipeline described in the RETFound paper [99] was initially employed. This pipeline consists of resizing the input images followed by a center crop to the network's expected input size(224x224). While effective for standardizing inputs, the center crop can result in the exclusion of peripheral regions of the retina, potentially discarding anatomically relevant structures.

To preserve the anatomical integrity of the retinal images, an alternative preprocessing approach was also applied. This method involved cropping each image based strictly on the detected retinal contour, ensuring that the entire retinal area was included, followed by zero-padding when necessary to produce square images suitable for model input. This approach maintains the full anatomical structure of the retina, which may be beneficial when evaluating models that rely on fine-grained spatial information.

Experiments were conducted using both preprocessing strategies standard RETFound resizing and center cropping, as well as the anatomy preserving preprocessing to assess the impact of input processing on the performance of RETFound fine-tuning for diabetic retinopathy classification.

6.2 Summary and Conclusions

In this chapter, the validation of synthetic retinal fundus images was conducted through downstream task performance. Fine-tuning the RETFound foundation model for diabetic retinopathy classification allowed for a practical evaluation of synthetic image utility. By comparing models trained on real datasets with models trained on augmented datasets including both real and synthetic images, the impact of synthetic data on classification robustness and generalization was empirically assessed. Additionally, RETFound facilitates label verification, ensuring that synthetic images generated for specific DR stages accurately correspond to their intended labels.

Furthermore, experiments explored the effect of different image preprocessing strategies on RETFound fine-tuning performance. While the original RETFound resizing and center cropping ensured standardized inputs, it occasionally excluded peripheral retinal regions. The anatomy-preserving preprocessing maintained full retinal structure, potentially enhancing model performance on spatially sensitive features.

Chapter 7

Experiments

This chapter describes the experimental setup employed to evaluate both the synthetic image generation process and its downstream utility for diabetic retinopathy (DR) classification. The experiments were divided into two main parts: fine-tuning of the ControlNet img2img pipeline to generate synthetic retinal fundus images; and fine-tuning of the RETFound model for DR classification using both real and augmented datasets.

7.1 ControlNet Fine-Tuning Experiments

Synthetic retinal fundus images were generated using the ControlNet img2img pipeline. Fine-tuning was carried out separately for each dataset (APTOS 2019, IDRiD, and Messidor-2) under comparable conditions. The inference parameters were initially fixed across all three datasets, as summarized in Table 7.1.

Table 7.1: Initial inference parameters for ControlNet fine-tuning.

Parameter	Value
Guide scale	7
Inference steps	50
Noise scheduler	ddim
Strength	0.30

7.1.1 Negative Prompts

Negative prompts provide an additional mechanism of control in diffusion-based generation by explicitly discouraging undesired features during image synthesis. When applied to retinal fundus image generation, negative prompts allow the suppression of class-specific features not intended for the target prompt.

In this work, three experimental configurations were explored for each dataset and parameter setting:

1. **No negative prompts:** images were generated without applying any negative control.
2. **Full negative prompts:** all other DR stages were provided as negative prompts. For instance, if the target prompt was “moderate diabetic retinopathy fundus image”, then “mild, severe, proliferative” were set as negatives.
3. **Targeted negative prompts:** negative prompts were applied only in cases of class regression. When generating images of a lower DR grade from a more severe real image. For example, when conditioning on a severe case while targeting mild DR, the negatives included “moderate, severe”.

This strategy enabled systematic evaluation of the effect of negative prompts.

7.1.2 Parameter Variations

During the experiments, certain structures inconsistent with retinal anatomy were observed in the generated images. Representative cases of these inconsistencies can be seen in Figures 7.1a, 7.1c. To mitigate these issues, additional fine-tuning was performed with adjusted parameter settings 7.2. Lowering the strength parameter reduced artefacts, producing more anatomically plausible images (7.1b, 7.1d).

Table 7.2: Modified inference parameters.(Lowering the strength parameter)

Parameter	Value
Guide scale	7
Inference steps	50
Noise scheduler	ddim
Strength	0.20

Further experiments were performed by increasing the number of inference steps. This adjustment, detailed in Table 7.3, was designed to refine the synthesis process, with the goal of enhancing the perceptual quality and fidelity of the generated retinal images.

Representative examples of these results are illustrated in Figure 7.2, where pairs of synthetic images generated with lower and higher inference steps are compared.

Table 7.3: Modified inference parameters. (Increasing the number of inference steps)

Parameter	Value
Guide scale	7
Inference steps	70
Noise scheduler	ddim
Strength	0.20

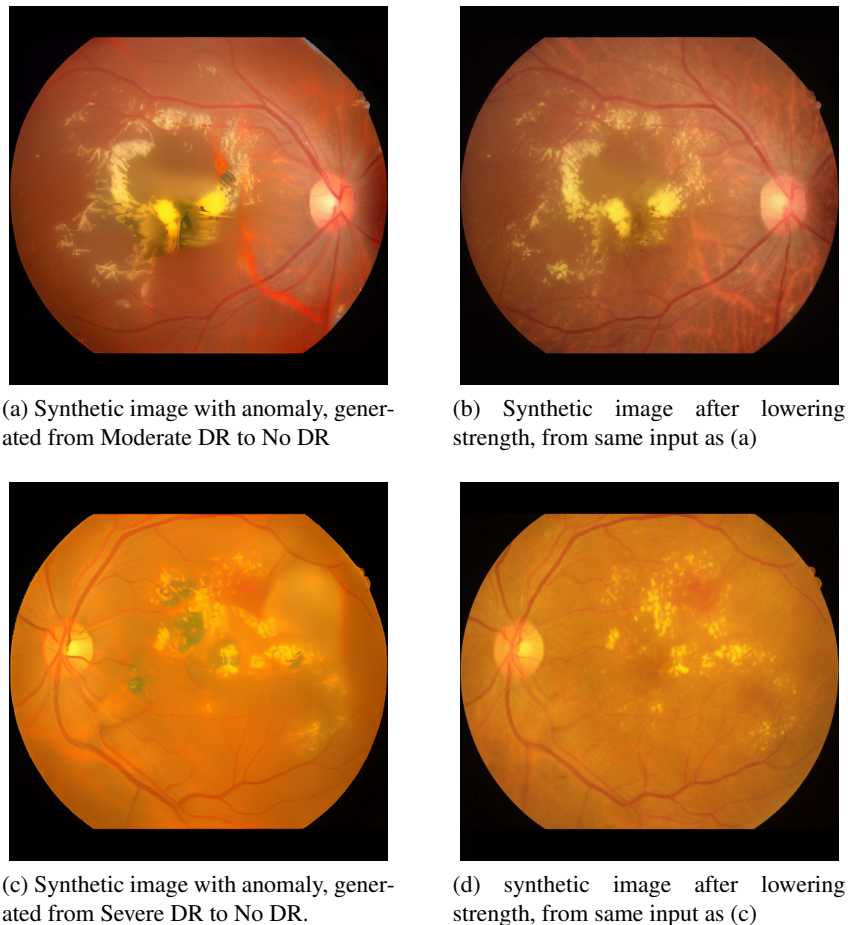
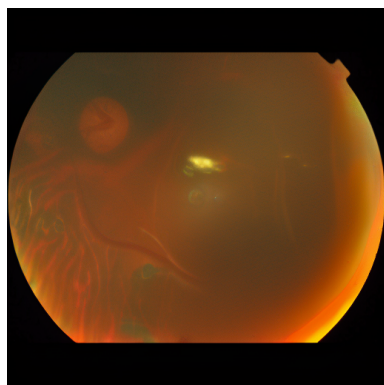
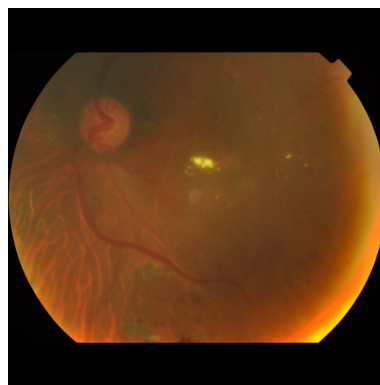


Figure 7.1: Examples of visual anomalies observed in synthetic images, compared with their counterparts obtained after reducing strength. Subfigures (a) and (b) show samples generated from the IDRiD dataset when synthesizing the No DR class from the same real Moderate DR image, while subfigures (c) and (d) illustrate the synthesis of No DR images from the same real Severe DR image. The reduction of the strength parameter mitigates the presence of artifacts, resulting in more realistic retinal fundus images.



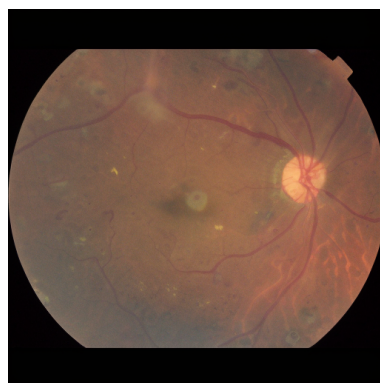
(a) Synthetic image generated with lower inference steps. Precision = 1, Recall = 0, GIQA = 0,7951.



(b) Synthetic image obtained from the same original input as (a) , using higher inference steps. Precision = 1, Recall = 1, GIQA = 0.8154



(c) Synthetic image generated with lower inference steps. Precision = 1, Recall = 0, GIQA = 0.9306.



(d) Synthetic image obtained from the same original input as (c) , using higher inference steps. Precision = 1, Recall = 1, GIQA = 0.9468

Figure 7.2: Examples from the APTOS dataset illustrating the effect of increasing the number of inference steps with the goal of enhancing perceptual quality and fidelity of synthetic retinal fundus images. All images were generated by conditioning from Moderate DR to Proliferative DR. Images (a) and (b) derived from the same real moderate DR image (a = lower steps, b = higher steps); Images (c) and (d) derived from the same real moderate DR image (c = lower steps, d = higher steps). For each subfigure the reported values (Precision, Recall, GIQA) allow quantitative comparison of image quality across configurations.

Finally, an additional set of experiments explored the combined effect of a higher guidance scale together with an increased number of inference steps. The parameter configuration is summarized in Table 7.4. This setting was tested to assess whether stronger conditioning during generation could produce images according to the desired visual characteristics

Table 7.4: Modified inference parameters. Increasing the guidance scale.

Parameter	Value
Guide scale	8.5
Inference steps	70
Noise scheduler	ddim
Strength	0.20

While consistent parameters were initially applied across datasets, dataset-specific behaviours were observed, making further adjustments almost unavoidable. To address these differences and improve the quality of the generated images, additional experiments for specific datasets were performed as follows:

7.1.2.1 IDRiD

Fine-tuning was also tested with parameters in Table 7.5.

Table 7.5: Modified inference parameters for IDRiD

Parameter	Value
Guide scale	8
Inference steps	60
Noise scheduler	ddim
Strength	0.30

7.1.2.2 Messidor-2

Fine-tuning was also tested with parameters in Table 7.6.

Table 7.6: Modified inference parameters for Messidor-2

Parameter	Value
Guide scale	8.5
Inference steps	60
Noise scheduler	ddim
Strength	0.15

These experiments aimed to identify parameter settings that maximized the anatomical plausibility, diversity, and perceptual quality of synthetic retinal fundus images for each dataset

7.2 RETFound Fine-Tuning Experiments

In the second part of the study, synthetic data were evaluated in terms of their contribution to downstream classification of DR. RETFound, was fine-tuned independently on each dataset (APTOS 2019, IDRiD, and Messidor-2). The data splits followed the protocol reported in the RETFound paper, ensuring comparability with prior work. For APTOS 2019, the split corresponds approximately to 60% for training, 10% for validation, and 30% for testing. IDRiD follows a distribution of about 60% training, 20% validation, and 20% testing. Finally, Messidor-2 adopts a split of roughly 60% training, 10% validation, and 30% testing. The Table 7.7 summarizes the distribution of samples across training, validation, and testing sets for all three datasets. The table also reports the class composition of each split according to the DR severity levels (no DR, mild, moderate, severe, proliferative), which highlights the inherent imbalance in the datasets and motivates the use of synthetic augmentation.

Table 7.7: Distribution of samples across training, validation and test splits for APTOS 2019, IDRiD and Messidor-2.

Dataset	Split	Total	No DR	Mild	Moderate	Severe	Proliferative
APTOS 2019	Train	2048	1010	207	559	108	164
	Val	514	253	52	140	27	42
	Test	1100	542	111	300	58	89
IDRiD	Train	328	107	16	108	59	39
	Val	84	27	4	28	15	10
	Test	105	34	7	32	19	13
Messidor-2	Train	1165	568	151	193	234	19
	Val	246	143	38	49	11	5
	Test	526	306	81	105	23	11

Two main experimental settings were established:

- **Baseline:** fine-tuning using only real dataset images.
- **Augmented:** fine-tuning using real plus synthetic images.

7.2.1 Image Preprocessing

RETFound requires inputs of size 224×224. Two preprocessing approaches were tested:

1. **Original RETFound preprocessing:** resizing and center-cropping.
2. **Anatomy-preserving preprocessing (described in Chapter 6):** cropping based on the retinal contour.

Initial experiments compared both strategies. As the anatomy-preserving preprocessing yielded superior results, all subsequent experiments used this approach.

7.2.2 Fine-Tuning and Evaluation

All fine-tuned models (both baseline and augmented) were trained under identical conditions. After training, performance was evaluated using standard classification metrics, enabling direct comparison between real-only and augmented datasets.

7.2.3 Complementary Experiments

Two complementary sets of experiments were conducted:

1. **Cross-dataset evaluation:** In addition to the standard train/test splits within the same dataset, as described in the RETFound paper, the original study also explored a cross-dataset approach, where models fine-tuned on one dataset were evaluated on a different dataset (for instance, trained on APTOS and tested on Messidor). In this work, both real-only and augmented (real + synthetic) versions of these cross-dataset experiments were performed. Although priority was given to the intra-dataset evaluation, as it is more directly comparable and the cross-dataset evaluation is more challenging due to differences in image characteristics, class distributions, and dataset-specific features.
2. **Quality-filtered augmentation:** instead of including all synthetic images, only those achieving maximum precision ($IP = 1$) according to pyMDMA quality metrics were used to augment the real datasets. This allowed assessment of whether filtering by quality improved downstream performance compared to unfiltered augmentation. To better illustrate the quality-filtered augmentation approach, Figure 7.3 presents examples of synthetic images with maximum precision ($IP = 1$) and low precision ($IP = 0$). These examples demonstrate the visual differences corresponding to the quality scores used for dataset augmentation.

7.3 Summary and Conclusions

This chapter presented a systematic set of experiments aimed at exploring the controllability and effectiveness of diffusion-based synthetic image generation with ControlNet. The first group of experiments focused on assessing how different parameter settings, such as strength, guidance scale, and inference steps, affect the fidelity, diversity, and anatomical plausibility of the generated retinal fundus images. These experiments demonstrated that parameter tuning was not merely optional but rather necessary, as distinct behaviours were observed across datasets. Adjustments in configuration proved essential for mitigating artifacts and enhancing structural consistency.

The second line of investigation concentrated on the utility of synthetic data when integrated into RETFound fine-tuning. By augmenting imbalanced datasets with carefully generated images, it was possible to assess the contribution of synthetic data to model training, particularly in challenging scenarios with limited representation of specific diabetic retinopathy grades.

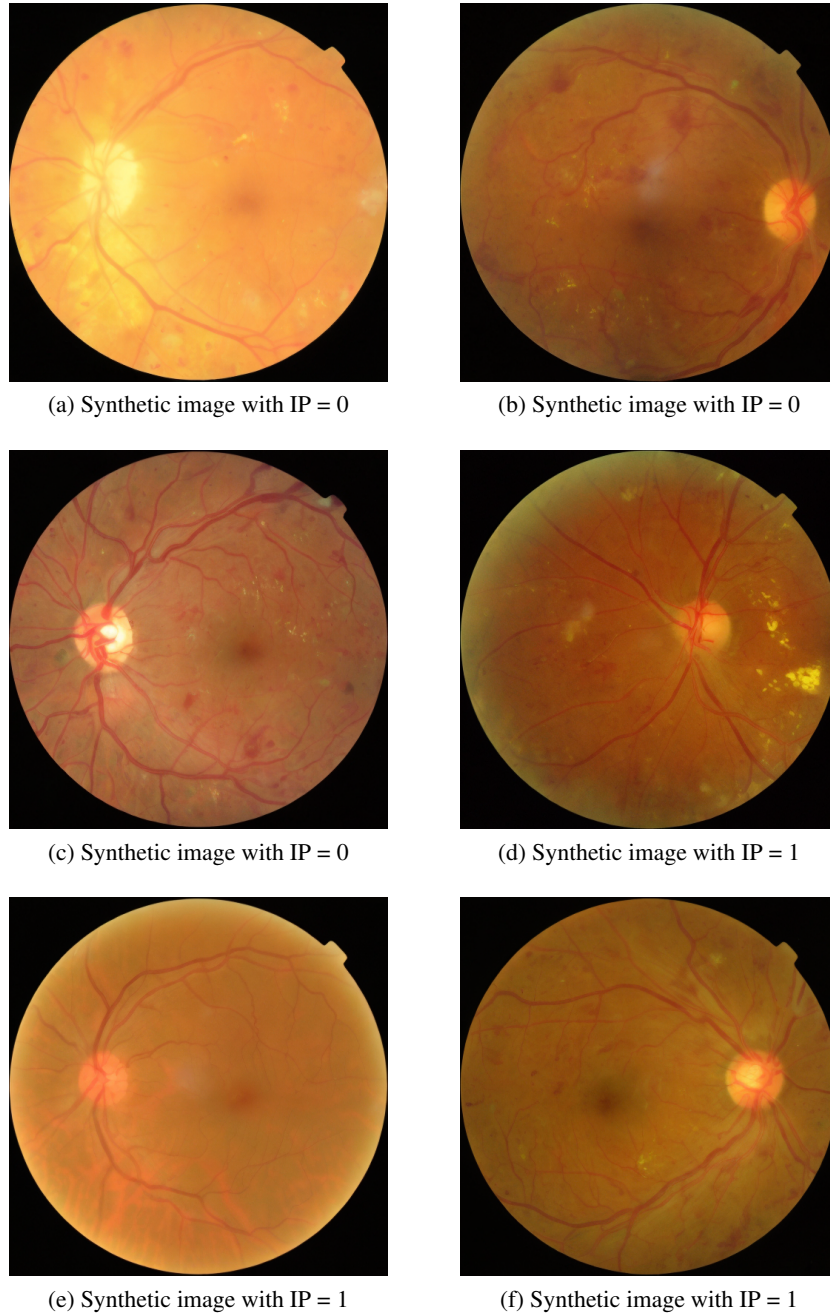


Figure 7.3: **Examples of synthetic retinal fundus images for quality-filtered augmentation.** Images generated from the Messidor dataset. These examples illustrate the differences in image quality considered during the augmentation process, highlighting the visual criteria behind the filtering strategy.

The following chapter presents the results and discusses the impact of synthetic augmentation across the three datasets. The subsequent section focus exclusively on the best results obtained for each dataset, omitting a analysis of the complementary experiments.

Chapter 8

Results and Discussion

This chapter presents the results and discussion of the validation experiments performed with synthetic retinal fundus images.

Given the extensive number of experiments performed, only the most representative outcomes are reported. For each dataset (APTOS 2019, IDRiD, and Messidor-2), the baseline performance obtained with real data is compared against the best-performing configuration achieved when augmenting the dataset with synthetic images.

In addition to quantitative results, selected synthetic examples are included to visually demonstrate the quality and variability of the generated images under different configurations.

The chapter aims to provide a comprehensive evaluation of the utility of synthetic data in improving diabetic retinopathy classification.

8.1 APTOS 2019 Dataset

For the APTOS 2019 dataset, the reported results correspond to the experimental configuration with inference parameters seen in Table 7.2, applying full negative prompts (Chapter 7). This setting was identified as the most effective for this dataset, from a total of 12 experiments, including four fine-tuning configurations, with one of these further evaluated under three variations of negative prompts. The configuration reported here was selected as the most effective among these experiments, leading to the representative distribution observed in the UMAP feature projection shown in Figure 8.4a.

This configuration produced the synthetic samples illustrated in Figure 8.5, and served as the basis for the augmented dataset used in the classification experiments.

The corresponding confusion matrices and evaluation metrics are presented below.

8.1.1 Confusion Matrices Analysis

Figure 8.1 presents the confusion matrices for both experimental settings. The augmented model (Figure 8.1b) exhibits improved recognition of minority classes, such as mild non-proliferative

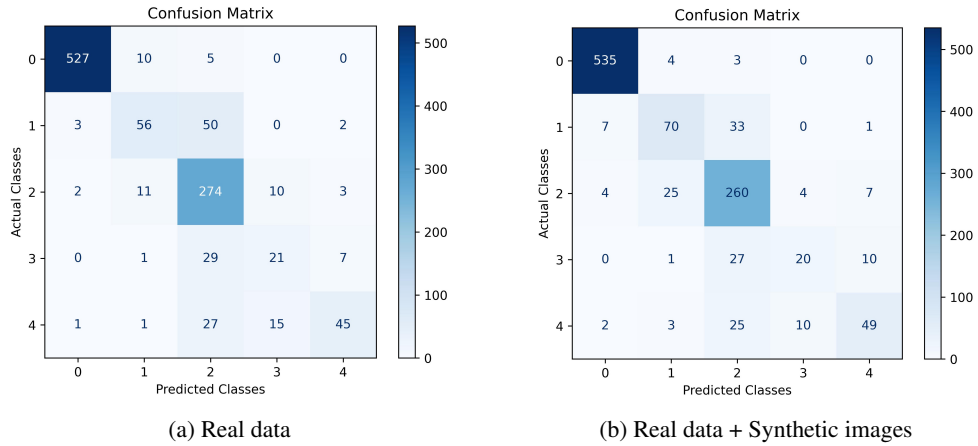


Figure 8.1: **Confusion matrices for the APTOS 2019 dataset showing the best results.** Class indices are defined as follows: 0 = No DR; 1 = Mild DR; 2 = Moderate DR; 3 = Severe DR; 4 = Proliferative DR.

DR and proliferative DR. However, the confusion matrix shows misclassifications in the moderate class.

Table 8.1 reports the quantitative results for the baseline and augmented models. The model already achieved strong baseline performance when trained solely on real data. However, the addition of synthetic images led to consistent improvements across all metrics, with the most notable gain observed in AUPR (+2.1%). This increase in AUPR may be attributed to the augmented dataset providing better representation of underrepresented DR classes, helping the model to discriminate positive cases more effectively and improving precision-recall performance.

Although the overall gains appear modest, they are meaningful given the high baseline performance, where further enhancements are inherently more challenging. These results suggest that the synthetic augmentation successfully contributed additional useful information, particularly benefiting the recognition of minority classes and enhancing the model’s robustness to class imbalance.

In terms of quality evaluation (Table 8.2), the synthetic images exhibit an Improved Precision (IP) of 0.58 and Improved Recall (IR) of 0.79, indicating moderate fidelity but a high level of diversity. Compared to the previous study on retinal image synthesis using Stable Diffusion, which reported IP = 0.50 and IR = 0.526 [19], it shows a small improvement in both realism and variability, demonstrating the effectiveness of controllable generation with ControlNet. The FID of 341.03 is relatively high, reflecting notable global distribution differences between real and synthetic images. Nevertheless, these synthetic samples still present substantial gains in diversity and perceptual quality (GIQA = 0.94), which are crucial for balancing underrepresented classes and supporting downstream training of diagnostic models.

Table 8.1: Classification performance of RETFound fine-tuning on the APTOS 2019 dataset using real data only versus real plus synthetic data. The best results are highlighted in bold.

Training Data	Accuracy	F1-score	AUROC	AUPR
Real only	0.8418	0.6919	0.9422	0.7214
Real + Synthetic	0.8491	0.7003	0.9475	0.7435

Table 8.2: Quality evaluation of synthetic images generated for the APTOS 2019 dataset using pyMDMA metrics.

Improved Precision (Fidelity)	Improved Recall (Diversity)	FID	GIQA (Perceptual)
0.58	0.79	341.03	0.94

8.2 IDRiD Dataset

For the IDRiD dataset, the best-performing results were achieved with the parameter configuration described in Table 7.5, without applying negative prompts (Chapter 7). A total of 15 experiments were conducted for the IDRiD dataset, comprising five fine-tuning configurations, each further evaluated with three variations of negative prompts. The parameter configuration reported here was selected as the best-performing among these experiments, as reflected in the UMAP visualization of real and synthetic feature distributions in Figure 8.4b.

This configuration generated the synthetic samples illustrated in Figure 8.6.

8.2.1 Confusion Matrices Analysis

Analysis of the confusion matrices for the IDRiD dataset (Figure 8.2) highlights the challenges posed by its highly imbalanced class distribution, with very few samples in the mild (class 1) and proliferative (class 4) categories. This scarcity led to poor classification performance for these underrepresented classes in the baseline model trained solely on real images (Figure 8.2a). Augmenting the dataset with synthetic images did not improve the classification of mild cases; however, a modest increase in correct predictions was observed for moderate (class 2) and severe (class 3) categories (Figure 8.2b).

The quantitative metrics in Table 8.3 indicate modest improvements in accuracy (+4.8%) and F1-score (+2.4%) when synthetic images were included in the training set. These gains, while limited, are meaningful given the challenging characteristics of the IDRiD dataset, which has fewer samples and a more pronounced class imbalance compared to APTOS. The slight decrease observed in AUROC and AUPR may result from the relatively small number of images of minority classes. Nevertheless, the increase in accuracy and F1-score suggests that synthetic augmentation provided additional useful information, particularly aiding the model in recognizing underrepresented DR classes. Overall, these results highlight the potential of controllable synthetic data to improve model robustness, even in small and imbalanced datasets.

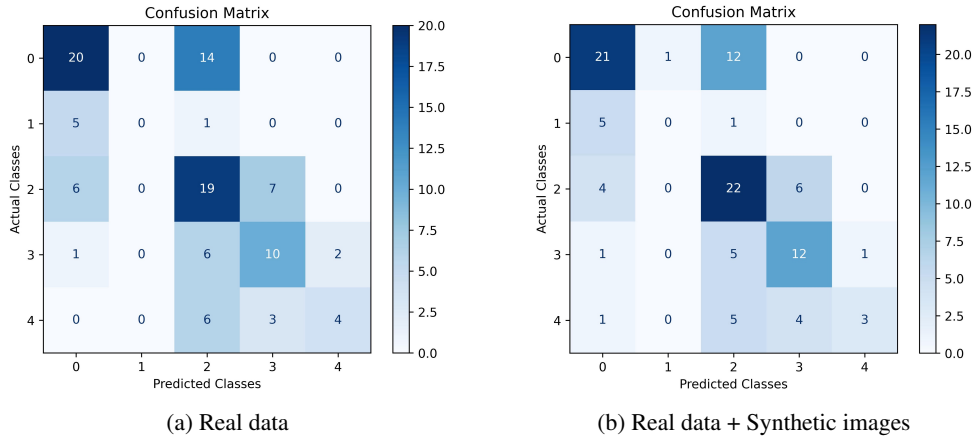


Figure 8.2: **Confusion matrices for the IDRiD dataset showing the best results.** Class indices are defined as follows: 0 = No DR; 1 = Mild DR; 2 = Moderate DR; 3 = Severe DR; 4 = Proliferative DR.

Table 8.3: Classification performance of RETFound fine-tuning on the IDRiD dataset using real data only versus real plus synthetic data. The best results are highlighted in bold.

Training Data	Accuracy	F1-score	AUROC	AUPR
Real only	0.5096	0.4054	0.8265	0.4995
Real + Synthetic	0.5577	0.4292	0.8196	0.4936

Examination of the quality metrics in Table 8.4, indicates that the synthetic images generated for the IDRiD dataset exhibit high fidelity, diversity, and perceptual quality. The Improved Precision (IP= 0.73) suggests that the generated images closely resemble real retinal images, while the Improved Recall (IR= 0.79) reflects good coverage of the variability present in the real dataset. The FID value (423.78) is higher than that reported for APTOS, which may be due to the smaller size of the IDRiD dataset and the resulting challenges in capturing the full distribution of real images.

Compared to previous retinal image synthesis study using Stable Diffusion, which reported IP = 0.50 and IR = 0.526 [19], these results demonstrate substantial improvements in both realism and variability. This highlights the effectiveness of controllable diffusion generation with ControlNet in producing high-quality synthetic retinal images, even in datasets with limited sample sizes.

Table 8.4: Quality evaluation of synthetic images generated for the IDRiD dataset using pyMDMA metrics.

Improved Precision (Fidelity)	Improved Recall (Diversity)	FID	GIQA (Perceptual)
0.73	0.79	423.78	0.73

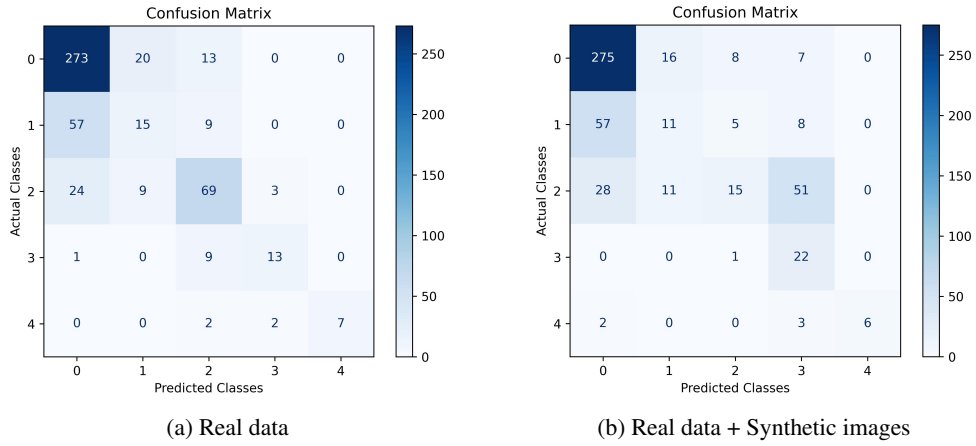


Figure 8.3: **Confusion matrices for the Messidor-2 dataset showing the best results.** Class indices are defined as follows: 0 = No DR; 1 = Mild DR; 2 = Moderate DR; 3 = Severe DR; 4 = Proliferative DR.

8.3 Messidor-2 Dataset

In the case of the Messidor-2 dataset, the outcomes presented were obtained using the inference settings in Table 7.4, without applying negative prompts (Chapter 7). A total of 15 experiments were performed for the Messidor-2 dataset, consisting of five fine-tuning configurations, each further tested with three variations of negative prompts. This configuration proved to be the most suitable for this dataset, as shown by the UMAP feature projection of real versus synthetic data in Figure 8.4c.

This configuration generated the synthetic samples illustrated in Figure 8.7.

8.3.1 Confusion Matrices Analysis

The confusion matrices for the Messidor-2 dataset (Figure 8.3) reveal a tendency for the model trained with both real and synthetic data to misclassify several cases as severe (class 3), indicating confusion between adjacent classes, particularly moderate (class 2).

Classification metrics presented in Table 8.5 indicate that the real-only model slightly outperforms the augmented model across most metrics, particularly in F1-score and AUPR. As a result, adding synthetic images to balance classes had a limited positive effect and, in some cases, introduced additional variability that may have slightly affected global performance metrics.

Table 8.5: Classification performance of RETFound fine-tuning on the Messidor-2 dataset using real data only versus real plus synthetic data. The best results are highlighted in bold.

Training Data	Accuracy	F1-score	AUROC	AUPR
Real only	0.7167	0.6289	0.8885	0.6889
Real + Synthetic	0.6255	0.4648	0.8789	0.06541

Evaluation of synthetic image quality (Table 8.6) shows that the synthetic images generated for the Messidor-2 dataset exhibit relatively low fidelity (IP= 0.58) but high perceptual quality (GIQA= 0.91) and diversity (IR= 0.79). The lower fidelity compared to IDRiD (IP= 0.73) may be attributed to the characteristics of the Messidor dataset, including the variability in image acquisition. The high IR value indicates that the synthetic images still capture a wide range of variability present in the real dataset.

Compared to previous study using Stable Diffusion for retinal image synthesis (IP = 0.50, IR = 0.526) [19], the results for Messidor-2 show improvement in both realism and variability, demonstrating that controllable generation with ControlNet can produce perceptually convincing images even when fidelity is limited.

Table 8.6: Quality evaluation of synthetic images generated for the Messidor-2 dataset using pyMDMA metrics.

Improved Precision (Fidelity)	Improved Recall (Diversity)	FID	GIQA (Perceptual)
0.58	0.79	592,81	0.91

8.4 Summary and Conclusions

This chapter presented a detailed evaluation of controllable diffusion for retinal fundus image synthesis across three public datasets: APTOS 2019, IDRiD, and Messidor-2. Both classification results and quality metrics were analyzed, complemented by confusion matrices to better capture the effect of synthetic augmentation on individual DR classes.

For APTOS 2019, the confusion matrices showed that augmentation improved recognition of minority classes, particularly mild and proliferative DR, although moderate DR remained prone to misclassification. Quantitatively, synthetic data led to consistent gains across all metrics, with the most notable improvement observed in AUPR, reflecting enhanced precision–recall performance. These improvements are meaningful given the already strong baseline model. Quality analysis revealed moderate fidelity but high diversity and perceptual quality, with better realism and variability compared to prior Stable Diffusion work.

For IDRiD, the confusion matrices highlighted severe difficulties with underrepresented classes such as mild and proliferative DR, which remained poorly classified even after augmentation. However, improvements were observed in moderate and severe DR, consistent with the quantitative gains in accuracy and F1-score. Quality evaluation indicated high fidelity and diversity, suggesting that synthetic images captured both structure and pathology, despite a relatively high FID reflecting distributional differences.

For Messidor-2, the confusion matrices showed a tendency of the augmented model to misclassify moderate DR as severe, indicating difficulties in discriminating adjacent classes. Quantitative results revealed that the real-only model slightly outperformed the augmented one. Despite this, quality evaluation confirmed that synthetic images achieved high perceptual quality and diversity, though fidelity remained low. Compared to previous generated images with Stable Diffusion,

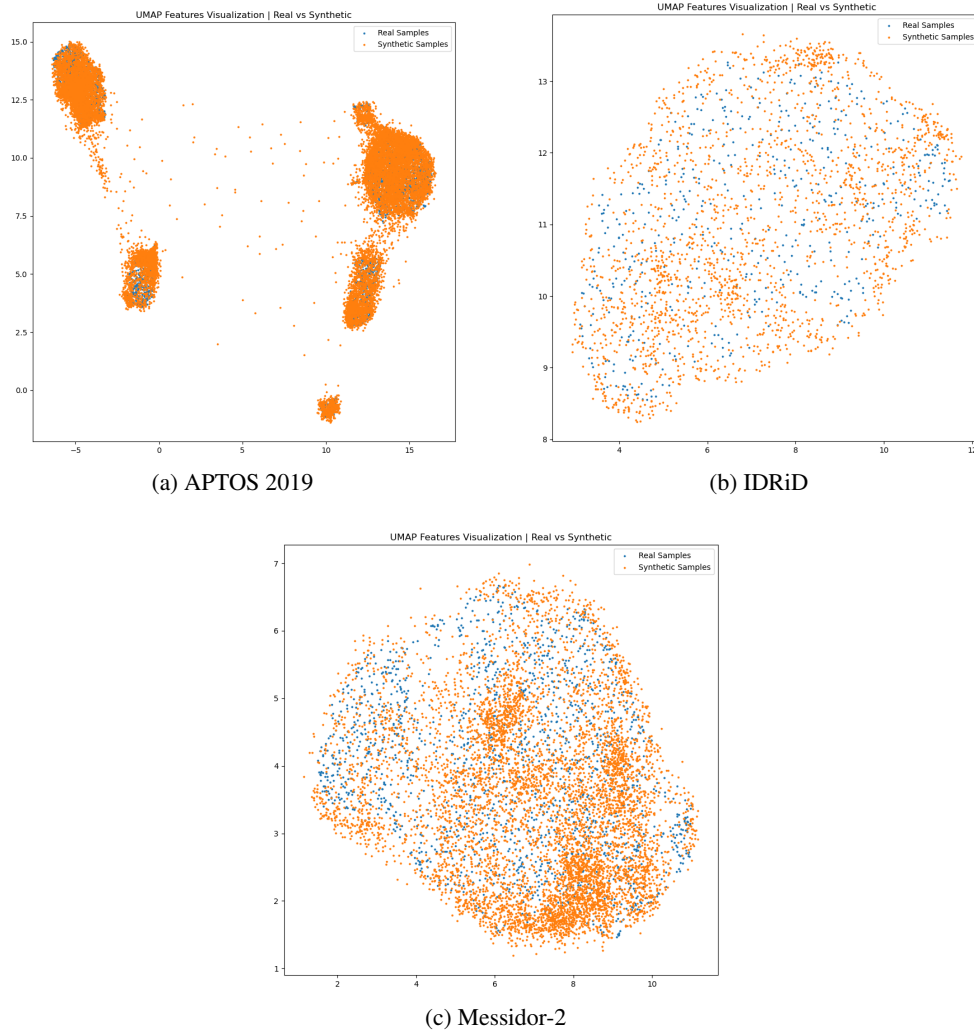


Figure 8.4: UMAP Visualization of Real vs Synthetic Features. This figure demonstrates the UMAP visualization of feature embeddings for real and synthetic retinal fundus images. Real samples are represented in the color blue and synthetic samples in orange, highlighting the overlap and alignment between datasets. This analysis supports the evaluation of fidelity and diversity in synthetic data generation and aids in identifying discrepancies or gaps in feature representation.

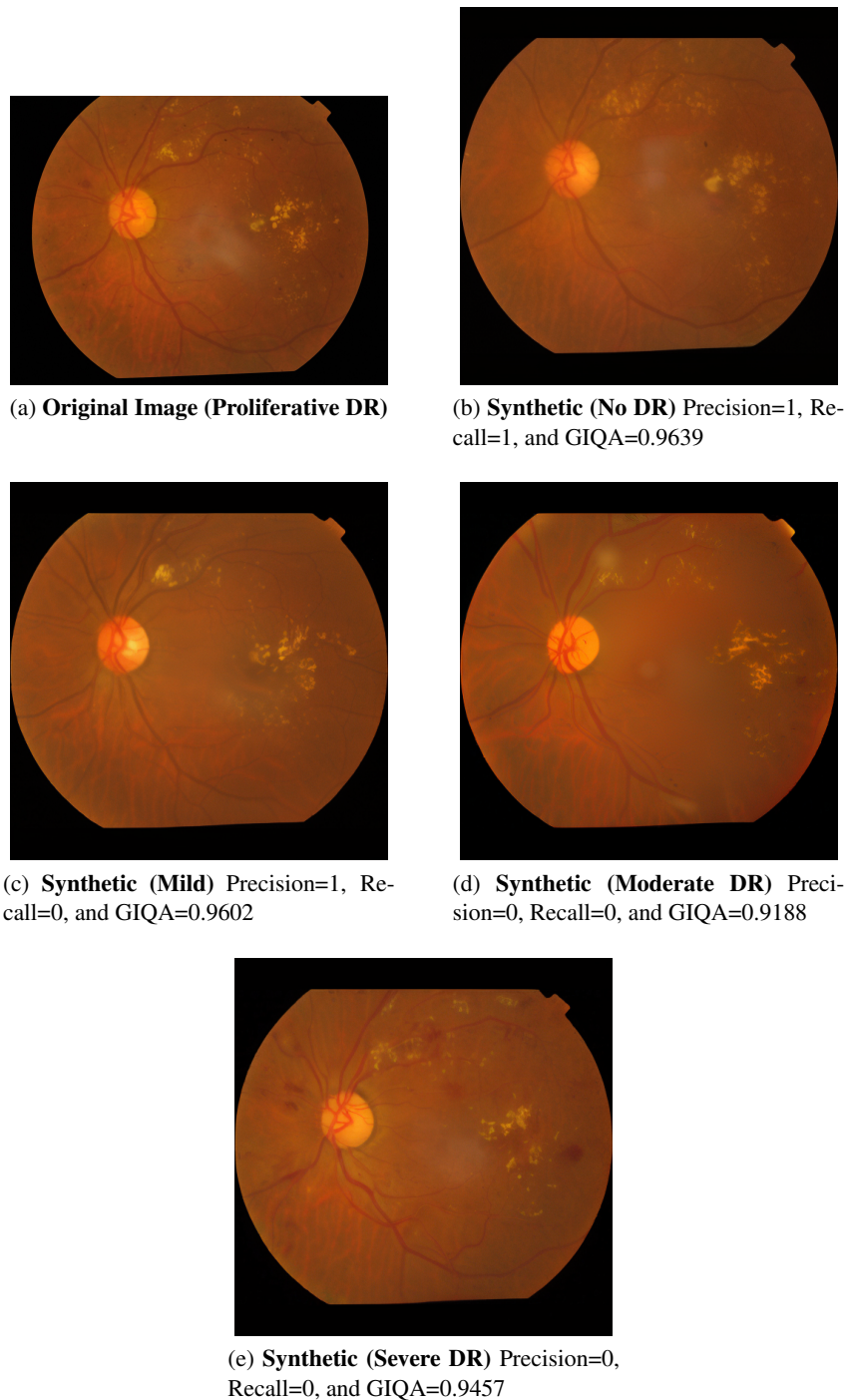


Figure 8.5: **Illustrative example of synthetic retinal fundus generation from APTOS 2019 dataset.** Real fundus image with its respective class, together with the synthetic images generated from it. Each image is annotated with its class label and associated evaluation metrics (precision, recall, and GIQA), enabling a direct assessment of visual quality.

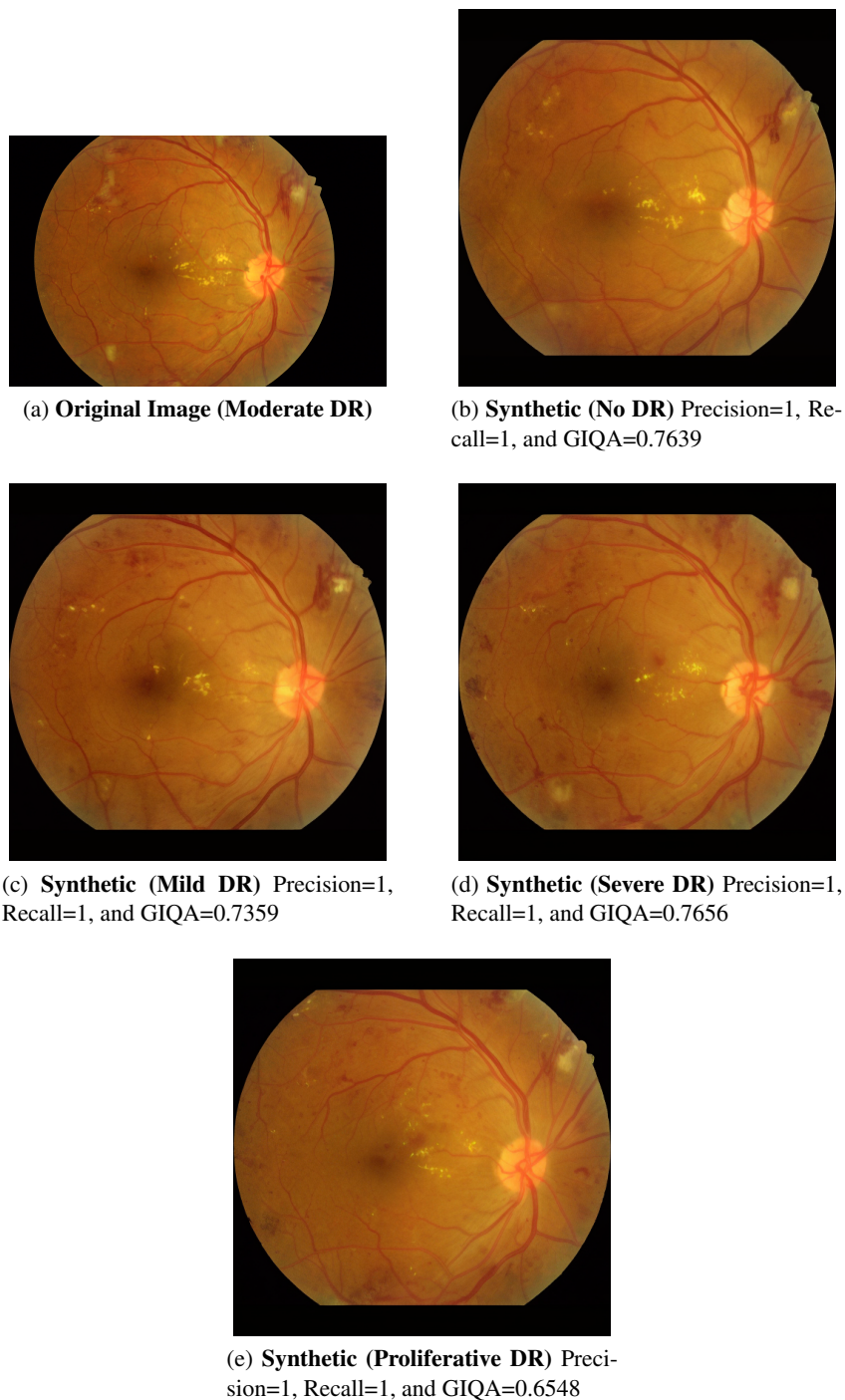


Figure 8.6: **Illustrative example of synthetic retinal fundus generation from IDRiD dataset.** Real fundus image with its respective class, together with the synthetic images generated from it. Each image is annotated with its class label and associated evaluation metrics (precision, recall, and GIQA), enabling a direct assessment of visual quality.

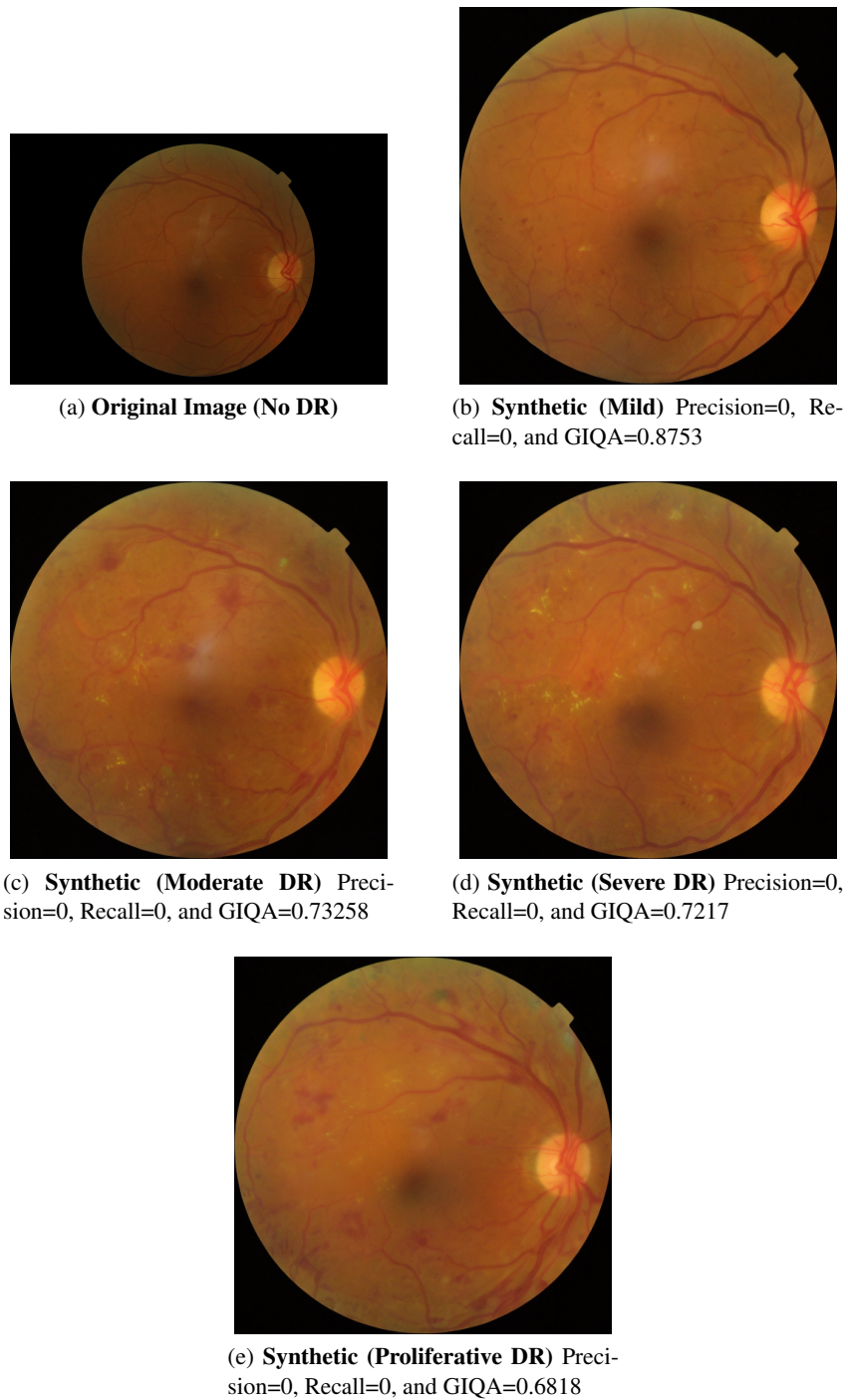


Figure 8.7: **Illustrative example of synthetic retinal fundus generation from Messidor-2 dataset.** Real fundus image with its respective class, together with the synthetic images generated from it. Each image is annotated with its class label and associated evaluation metrics (precision, recall, and GIQA), enabling a direct assessment of visual quality.

improvements were still observed in realism and variability, demonstrating that controllable generation can produce perceptually convincing images even in challenging datasets.

In summary, confusion matrix analysis revealed that synthetic augmentation is particularly effective in supporting minority classes, though challenges remain in consistently improving all DR severity levels. Combined with quantitative and quality metrics, the results confirm that controllable diffusion enhances diversity and clinical plausibility of synthetic images, but its utility in classification strongly depends on dataset characteristics and class distribution.

Chapter 9

Conclusion

This thesis investigated the synthetic generation of retinal fundus images using a controllable diffusion model, with a particular focus on their application in diabetic retinopathy diagnosis and classification. The work was motivated by the pressing need for large-scale, balanced, and diverse datasets, as real-world clinical data often suffer from limitations such as class imbalance, scarcity of annotated cases, and privacy concerns.

A key contribution of this work lies in demonstrating the value of controllability in diffusion-based generation, achieved through the integration of ControlNet into the Stable Diffusion backbone. By conditioning the generative process on anatomical masks and disease-specific prompts, ControlNet enabled the synthesis of retinal fundus images that not only preserved the global retinal structure but also reflected targeted pathological features associated with different DR severity levels. This level of control was particularly important for addressing class imbalance, as it allowed the focused generation of underrepresented categories while maintaining clinical plausibility.

The evaluation framework combined fidelity, diversity, and utility metrics. Utility was assessed through the fine-tuning of RETFound across three public datasets: APTOS 2019, IDRiD, and Messidor-2. The results highlighted that there is no single optimal configuration for synthetic augmentation. Instead, outcomes varied according to dataset characteristics, including class distribution and image quality.

On APTOS 2019, synthetic augmentation consistently improved performance, reflecting improved recognition of minority classes. For IDRiD, modest but meaningful gains in accuracy and F1-score demonstrated the potential of synthetic augmentation to support learning in small and highly imbalanced datasets. For Messidor-2, augmentation did not improve performance and sometimes led to confusion between adjacent severity levels, showing that synthetic images may also introduce variability that does not always benefit classification.

Quality evaluation confirmed that synthetic images exhibited strong diversity and perceptual quality across all datasets. The results achieved showed clear improvements in both fidelity and variability, underscoring the value of controllability. However, fidelity varied between datasets: higher in IDRiD, moderate in APTOS, and lower in Messidor-2, reflecting the sensitivity of generative performance to dataset size and imaging conditions.

Overall, this thesis provides evidence that diffusion-based synthetic generation, when coupled with controllability mechanisms, holds substantial promise for medical imaging. However, the effectiveness of synthetic augmentation depends on fine-grained parameter tuning, preprocessing strategies, and an awareness of dataset-specific characteristics.

9.1 Future Work

While this thesis has demonstrated the potential of controllable diffusion models for synthetic retinal fundus image generation, several directions remain open for future exploration. A key avenue lies in extending the level of control during synthesis, particularly regarding pathological findings. Current controllability enabled alignment with global retinal structures and disease severity levels, but finer control over specific lesions, such as microaneurysms, hemorrhages, and exudates, would substantially improve the clinical relevance of synthetic data.

Incorporating prior knowledge about the prevalence, spatial distribution, and co-occurrence of lesions across different stages of diabetic retinopathy would allow the generation of more accurate and representative synthetic classes. This would result in datasets that better reflect the true structure of disease progression and improve the reliability of synthetic augmentation.

Also, fine-grained control of lesions could enable the targeted synthesis of clinical cases that are underrepresented or even absent in available datasets. For instance, if severe cases with abundant hard exudates are often missing in public collections, limiting the capacity of diagnostic models to learn from such scenarios. The ability to enforce these specific lesion patterns during image generation would fill these gaps, creating richer datasets and supporting the training of models robust to rare but clinically significant conditions.

References

- [1] MD Abdhish R. Bhavsar. Diabetic retinopathy clinical presentation, 2024. Accessed: 2024-10-27.
- [2] Michael D Abramoff, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA ophthalmology*, 131(3):351–357, 2013.
- [3] ADCIS. Messidor: Methods for evaluating segmentation and indexing techniques dedicated to retinal ophthalmology, 2014. Accessed: 2024-12-20.
- [4] Zahra Amini, Rahele Kafieh, Elaheh Mousavi, and Hossein Rabbani. 13 - diabetic retinopathy detection in ocular imaging by dictionary learning. In Ayman S. El-Baz and Jasjit S. Suri, editors, *Diabetes and Fundus OCT*, Computer-Assisted Diagnosis, pages 343–378. Elsevier, 2020.
- [5] P Andreini, S Bonechi, M Bianchini, A Mecocci, F Scarselli, and A Sodi. A two stage gan for high resolution retinal image generation and segmentation. arxiv 2019. *arXiv preprint arXiv:1907.12296*, 2023.
- [6] SOPHIE J. BAKRI and PETER K. KAISER. Chapter 21 - diabetic retinopathy. In David Huang, Peter K. Kaiser, Careen Y. Lowder, and Elias I. Traboulsi, editors, *Retinal Imaging*, pages 233–240. Mosby, Philadelphia, 2006.
- [7] Ana Barragán-Montero, Umair Javaid, Gilmer Valdés, Dan Nguyen, Paul Desbordes, Benoit Macq, Siri Willems, Liesbeth Vandewinckele, Mats Holmström, Fredrik Löfman, et al. Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Medica*, 83:242–256, 2021.
- [8] Artur Gomes Barreto, Juliana Martins de Oliveira, Francisco Nauber Bernardo Gois, Paulo Cesar Cortez, and Victor Hugo Costa de Albuquerque. A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks. *Bioengineering*, 10(9):1098, 2023.
- [9] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.
- [10] Jasper Kyle Catapang, Darby E Santiago, and Iris Thiele Isip-Tan. Improving detection of diabetic retinopathy in low-resolution images via latent diffusion. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 53–58. IEEE, 2023.

- [11] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.
- [12] Mark A Chia, Fares Antaki, Yukun Zhou, Angus W Turner, Aaron Y Lee, and Pearse A Keane. Foundation models in ophthalmology. *British Journal of Ophthalmology*, 2024.
- [13] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abramoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017.
- [14] Coen De Vente, Koenraad A Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, et al. Airops: Artificial intelligence for robust glaucoma screening challenge. *IEEE transactions on medical imaging*, 43(1):542–557, 2023.
- [15] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pages 231–234, 2014.
- [16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [17] Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. <https://kaggle.com/competitions/diabetic-retinopathy-detection>, 2015. Kaggle.
- [18] Jan Ehrhardt and Matthias Wilms. Autoencoders and variational autoencoders in medical image analysis. In *Biomedical Image Synthesis and Simulation*, pages 129–162. Elsevier, 2022.
- [19] Ivo Facoco, Goncalo Mesquita, Francisca Lucio, and Luis Rosado. Adapting stable diffusion models for domain-specific medical imaging: A case study in synthetic retinal fundus image generation. *ECML-PKDD*, Springer, 2025.
- [20] Ivo S Façoco, Joana Rebelo, Pedro Matias, Nuno Bento, Ana C Morgado, Ana Sampaio, Luís Rosado, and Marília Barandas. pymdma: Multimodal data metrics for auditing real and synthetic datasets. *SoftwareX*, 31:102256, 2025.
- [21] Zehao Fang et al. Conditional diffusion model for x-ray segmentation data generation. *Journal of Artificial Intelligence Practice*, 7(1):7–10, 2024.
- [22] Jimena Fernández-Carneado, Ana Almazán-Moga, Dolores T Ramírez-Lamelas, Cristina Cuscó, José Ignacio Alonso de la Fuente, J Carlos Pastor, María Isabel López Gálvez, and Berta Ponsati. Quantification of microvascular lesions in the central retinal field: could it predict the severity of diabetic retinopathy? *Journal of Clinical Medicine*, 12(12):3948, 2023.
- [23] Donald S Fong, Lloyd P Aiello, Frederick L Ferris III, and Ronald Klein. Diabetic retinopathy. *Diabetes care*, 27(10):2540–2553, 2004.
- [24] Barbara Davis Center for Diabetes. Discussion and pictures of diabetic retinopathy lesions, 2024. Accessed: 2025-01-10.

- [25] Timothy HM Fung, Bakula Patel, Emma G Wilmot, and Winfried MK Amoaku. Diabetic retinopathy for the non-ophthalmologist. *Clinical Medicine*, 22(2):112–116, 2022.
- [26] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [28] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giga: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 369–385. Springer, 2020.
- [29] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Priorgan: Real data prior for generative adversarial nets. *arXiv preprint arXiv:2006.16990*, 2020.
- [30] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. *arXiv preprint arXiv:2409.11169*, 2024.
- [31] John T Hancock, Taghi M Khoshgoftaar, and Justin M Johnson. Evaluating classifier performance with highly imbalanced big data. *Journal of Big Data*, 10(1):42, 2023.
- [32] Siti Syafinah Ahmad Hassan, David BL Bong, and Mallika Premseenthil. Detection of neo-vascularization in diabetic retinopathy. *Journal of digital imaging*, 25:437–444, 2012.
- [33] Mohammad Havaei, Ximeng Mao, Yiping Wang, and Qicheng Lao. Conditional generation of medical images via disentangled adversarial inference. *Medical image analysis*, 72:102106, 2021.
- [34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [35] Alejandro Ungria Hirte, Moritz Platscher, Thomas Joyce, Jeremy J Heit, Eric Tranvinh, and Christian Federau. Realistic generation of diffusion-weighted magnetic resonance brain images with deep generative models. *Magnetic Resonance Imaging*, 81:60–66, 2021.
- [36] Indu Ilanchezian, Valentyn Boreiko, Laura Kühlewein, Ziwei Huang, Murat Seçkin Ayhan, Matthias Hein, Lisa Koch, and Philipp Berens. Development and validation of an ai algorithm to generate realistic and meaningful counterfactuals for retinal imaging based on diffusion models. *PLOS Digital Health*, 4(5):e0000853, 2025.
- [37] Shahzaib Iqbal, Tariq M. Khan, Khuram Naveed, Syed S. Naqvi, and Syed Junaid Nawaz. Recent trends and advances in fundus image analysis: A review. *Computers in Biology and Medicine*, 151:106277, 2022.
- [38] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

- [39] Satya Jena and Kalyan Tripathy. *Vitreous Hemorrhage*. StatPearls [Internet]. StatPearls Publishing, Treasure Island, FL, 2024. [Updated 2023 Aug 25].
- [40] Venkata M. Kanukollu and Saba S. Ahmad. *Retinal Hemorrhage*. StatPearls [Internet]. StatPearls Publishing, Treasure Island, FL, 2024. [Updated 2023 Aug 8].
- [41] Sinan Kaplan, Lasse Lensu, Lauri Laaksonen, and Hannu Uusitalo. Evaluation of unconditioned deep generative synthesis of retinal images. In *Advanced Concepts for Intelligent Vision Systems: 20th International Conference, ACIVS 2020, Auckland, New Zealand, February 10–14, 2020, Proceedings 20*, pages 262–273. Springer, 2020.
- [42] Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. <https://kaggle.com/competitions/aptos2019-blindness-detection>, 2019. Kaggle.
- [43] Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U Gamble, Hari M Trivedi, Cody C Wyles, Andrew B Sellergren, Saptarshi Purkayastha, Bradley J Erickson, et al. Synthetically enhanced: unveiling synthetic data’s potential in medical imaging research. *EBioMedicine*, 104, 2024.
- [44] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [45] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024.
- [46] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.
- [47] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.
- [48] Badia Ouissam Lakas, Chemousse Berdjouh, Khadra Bouanane, Mohammed Lamine Kherfi, Oussama Aiadi, Lamri Laouamer, and Abbas Cheddad. Enhancing diabetic retinopathy grading with advanced diffusion models. In *International Congress on Information and Communication Technology*, pages 215–227. Springer, 2024.
- [49] Kristine E Lee, Barbara Eden Kobrin Klein, Ronald Klein, and Stacy M Meuer. Association of retinal vessel caliber to optic disc and cup diameters. *Investigative ophthalmology & visual science*, 48(1):63–67, 2007.
- [50] Haijun Lei, Zhihui Tian, Hai Xie, Benjian Zhao, Xianlu Zeng, Jiuwen Cao, Weixin Liu, Jiantao Wang, Guoming Zhang, Shuqiang Wang, et al. Lac-gan: Lesion attention conditional gan for ultra-widefield image synthesis. *Neural Networks*, 158:89–98, 2023.
- [51] Jing Li. Area under the roc curve has the most consistent evaluation for binary classification. *PloS one*, 19(12):e0316019, 2024.
- [52] Yuheng Li. Towards controllable diffusion models with gligen, 2023. Accessed: 2024-12-17.

- [53] Kang Lin and Reinhard Heckel. Robustness of deep learning for accelerated mri: Benefits of diverse training data. *arXiv preprint arXiv:2312.10271*, 2023.
- [54] Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G Baraniuk, and Michael J Henry. Evaluating generative networks using gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 279–288, 2023.
- [55] Matthew McDermott, Haoran Zhang, Lasse Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at auroc and auprc under class imbalance. *Advances in Neural Information Processing Systems*, 37:44102–44163, 2024.
- [56] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.
- [57] Team Mosaic ML. composer. <https://github.com/mosaicml/composer/>, 2021.
- [58] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.
- [59] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarbuerger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.
- [60] Loc X Nguyen, Pyae Sone Aung, Huy Q Le, Seong-Bae Park, and Choong Seon Hong. A new chapter for medical image generation: the stable diffusion method. In *2023 International Conference on Information Networking (ICOIN)*, pages 483–486. IEEE, 2023.
- [61] Jamie A O’Reilly and Fawad Asadi. Pre-trained vs. random weights for calculating fréchet inception distance in medical imaging. In *2021 13th Biomedical Engineering International Conference (BMEiCON)*, pages 1–4. IEEE, 2021.
- [62] Achraf Oussidi and Azeddine Elhassouny. Deep generative models: Survey. In *2018 International conference on intelligent systems and computer vision (ISCV)*, pages 1–8. IEEE, 2018.
- [63] Julia Palik, Mauricio Rivera Celestino, David Gomez-Triana, Nicholas Marsh, and Ida Rødningen. Disarmament, demobilization, and reintegration in peace agreements (1975–2021): Introducing the ddr dataset. *Journal of Peace Research*, page 00223433241294148, 2025.
- [64] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.
- [65] Suraj Patil, Pedro Cuenca, Nathan Lambert, and Patrick von Platen. Stable diffusion with diffusers. *Hugging Face Blog*, 2022. [<https://huggingface.co/blog/rlhf>](https://huggingface.co/blog/stable_diffusion).

- [66] Haonan Peng, Shan Lin, Daniel King, Yun-Hsuan Su, Waleed M Abuzeid, Randall A Bly, Kris S Moe, and Blake Hannaford. Reducing annotating load: Active learning with synthetic images in surgical instrument segmentation. *Medical Image Analysis*, 97:103246, 2024.
- [67] Mehran Pesteie, Purang Abolmaesumi, and Robert N Rohling. Adaptive augmentation of medical data using independently conditional variational auto-encoders. *IEEE transactions on medical imaging*, 38(12):2807–2820, 2019.
- [68] Walter Hugo Lopez Pinaya. Controllable medical image generation with controlnets: Guide on using controlnets to control the generation process of latent diffusion models, 2023. Accessed: 2024-10-27.
- [69] Simon Playe. Generative ai: Learn how to control image generation with stable diffusion, 2024. Accessed: 2024-12-16.
- [70] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018.
- [71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [72] Pedro Romero-Aroca, Eugeni Garcia-Curto, Jordi Pascual-Fontanilles, Aida Valls, Antonio Moreno, and Marc Baget-Bernaldiz. Distribution of microaneurysms and hemorrhages in accordance with the grading of diabetic retinopathy in type diabetes patients. *Diagnostics*, 14(14):1547, 2024.
- [73] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir 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*, pages 22500–22510, 2023.
- [74] Amir Safarpour, Shivam Kalra, and Hamid R Tizhoosh. Generative models in pathology: synthesis of diagnostic quality pathology images. *The Journal of Pathology*, 253(2):131–132, 2021.
- [75] Mariko Sasaki, Ryo Kawasaki, Jonathan E Noonan, Tien Yin Wong, Ecosse Lamoureux, and Jie Jin Wang. Quantitative measurement of hard exudates in patients with diabetes and their associations with serum lipid levels. *Investigative ophthalmology & visual science*, 54(8):5544–5550, 2013.
- [76] Turab Selcuk and Ahmet Alkan. Detection of microaneurysms using ant colony algorithm in the early diagnosis of diabetic retinopathy. *Medical hypotheses*, 129:109242, 2019.
- [77] Fangxin Shang, Jie Fu, Yehui Yang, Haifeng Huang, Junwei Liu, and Lei Ma. Synfundus-1m: A high-quality million-scale synthetic fundus images dataset with fifteen types of annotation. *arXiv preprint arXiv:2312.00377*, 2023.
- [78] Dereje Shenkut and Vijayakumar Bhagavatula. Fundus gan-gan-based fundus image synthesis for training retinal image classifiers. In *2022 44th annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC)*, pages 2185–2189. IEEE, 2022.
- [79] Sadi Can Sonmez, Mertcan Sevgi, Fares Antaki, Josef Huemer, and Pearse A Keane. Generative artificial intelligence in ophthalmology: current innovations, future applications and challenges. *British Journal of Ophthalmology*, 108(10):1335–1340, 2024.

- [80] Fu Steins. Stable diffusion controlnet clearly explained, 2023. Accessed: 2024-11-03.
- [81] Nathan Lambert Patrick von Platen Suraj Patil, Pedro Cuenca. Stable diffusion with diffusers, 2022. Accessed: 2025-07-04.
- [82] Benjamin Sykes, Loic Simon, and Julien Rabin. Unifying and extending precision recall metrics for assessing generative models. *arXiv preprint arXiv:2405.01611*, 2024.
- [83] Haileleol Tibebe, Aadin Malik, and Varuna De Silva. Text to image synthesis using stacked conditional variational autoencoders and conditional generative adversarial networks. In *Science and Information Conference*, pages 560–580. Springer, 2022.
- [84] Bambang Krismono Triwijoyo, Boy Subirosa Sabarguna, Widodo Budiharto, and Edi Abdurachman. 2 - deep learning approach for classification of eye diseases based on color fundus images. In Ayman S. El-Baz and Jasjit S. Suri, editors, *Diabetes and Fundus OCT*, Computer-Assisted Diagnosis, pages 25–57. Elsevier, 2020.
- [85] Hristina Uzunova, Sandra Schultz, Heinz Handels, and Jan Ehrhardt. Unsupervised pathology detection in medical images using conditional variational autoencoders. *International journal of computer assisted radiology and surgery*, 14:451–461, 2019.
- [86] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022.
- [87] Charles P Wilkinson, Frederick L Ferris III, Ronald E Klein, Paul P Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R Pararajasegaram, Juan T Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003.
- [88] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P Yung, Ankit B Patel, et al. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2024.
- [89] Yu Xia, Chi-Hua Wang, Joshua Mabry, and Guang Cheng. Advancing retail data science: Comprehensive evaluation of synthetic data. *arXiv preprint arXiv:2406.13130*, 2024.
- [90] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.
- [91] Zhiyong Yang, Qianqian Xu, Shilong Bao, Xiaochun Cao, and Qingming Huang. Learning with multiclass auc: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7747–7763, 2021.
- [92] Aram You, Jin Kuk Kim, Ik Hee Ryu, and Tae Keun Yoo. Application of generative adversarial networks (gan) for ophthalmology image domains: a survey. *Eye and Vision*, 9(1):6, 2022.
- [93] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, and Vishal M Patel. Scenecomposer: Any-level semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22468–22478, 2023.

- [94] Juzhao Zhang, Senlin Lin, Tianhao Cheng, Yi Xu, Lina Lu, Jiangnan He, Tao Yu, Yajun Peng, Yuejie Zhang, Haidong Zou, et al. Retfound-enhanced community-based fundus disease screening: real-world evidence and decision curve analysis. *NPJ digital medicine*, 7(1):108, 2024.
- [95] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [96] Tao Zhou. Discriminating abilities of threshold-free evaluation metrics in link prediction. *Physica A: Statistical Mechanics and its Applications*, 615:128529, 2023.
- [97] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3):818–828, 2021.
- [98] Yi Zhou, Boyang Wang, Lei Huang, Shanshan Cui, and Ling Shao. A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3):818–828, 2020.
- [99] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.