Algerian Democratic and Popular Republic
Ministry of Higher Education and Scientific Research

KASDI MERBAH UNIVERSITY OUARGLA
Faculty of New Information and Communication Technologies
Department of Computer Science and Information Technology

Thesis Submitted in Candidacy for a Master Degree in computer science, option

Artificial Intelligence and Data Science

by Chemousse Berdjouh & Badia Ouissam Lakas

Diffusion Models for Data Augmentation of Medical Images

Jury Members:

Prof.	Mohamed El-Amine Abderrahim	Jury Chair	UKM Ouargla
Dr.	Khadra Bouanane	Supervisor	UKM Ouargla
Dr.	Belal Khaldi	Reviewer	UKM Ouargla

Academic year: 2022/2023

Acknowledgment

To each and every person who has helped us throughout the journey of completing our Master's degree graduation thesis, we extend our deepest gratitude.

First, we would like to express our sincerest gratitude to our supervisor, ***DR. Khadra Bouanane****, for their invaluable guidance, steadfast support, and expertise throughout the entirety of the research process. Their mentorship and thought-provoking feedback have been integral in determining the trajectory and caliber of this thesis.*

We would like to wholeheartedly thank the faculty members and staff at ***The Department of Computer Science and Information Technology*** *for their unwavering dedication to academic excellence and their tireless efforts in establishing a stimulating learning environment.*

Lastly, we would like to recognize the wider ***scientific community and experts in our field*** *whose groundbreaking work and publications provided the basis for this study. Their contributions have been essential in fostering progress in knowledge and were instrumental in configuring the research landscape.*

In closing, we are profoundly grateful to all those who contributed in any way to the successful completion of our Master's degree graduation thesis. Their assistance, leadership, and encouragement have been invaluable, and we are immensely privileged to have interacted with such remarkable people.

Thank you all.

Badia Ouissam Lakas & Chemousse Berdjouh

Dedication

In the pursuit of this monumental work that has consumed a significant portion of my academic journey, I attribute its accomplishment to the unwavering guidance of the Lord God. Thus, first and foremost, I offer my heartfelt gratitude to God.

I am deeply indebted to my beloved mother and my supportive father, whose unwavering encouragement and support have been my constant companion throughout my educational years. They have shared in both my moments of joy and disappointment, and I am profoundly grateful for their steadfast presence. Their unwavering belief in me and their unwavering support has propelled me to discover my true potential, become the best version of myself, and exert the utmost effort in all my endeavors.

*I would also like to express my profound gratitude to my siblings "**Anfal**", "**Mohamed**" and "**zinou**", whose unwavering support and inspiration have been a source of strength and motivation. Their presence by my side has been instrumental in my journey, and I am grateful for their unwavering support. I cannot overlook the role of my feline companion, "**Mousha**", who has been a constant presence in my life, faithfully waking up with me each morning.*

*In expressing my gratitude, I cannot end without mentioning the wonderful friends and classmates, "**Chemousse**" and "**Anfel**". Their presence and friendship have been invaluable to me throughout this journey.*

Furthermore, I extend my appreciation to every individual who has influenced my academic and professional journey, to all those who have played a part in my educational and professional development, instilling in me the belief that the bitter taste of hardship is temporary and will eventually yield a sweeter reward. I offer my sincerest gratitude.

Badia Ouissam

To embark on this profound journey that has consumed a significant portion of my academic life, I am deeply indebted to the divine guidance of the Lord God. Therefore, I begin by expressing my utmost gratitude to God.

I extend my heartfelt appreciation to my beloved mother and my ever-supportive father "May Allah accept him in His boundless mercy", whose unwavering love and encouragement have been my guiding light throughout my educational years. They have stood by me through every triumph and tribulation, sharing in the joys and sorrows of my educational pursuits. Their unwavering belief in my potential and their unwavering support have propelled me to unearth my true capabilities and strive for excellence with unwavering determination.

I would be remiss not to express my deepest appreciation to all those individuals who have played a role in shaping my educational and professional endeavors. Their wisdom, guidance, and unwavering belief in my abilities have fueled my determination and pushed me beyond my limits. Despite the hardships and challenges faced along the way, their unwavering faith in my potential has served as a constant reminder that perseverance leads to eventual triumph.

To each person who has contributed to my growth, education, and professional development, I offer my sincere and heartfelt gratitude. This dedication is a testament to the immense impact that love, support, and unwavering faith can have on one's journey. May it serve as a reminder to all that our collective efforts and belief in one another can pave the way for remarkable achievements.

With deep gratitude and profound appreciation, I dedicate this work to all those who have played a part in shaping my path. May their unwavering support and belief in me be forever etched in my heart as a reminder of the transformative power of love and encouragement.

Chemousse

Abstract

Recently, there has been a substantial surge in interest surrounding diffusion probabilistic models, which are considered a prominent class of generative models, particularly in the realm of deep learning. These models have garnered significant attention due to their potential applications in a range of deep-learning problems.

The primary objective of this thesis is to assess the effectiveness of Diffusion models as a data augmentation technique in the context of medical image analysis. Furthermore, it aims to conduct a comparative analysis of the performance exhibited by deep-based classifiers trained on two distinct datasets. One dataset was augmented using diffusion models, while the other dataset underwent traditional data augmentation techniques.

Utilizing the IDRID dataset for the purpose of diabetic retinopathy diagnosis, the acquired outcomes substantiate the efficacy of Diffusion models as a data augmentation methodology for medical images, in contrast to the traditional data augmentation technique which is predominantly employed. The integration of diffusion model augmented data yielded superior performance for both classifiers, namely the Fine-tuned Resnet50 and the proposed CNN, surpassing the performance of classifiers trained using traditional data augmentation.

KEYWORDS

Diffusion Models, Data augmentation, Diabetic retinopathy, deep learning classifier, IDRID dataset, Medical images.

ملخص

مؤخرا، ارتفع الاهتمام حول نملذج الانتشار الاحتمالية، التي تصنف تحت نماذج التوليد تحت مجال التعلم العميق. هذا الاهتمام كان بسبب استخدامها في عدة تطبيقات في مجال التعلم الآلي. الهدف الأساسي من هذه المذكرة هي التجربة و التأكد من فعالية استعمال نماذج الانتشار ك تقنية لزيادة البانات في المجالات الطبية. ايضا، الهدف هو اجراء مقارنة بين اداء نموذج يعتمد على التعلم العميق متعلم ببيانات متزايدة عن طريق نموذج الانتشار الاحتمالي و نفس النموذج مدرب ببيانات متزايدة بالطريقة التقليدية المعتمدة لزيادة الصور الطبية. استخدام مجموعة بيانات IDRID في تشخيص مرض السكري في شبكية العين، تظهر النتائج المحصل عليها فعالية نماذج الانتشار كتقنية لزيادة البيانات في صور الطبية، بالمقارنة مع تقنية زيادة البيانات التقليدية التي تستخدم بشكل رئيسي. تكامل بيانات الزيادة المعززة بنماذج الانتشار أدى إلى تحقيق أداء متفوق لكلا النمطين المصنفين، وهما Resnet50 المعدل بشكل دقيق والشبكة العصبية المقترحة، متفوقاً على أداء المصنفات المدربة باستخدام زيادة البيانات التقليدية

الكلمات المفتاحية: نموذج الانتشار الاحتمالي، زيادة البيانات، مرض السكري في شبكة العين، مصنف عبر التعلم العميق، الصور الطبية ، IDRID dataset

RÉSUMÉ

Dans les années dernières,il y a eu une augmentation substantielle de l'intérêt entourant les modèles probabilistes de diffusion, qui sont considérés comme une classe de modèles génératifs, en particulier dans le domaine de l'apprentissage profond. Ces modèles ont suscité une attention considérable en raison de leurs applications éventuelles dans une gamme de problèmes d'apprentissage profond.

L'objectif principal de cette thèse est d'évaluer l'efficacité des modèles de diffusion en tant que technique d'augmentation de données dans le contexte de l'analyse d'images médicales. En outre, l'objectif est de réaliser une analyse comparative des performances des classifieurs basés sur l'apprentissage profond entraînés sur deux ensembles de données distincts. Un ensemble de données a été augmenté en utilisant des modèles de diffusion, tandis que l'autre ensemble de données a été soumis à des techniques d'augmentation de données traditionnelles.

En utilisant IDRID dataset dans le but du diagnostic de la rétinopathie diabétique, les résultats obtenus confirment l'efficacité des modèles de diffusion en tant que méthodologie d'augmentation de données pour les images médicales, par rapport à la technique traditionnelle d'augmentation de données qui est principalement utilisée. L'intégration de données augmentées par des modèles de diffusion a conduit à des performances supérieures pour les deux classifieurs, à savoir le Resnet50 fine-tuned et le CNN proposé, dépassant les performances des classifieurs entraînés à l'aide de l'augmentation de données traditionnelle.

MOTS-CLÉS:

Les modèles de diffusion, L'augmentation de données, La rétinopathie diabétique, Classifieur d'apprentissage profond, IDRID dataset, Les images medicales.

Contents

List of Figures

General Introduction

Machine learning (ML) has significantly impacted various domains of our modern society over the past few decades [3] where it has demonstrated cognitive capabilities comparable to humans and has even surpassed human performance, exhibiting superhuman abilities [4, 5]. As a result, both ML and its subset, deep learning (DL), have found increasing applications in the medical field [3], and the ability to automatically process medical images has led researchers to develop systems for automated analysis and diagnosis [6]. This advancement has paved the way for the creation of efficient and accurate tools to enhance the capability of the tasks of medical image analysis [6]. The success of machine learning in the medical field can be attributed to the availability of high-quality and large-scale training data. However, it is important to note that the collection and annotation of a specific medical dataset can be a time-consuming and expensive process, requiring careful attention and concentration [7], while given the sensitive nature of this field, precision and accuracy are paramount in the diagnostic process, which often requires a significant time commitment and concentration [7].

One approach to address the limited availability of medical datasets and the cumbersome task of collecting and annotating medical images is through data augmentation techniques. Data augmentation involves generating additional images, either by augmenting existing ones or creating new ones from scratch, utilizing various methods [8]. This process aims to improve the diversity and quantity of the available annotated data, thereby enhancing the robustness and generalization capabilities of machine learning models in medical image analysis tasks.

In the context of data augmentation, a range of methods and tools have emerged to underline this process, encompassing both traditional approaches and deep learning-based techniques. Focusing on deep generative models, these latter have gained attention due to their ability to generate high-quality samples. Noteworthy among these models are diffusion models, which have demonstrated state-of-the-art performance in image generation tasks [9].

Based on an extensive literature review that we conducted on the diverse applications of diffusion models in the medical domain, it was observed that most of the previous research focused predominantly on utilizing diffusion models to process medical images. Brain datasets were the most commonly used images in the studies reviewed, with six papers using them for a variety of tasks [10–15]. Four papers used chest images [12, 16–18]as a consequence of the study on the paper [12] used two diffrent types of datasets, while one paper used retinal images [19]. and mentioning the data augmentation task,, two papers applied data augmentation to dermatology images [20, 21].

However, in the context of data augmentation, only a few investigations were conducted on dermatology datasets.

In this work, we aim to assess the employment of diffusion models as a data augmentation technique for the diabetic retinopathy dataset. This choice aims to facilitate a new study with unique perspectives and research contributions.

To do so, the augmented dataset is utilized to train deep-learning models for disease diagnosis. The diagnostic outcomes that are obtained from these models are compared to those achieved using the widely recognized data augmentation techniques, serving as a validation of the effectiveness of the proposed approach.

Our main contributions are summarized as follows:

- We conducted an extensive and comprehensive literature review on the use of diffusion models in the medical field. We have to mention that, although we used the categorization given in [21], our investigation provides more details about the state of the artworks.
- We make use of diffusion models as a data augmentation technique. We assess the quality of the obtained images by evaluating two classifiers on the task of diabetic retinopathy diagnosis.

Our thesis is organized as follows:

Chapter 1 provides a brief overview of diverse data types and associated data collection devices, emphasizing the critical role of data in the advancement of ML models. The chapter also discusses the preprocessing and preparation of medical datasets and presents notable ML tasks specific to the medical domain. Additionally, it explores the challenges encountered when applying ML in the medical field.

Chapter 2 presents a brief overview of data augmentation surveys and aligns the methods utilized in this thesis.

Chapter 3 discusses variants of diffusion models that have been refined from diffusion probabilistic models. Moreover, we present a comprehensive study of the state-of-the-art in the application of diffusion models to medicine field. We review the various tasks that diffusion models have been used for in medicine and mention the various types of medical images that diffusion models have been used with (Brain images, chest images, etc).

Chapter 4 provides a detailed description of the methodology that is used for augmenting data set images as well as the architecture of the classifiers used for retinopathy diagnosis.

Chapter 5 concludes this thesis by presenting and discussing the results of the disease grading using the methodology that was described in the previous chapter.

CHAPTER 1

Data in Machine Learning field

1 INTRODUCTION

Machine learning refers to the ability of systems to acquire knowledge from specific training data, enabling the automated construction of analytical models and the resolution of associated tasks [4]. In recent years, the field of machine learning has experienced a transformative phase, characterized by a growing demand for precise and effective models that possess cognitive capabilities similar to humans [4], and even surpass human performance, exhibiting superhuman capabilities [5]. In order to attain such achievements, models are iteratively trained and their parameters are adjusted until optimal outcomes are attained. However, the fundamental determinant in developing an optimal model lies in the quality and quantity of data employed for training purposes [22]. Therefore, it is of paramount importance to meticulously collect, filter, preprocess, and visualize the data prior to training any model. The subsequent sections will provide some diverse data types and the associated data collection devices, highlighting the crucial role of data in the progress of machine learning models. Additionally, we will address the preprocessing and the preparation of medical datasets and showcase notable machine learning tasks specific to the medical domain. Furthermore, we will discuss the challenges encountered by machine learning in the medical field.

2 DATA COLLECTION

Data, within the realm of machine learning, represents the foundational substrate upon which significant patterns are automatically discerned. Over the preceding decades, machine learning has emerged as an omnipresent tool deployed across a diverse spectrum of tasks demanding the extraction of information from extensive datasets [23]. The availability of data assumes paramount importance in the development of machine learning models. With the proliferation of various online platforms, such as Kaggle [1], IEEE [2], UCI Machine Learning Repository [3], Data.gov [4], GitHub [5], and TensorFlow [6], data can be readily accessed and downloaded. While the acquisition of private data may necessitate specific permissions, it does not impede the accessibility of other data sources. In certain scenarios, developers must procure data themselves, particularly when it pertains to private datasets or those that are not publicly accessible.

[1] https://www.kaggle.com

[2] https://www.ieee.org

[3] https://archive.ics.uci.edu/ml/index.php

[4] https://data.gov

[5] https://github.com

[6] https://www.tensorflow.org

Textual data can be procured from social media platforms (e.g., Facebook [7], Twitter [8]) [24], legal documents such as scientific papers [25], web pages, and online reviews [26].

Audio data can be obtained from diverse resources such as podcasts [27], music recordings [28], and voice recordings [29].

For video data, various sources can be utilized, including smartphone video recordings [30], home videos, surveillance footage [31], news footage [32], or motion capture data [33].

In the case of image data, a plethora of devices can be employed, encompassing digital cameras [34], smartphones [35], and satellites [36]. Medical imaging which can be considered as a subset of the image dataset warrants the use of distinct devices for image acquisition, such as the Methylammonium lead iodide (MAPbI3) perovskite-based semiconductor detector for X-ray and CT scans [37], or magnetic resonance imaging (MRI) scanner systems for MRI scans [38].

The use of data necessitates meticulous planning, entailing the delineation of objectives and requirements, meticulous selection of appropriate data sources, and meticulous assurance of data quality and consistency, particularly for sensitive datasets such as medical datasets. Furthermore, ethical and legal considerations, such as data privacy and intellectual property rights, assume critical significance [39].

3 MEDICAL DATASET

Medical imaging encompasses a range of procedures that facilitate the visual representation of anatomical and physiological information pertaining to the human body. Its primary objective is to support radiologists and clinicians in enhancing the efficiency of diagnostic and therapeutic procedures. As an integral component of disease diagnosis and management, medical imaging comprises diverse imaging modalities, each offering unique insights into the underlying pathologies and conditions [40].

To date, both academia and industry have predominantly relied on limited, publicly available datasets and data obtained through commercial products [39]. However, due to privacy and ethical constraints, institutions that possess medical data face challenges in making it publicly accessible [41]. Furthermore, researchers who specialize in applying deep learning (DL) methods to medical image analysis often lack a medical background, typically computer scientists. As a result, they encounter obstacles in independently collecting data due to restricted access to medical equipment and patients, as well as in annotating the acquired data due to a lack of relevant medical

[7] https://www.facebook.com

[8] https://www.twitter.com

knowledge [41].

Therefore, in the context of utilizing medical images for the development of machine learning algorithms, a well-defined sequence of steps becomes crucial. Firstly, it is imperative to obtain local ethical committee approval before employing medical data for algorithm development [42]. This approval signifies adherence to ethical guidelines and ensures the preservation of patient privacy. Subsequently, an institutional review board evaluates the associated risks and benefits of the study to safeguard patient safety and well-being [42]. Following ethical approval, the de-identification of collected data becomes essential to maintain patient confidentiality and privacy [43]. Additionally, the implementation of secure storage measures is necessary to prevent unauthorized access or breaches.

Another critical aspect involves establishing the connection between medical images and their corresponding ground-truth information. Ground-truth information may encompass one or more labels, segmentation masks, or electronic phenotypes such as biopsy or laboratory results [42]. This linkage enables accurate training and evaluation of machine learning models, facilitating meaningful analysis and interpretation.

In conclusion, it is crucial to adhere to a systematic process when utilizing medical images for the development of machine learning algorithms. This process involves obtaining ethical approval, de-identifying the data, implementing secure storage measures, and establishing connections between images and ground-truth information to enable effective analysis and interpretation [42].

4 Machine Learning tasks in the medical field

During the period spanning from the 1970s to the 1990s, the field of medical image analysis primarily relied on a sequential approach that involved the application of low-level pixel processing techniques (such as edge and line detector filters, region growing) and mathematical modeling (including the fitting of lines, circles, and ellipses). These methods were utilized to construct rule-based systems capable of addressing specific tasks within medical image analysis [6]. However, in recent years, there has been a substantial increase in the number of publications utilizing computer vision techniques for the analysis of static medical imagery, with the volume of such publications rising from hundreds to thousands [6]. Machine learning algorithms have been developed to optimize workflow and provide support to medical specialists, including doctors, surgeons, analysts, and radiologists. Furthermore, these algorithms have the potential to alleviate challenges in patient care, particularly in scenarios where access to medical experts is limited or unavailable at certain medical centers. Notably, the efficacy of machine learning and deep learning algorithms in various medical tasks has surpassed the performance of medical specialists in certain cases.

Significant progress has been made in the medical domain, leading to improvements in workflow and the resolution of diverse challenges. For example, Computer-Aided Detection (CAD) algorithms have been employed for the purpose of triaging screening mammograms by detecting and annotating suspicious findings, thereby enhancing the sensitivity of radiologists [44]. Deep learning neural networks have facilitated the reduction or elimination of gadolinium-based contrast media usage in MRI scans [45]. Additionally, noise reduction techniques have been applied to CT images to decrease radiation dosage during CT imaging [46].

Machine learning techniques have also demonstrated successful outcomes in automatic lesion detection across various imaging modalities. Examples include the identification of pulmonary malignant neoplasms, active tuberculosis, pneumonia, and pneumothorax in thoracic diseases [47], the identification of intracranial hemorrhages [48], Retinal image analysis, encompassing blood vessel segmentation [49] and diabetic retinopathy detection [50], the prediction of Alzheimer's disease diagnosis [51] and urinary stone detection [52]. Moreover, automatic quantification of medical images has enabled the assessment of skeletal maturity based on pediatric hand radiographs [53], coronary calcium scoring using CT images [54], prostate cancer classification using MRI images [55], ventricle segmentation from cardiac MRI scans [56] and innovation in surgery simulation, allowing surgeons to predict and explore potential solutions before performing operations, thus saving time and costs [57].

The scarcity of image data for training and evaluating artificial intelligence (AI) algorithms poses a substantial constraint and challenge within the field of medical image analysis. Scholars and professionals emphasize the criticality of obtaining ample and diverse datasets to enhance the performance and generalizability of AI algorithms in the context of analyzing medical images. Access to such datasets is essential for advancing the capabilities of AI algorithms and enabling more accurate and reliable analysis of medical images.

5 Data unavailability in the medical field

Data availability constitutes a fundamental consideration in the realm of machine learning (ML) applied to healthcare. Sufficient and diverse data are indispensable for developing high-performance ML models and evaluating their generalizability with confidence. However, data availability encounters various constraints that hinder these objectives. Such limitations may arise from factors like incomplete digitization, as observed in pathology where a majority of slides remain unscanned, inaccessibility due to patient privacy concerns or commercial restrictions [58], or insufficiency concerning diseases affecting a small patient population, thereby impeding both diagnosis and treatment [59].

According to the European Union (EU) definition, diseases are categorized as rare when their prevalence is below 5 individuals per 10,000 [60]. This scarcity of data engender adverse effects, including delayed disease diagnosis, hindered treatment efficacy, and potential fatality, exemplified by Fibrodysplasia ossificans progressiva (FOP), a rare genetic disorder that causes systemic and progressive ossification in various fibrous tissues, which was examined in a study involving a cohort of four patients. However, due to certain limitations, the images of one patient were excluded from the study. Among the remaining three patients, two underwent three separate CT scans each, while the third patient underwent four CT scans [61].

Moreover, the acquisition of well-annotated medical data is a resource-intensive and time-consuming process [62], necessitating the involvement of specialists. For instance, in the study [7], three datasets comprising 11,852 image samples from 872 patients were collected from three medical centers and were utilized to evaluate the method's applicability in real-world scenarios. The annotations of these datasets were meticulously curated by three experienced radiologists with over a decade of expertise in interpreting breast MR images and which coasts a significant amount of time, effort, and financial resources to be realized. However, the labor-intensive and resource-demanding nature of the annotation process poses limitations on the availability of adequately annotated datasets.

Overall, data availability poses significant challenges when applying ML to healthcare, as relying on limited-sized data for training machine learning algorithms proves inadequate. This insufficiency is demonstrated by the study on Soft Tissue Sarcoma conducted on the Cancer Imaging Archive dataset of anatomical MR imaging data from 51 patients [63], indicating that ML algorithms typically necessitate substantial volumes of data with balanced class distributions to achieve optimal performance.

To address this predicament, a robust and comprehensive approach is required to handle the unavailability of data. This approach aims to augment patient care for individuals afflicted with rare diseases and expedite diagnosis in cases burdened by extensive waiting lists, particularly in densely populated urban areas. By adopting such an approach, more efficient referrals can be facilitated, leading to enhanced healthcare delivery in terms of quality and timeliness [59].

Furthermore, the unavailability of data is a significant factor that impacts the development of the machine learning domain in the medical field. This limitation poses a challenge to the effective implementation and advancement of machine learning models, hindering their potential in providing additional support and augmenting the expertise of medical practitioners.

6 CONCLUSION

In conclusion, recent years have witnessed significant advancements in machine learning, resulting in the development of models that exhibit cognitive capabilities comparable to humans and even surpass human performance in certain tasks.

The quality and quantity of data utilized for training machine learning models play a pivotal role in achieving optimal outcomes. Data collection encompasses accessing diverse sources and employing specific devices for various data types, including textual, audio, video, and image data.

In the realm of medical imaging, ethical considerations, and data privacy assume paramount importance, particularly when dealing with medical datasets. Machine learning techniques have greatly benefited the field of medical image analysis, facilitating improved workflow and addressing challenges in patient care. However, the scarcity of labeled data remains a significant constraint, impeding the progress of AI algorithms in medical image analysis.

Data availability presents challenges in healthcare applications of machine learning, including factors such as incomplete digitization, limited accessibility, and diseases affecting small patient populations. The scarcity of data negatively impacts disease diagnosis, treatment efficacy, and patient outcomes. Furthermore, the acquisition of well-annotated medical data is a resource-intensive and time-consuming process, further limiting the availability of adequately annotated datasets.

The field of machine learning in healthcare holds immense potential, yet it faces challenges related to data availability. Therefore, in the subsequent chapter, we will delve into one of the prominent strategies employed to tackle the challenge of data unavailability.

CHAPTER 2

Data Augmentation tools

1 Introduction

As stated in Chapter 1, the adequacy or bias in the training data can significantly impact the generalization performance, irrespective of advancements in model design and training methodologies. Conversely, employing comprehensive, diverse, and representative datasets consistently results in satisfactory performance, even with less complex algorithms [64]. Data augmentation is a commonly employed approach to tackle the scarcity of training datasets, wherein new samples are generated either from existing data or from scratch [8]. Furthermore, this chapter aims to present a brief overview of data augmentation surveys. To align with the methods utilized in this thesis, we have adopted a taxonomy from relevant papers [8, 21, 65–68].

2 Data Augmentation

Data augmentation involves generating additional training samples from existing data or creating them from scratch using various methods [8]. Expanding the training data can be achieved through two main approaches: the manipulation of existing training data, which is called **the traditional (classical) image data augmentation**, and the generation of new data samples, called **the deep-based learning augmentation** [65]. Figure 2.1 portrays the employed taxonomy for various data augmentation tools including the utilized ones.

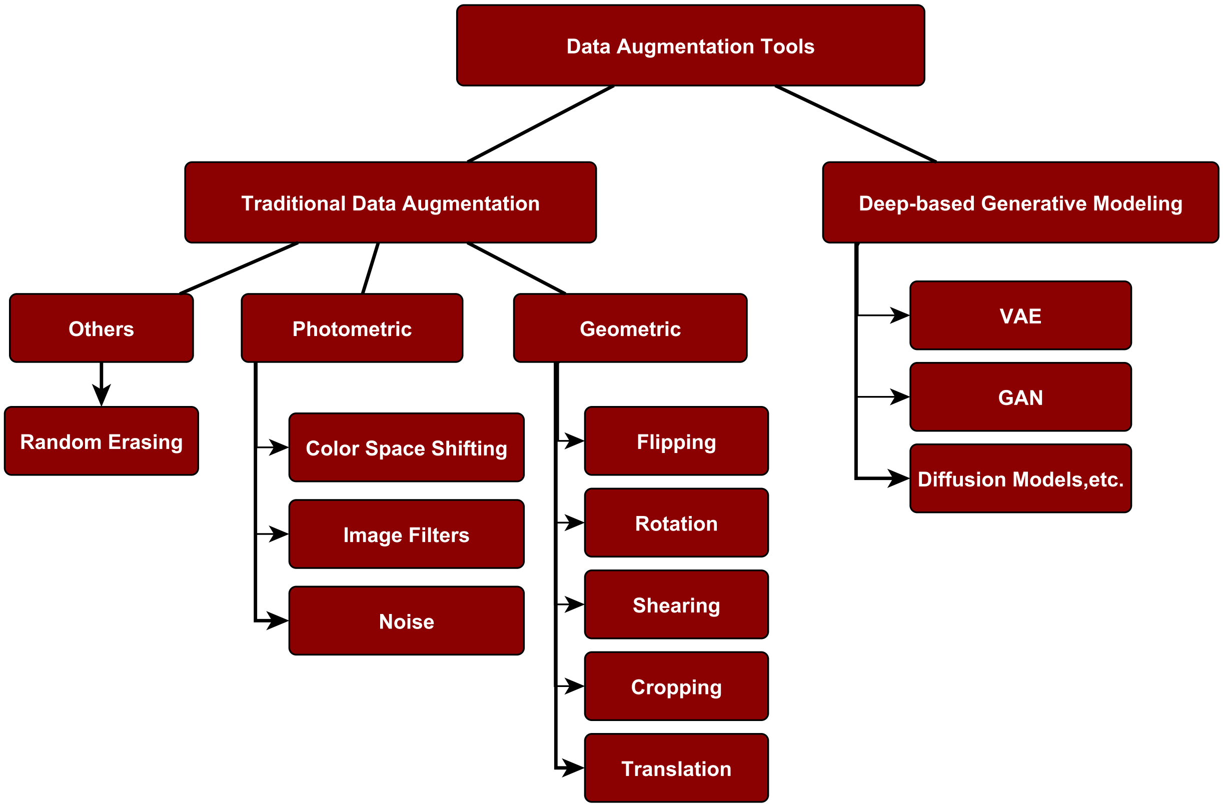

Figure 2.1: Taxonomy of data augmentation tools

2.1 TRADITIONAL DATA AUGMENTATION

The traditional image transformation methods, comprising both photometric and geometric techniques and other methods, such as random erasing [8,65]. These techniques are employed to apply various transformations to the existing training data, thereby enhancing its diversity and expanding the size of the training set. By introducing such variations, the training data is effectively augmented to simulate real-world scenarios, leading to improved generalization capabilities of machine learning models [69–71].

GEOMETRIC TRANSFORMATION

Geometric transformations encompass a set of image data augmentation techniques utilized to alter the geometric properties of images [8]. This category of transformations comprises operations such as flipping, rotation, shearing, cropping, and translation. Unlike other augmentation methods, geometric transformations do not modify the pixel values but instead focus on repositioning the pixels within the image [8]. The primary objective of these transformations is to introduce variations in the training data that accurately reflect real-world changes in appearance, including variations in viewpoint, non-rigid deformations, perspective adjustments, and changes in scale [8].

Image flipping is a geometric transformation that entails reflecting an image across its vertical axis, horizontal axis, or both axes simultaneously [65]. By employing flipping techniques, users can augment the dataset without the need for artificial processing [65]. This augmentation technique encompasses various methods, including vertical flipping, horizontal flipping, and the combined approach involving both vertical and horizontal flipping. Vertical flipping entails rotating the image upside down, with the y-axis positioned at the top and the x-axis at the bottom. The transformation is described by the f_x and f_y values, which represent the current coordinates of each pixel after flipping along the vertical axis (Eq 2.1).

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.1}$$

On the other hand, horizontal flipping requires the image to be horizontally rotated, resulting in its left and right sides being reversed. The f_x and f_y components determine the pixel's new location after reflection along the horizontal y-axis (Eq 2.2).

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.2}$$

Furthermore, vertical and horizontal flipping combines both of these transformations, resulting in a horizontally and vertically rotated image where both the horizontal and vertical columns are preserved. The f_x and f_y coordinates represent the current coordinates of each pixel after reflection along the vertical and horizontal axes, respectively (Eq 2.3).

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.3}$$

Rotation is a classical geometric image data augmentation technique whereby the image is rotated around an axis, either clockwise or counterclockwise, by angles ranging from 1 to 359 degrees [65]. The rotation process can be applied to images incrementally by a specified angle degree. For instance, rotating an image at approximately 30-degree intervals would result in a set of 11 images with rotation angles of 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, and 330 degrees. The rotation equation, as depicted in Eq. 2.4, describes the transformation of the pixel's new position (f_x, f_y) after the rotation process, based on the initial coordinates (x, y) of the raw image.

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.4}$$

Shearing is a geometric transformation that alters the shape of an object within an image by modifying its dimensions along both the x and y directions [65]. This technique is particularly useful for distorting the original shape of an object. Shearing can be classified into two types: shearing along the x-axis and shearing along the y-axis. Equation 2.5 defines the shearing transformation along the x-axis, while Equation 2.6 describes the shearing transformation along the y-axis. The f_x and f_y variables represent the new position of each pixel after the shearing operation, whereas x and y denote the coordinates of the corresponding pixel in the original image.

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} 1 & \text{shX} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.5}$$

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \text{shY} & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.6}$$

Cropping is a classical geometric image data augmentation technique used to magnify the original image. Cropping involves two distinct methods [65] (sometimes referred to as "zooming" or "scaling" in scientific research). The first method involves selecting a region within the image, starting from a specific X, Y location and ending at another X, Y location. For instance, if the image size is 200x200 pixels, a cropping operation may involve cutting the image from the location (0, 0) to (150, 150), or from (50, 50) to (200, 200). The second method entails scaling the cropped image back to its original size. Following the previous example, the cropped image would be resized to 200x200 pixels. Equation 2.7 depicts the scaling equation, where f_x and f_y represent the new coordinates of each pixel after the scaling operation, and x and y represent the coordinates of the original location within the image.

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} \text{Xscale} & 0 \\ 0 & \text{Yscale} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.7}$$

The translation is a fundamental operation that involves displacing an object within an image from one position to another. In the geometric image data augmentation, translation is typically performed while considering the preservation of image data [65]. This can be achieved by leaving a portion of the image white or black after the translation, preserving the original image data, or introducing randomness or Gaussian noise. Translation can be performed in the X direction, Y direction, or both simultaneously (X and Y direction). Translating images in different directions, such as left, right, up, or down, can be particularly useful for mitigating positional bias in the data. Equation 2.8 presents the translation equation, where fx and fy denote the new coordinates of each pixel after the translation operation, and x and y represent the coordinates of the original location within the image.

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} \mathrm{x} \\ \mathrm{y} \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix} \tag{2.8}$$

Photometric Transformation

Photometric transformations represent a class of image data augmentation methodologies employed to manipulate the pixel values of images while retaining their fundamental structure and geometry [70, 72]. Their principal objective is to modify the visual appearance and attributes of images, encompassing elements such as brightness, contrast, color, and texture [70, 72]. In the realm of computer vision tasks, photometric transformations assume a critical role in the augmentation of training data. By introducing variations in the visual appearance of images, these transformations strive to enhance the generalization capacity and robustness of machine learning models. Their purpose extends to simulating real-world scenarios, accommodating diverse lighting conditions, and bolstering the diversity exhibited within the training dataset [70, 72]. The photometric transformation encompasses various methods, including color space-shifting, image filters, and noise [65].

Color space-shifting is a classical photometric data augmentation technique within the realm of color manipulation [65]. A color space is a mathematical construct used to represent and manipulate colors based on their properties such as brightness and hue. In human perception, colors are distinguished based on the quantities of red, green, and blue light emitted by the phosphor panel. In classical photometric data augmentation, color space shifting is considered an important technique for augmenting the number of images and revealing hidden features that may be obscured under a specific color space. Several well-known color spaces are commonly used, including CMY(K) (Cyan-Magenta-Yellow-Black), YIQ, YUV, YCbCr, YCC (Luminance/Chrominance), HSL (Hue-Saturation-Lightness), and RGB (Red-Green-Blue). Color space shifting allows for random and intelligent modifications. Adjusting pixel values by a constant value can enhance the visibility of bright or dark images. Furthermore, color space manipulation enables independent processing of individual RGB color channels. Another approach involves constraining pixel values to a minimum or maximum range. These techniques contribute to enhancing the color appearance of optical photographs without the need for sophisticated tools.

Image Filters are numerous widely used image processing techniques, including histogram equalization, brightness adjustments, sharpening, blurring, and filters, that have gained significant popularity [65]. These techniques and filters employ the application of an n×m matrix across the entire image. Histogram equalization is a method used to modify image intensities in order to enhance contrast. On the other hand, white balancing aims to alter the image such that it appears illuminated by a neutral light source [65]. Special operations are often conducted separately in different spectral domains of the signal. Sharpening filters are spatial filters utilized to emphasize fine details or enhance blurred features within an image [65]. Conversely, blurring involves an averaging process that integrates pixel values with those of neighboring pixels. Combining the sharpening and blurring filters may result in a distorted image or the accentuation of high-contrast horizontal or vertical edges, which can aid in the identification of image details [65]. These aforementioned filters are applied through matrix multiplication between the original image and the corresponding filter matrix.

Noise is a fundamental element in image augmentation techniques, often employed to enhance the realism and robustness of image processing algorithms [65]. Various types of noise can be introduced to images, each serving a unique purpose. Gaussian noise, for instance, is commonly used as it introduces color value variations by applying a noise matrix derived from a standard distribution. Poisson noise, on the other hand, is inherent in electromagnetic frequencies encountered in applications like X-ray and gamma-ray machines that emit photons continuously. Another type, salt, and pepper noise, involves modifying pixel values in specific regions of an image. Lastly, speckle noise, found in optical devices such as lasers, radar systems, and sonar, can be both multiplicative and additive. These diverse forms of noise play a crucial role in simulating real-world conditions and enhancing the overall resilience of image processing algorithms.

RANDOM ERASING

The random erasing technique, an image data augmentation method, is distinct from geometric transformations commonly employed in image processing [65]. Random erasing operates on the fundamental principle of randomly selecting a square region within an image and removing its contents. Empirical evidence supports the efficacy of this technique, illustrating its positive influence on image augmentation and subsequent performance enhancements across diverse tasks. By selectively erasing regions within an image, random erasing introduces perturbations that encourage robustness and generalization in deep learning models, ultimately improving their ability to handle variations and challenges present in real-world scenarios.

2.2 Deep-based Generative Modeling

Generative modeling techniques are essential in the generation of synthetic data samples that closely resemble real data instances, capturing their statistical properties and characteristics [8]. These models achieve this by learning the underlying distribution of the training data, enabling the generation of novel samples that exhibit similar statistical properties to the original data [8]. By comprehending the intricate patterns and dependencies within the training dataset, generative models excel at generating synthetic data instances that effectively mimic the complexities observed in real-world data [8].

The application of generative modeling for data augmentation offers several advantages in the context of machine learning tasks. It facilitates the expansion of the training dataset by creating additional samples, thereby increasing its size and diversity [8]. This augmentation technique effectively addresses the challenges associated with limited training data and has the potential to enhance the generalization performance of machine learning models.

By harnessing the power of generative models, researchers can generate new samples with desired variations, such as different poses, lighting conditions, or object appearances. This capability allows for the creation of synthetic data that can supplement the original dataset and enhance the model's capacity to handle diverse real-world scenarios. In this chapter, we provide a comprehensive overview of three fundamental generative models: Vanilla Variational Autoencoders (VAEs) [73], Generative Adversarial Networks (GANs) [74], Diffusion Probabilistic models(DMs) [75].

Vanilla Variational Autoencoder (VAE)

A vanilla variational autoencoder (VAE) [73] is a probabilistic generative model [76] that is built upon the concept of an "autoencoder" in deep learning [77]. It comprises two fundamental components, namely the encoder and the decoder [76]. During training, the VAE aims to minimize the reconstruction error between the input and the decoded/reconstructed data, following the standard autoencoder reconstruction objective. Additionally, it incorporates a variational objective term to encourage the learning of a latent space distribution resembling a standard normal distribution [77].

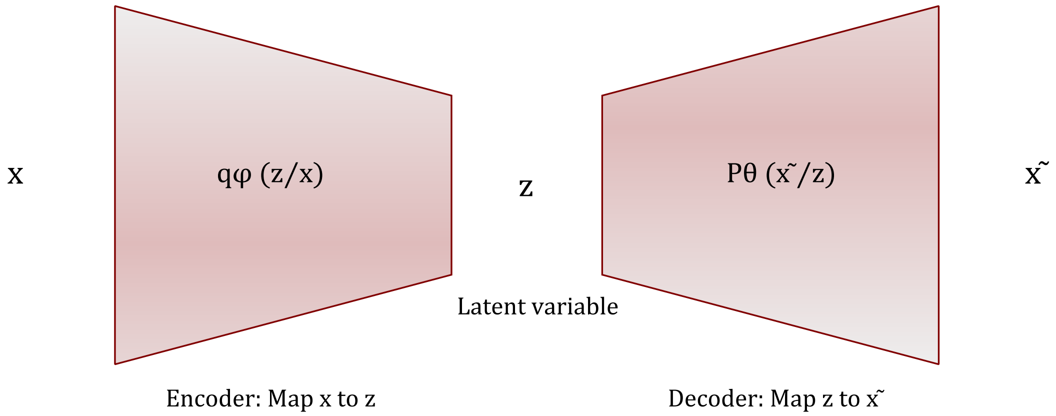

Figure 2.2: Variation AuotoEncoder architecture

The basic idea of the VAE for image generation can be summarized as the following:
The probabilistic encoder component of the VAE probabilistically encodes the input data samples, denoted as x, from a specific dataset X, into a latent variable z from a conditional distribution $q_{\varphi}(z \mid x)$. This encoding process involves the generation of a distribution of representations through a probabilistic encoder, which is implemented using a neural network and its associated parameters [78].
The generator component, known as **the probabilistic decoder**, within the VAE, generates data $\tilde{x}$ through a random process that involves the latent variable z. This process begins by sampling a value z from a normal distribution. Subsequently, a value $\tilde{x}$ is generated from a conditional distribution $p_{\theta}(\tilde{x} \mid z)$, which is parameterized by θ [76]. The conditional distribution $p_{\theta}(\tilde{x} \mid z)$ determines the relationship between the generated data $\tilde{x}$ and the latent variable z.

$$z \sim \mathcal{N}(0, I)$$
$$x \sim p_{\theta}(\tilde{x} \mid z)$$

The training: Throughout the training process, the VAE aims to maximize the marginal likelihood of the reconstructed data $\tilde{x}$. However, due to the integral of the marginal likelihood $p_{\theta}(\tilde{x}) = \int p_{\theta}(z)p_{\theta}(\tilde{x} \mid z)\, dz$ is intractable, the involvement of posterior inference $p_{\theta}(z|\tilde{x}) = \frac{p_{\theta}(\tilde{x}|z)p_{\theta}(z)}{p_{\theta}(\tilde{x})}$ becomes intractable, researchers employ backpropagation and stochastic gradient descent [73] to optimize the variational lower bound of log likelihood [73].
By utilizing these techniques, the VAE seeks to approximate and improve the likelihood estimation, thereby enhancing the overall training process.

$$\log p_{\theta}(\tilde{x}) \geq L_{\text{vae}} = E_{q_{\phi}(z|x)}[\log p_{\theta}(\tilde{x} \mid z)] - D_{\text{kL}}(q_{\phi}(z \mid x) \| p(z))) \tag{2.9}$$

Where in 2.9, $p_(z)$ is the prior distribution (Unit Gaussian), the posterior distribution of latent variable z given data x is approximated by the variational posterior, $q_{\varphi}(z \mid x)$, which is parameterized by an encoder network, and $p_{\theta}(\tilde{x} \mid z)$ is a probabilistic decoder parameterized by a neural network to generate data $\tilde{x}$ given the latent variable z (a reconstruction term) [73].
To optimize the VAE model, the lower bound of the marginal likelihood L_{vae} is maximized using techniques such as the reparameterization trick [73] and the Stochastic Gradient Variational Bayes (SGVB) estimator [73]. By applying these methods, the VAE model can effectively learn the process of generating data based on a random latent variable z sampled from a normal distribution.

GENERATIVE ADVERSARIAL NETWORKS (GAN)

Generative Adversarial Networks (GANs) [74] have emerged as prominent models for effectively capturing the complexity of real-world data. Comprising a generator (G) and a discriminator (D), GANs employ adversarial optimization to achieve their objectives [79].

In GANs, the generator and discriminator engage in a competitive interplay, each enhancing its own capabilities. The generator's role is to learn and capture the underlying distribution of authentic samples, while the discriminator, often implemented as a binary classifier, evaluates the likelihood of a given sample originating from the real dataset. By functioning as a critic, the discriminator strives to refine its ability to discern synthetic samples from genuine ones [80].

Conceptually, the generative model (G) can be likened to a team of counterfeiters producing and utilizing fake currency without detection, while the discriminative model (D) represents the police force aiming to identify counterfeit currency [74]. This competitive game pushes both teams to continuously refine their methods until the counterfeits become indistinguishable from genuine articles [74]. The following Figure 2.3 illustrates the generator and the discriminator of GANs.

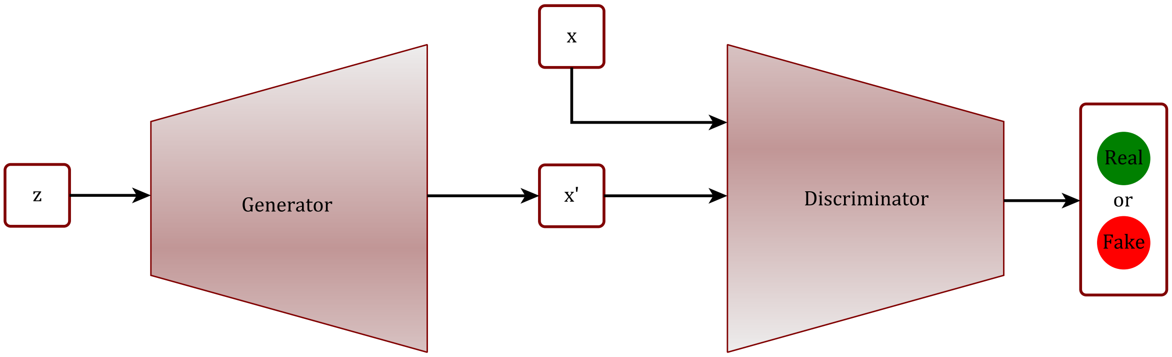

Figure 2.3: General GANs Architecture

To estimate the underlying distribution of real data instances, a prior distribution $p_z(z)$ is established for the input noise variables. The generator component, denoted as $G(z; \theta_g)$, employs a differentiable multilayer perceptron function parameterized by θ_g to map the noise variables z to the data space. Simultaneously, the discriminator component, denoted as $D(x; \theta_d)$, is introduced as a multilayer perceptron that outputs a single scalar.

The optimization process involves a minimax game aimed at minimizing the classification loss function, given by:

$$\min_G \max_D V(D,G) = \underbrace{\mathbb{E}x \sim p_{\text{data}(x)}[\log D(x)]}_{I} + \underbrace{\mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]}_{II} \tag{2.10}$$

In Equation 2.10, $D(x)$ represents the probability that sample x is drawn from the original data distribution rather than the generator's distribution. The classification loss consists of two terms: the first term (I), influenced by real images, is maximized when the discriminator correctly classifies them as real; the second term (II), influenced by generated images, is maximized when the discriminator accurately classifies them as fake.

During training, the generator aims to minimize the II term to generate realistic images that can deceive the discriminator, while the discriminator is trained to maximize the I term to improve its classification accuracy. The generator and discriminator are trained iteratively in an adversarial manner. When training the generator, the discriminator is kept fixed, and the classification loss is minimized. Conversely, when training the discriminator, the generator is fixed, and the classification loss is minimized from the discriminator's perspective.

During the inference phase, the discriminator is no longer involved. Instead, random noise samples z are drawn from a normal distribution $(z \sim \mathcal{N}(0,1))$, and the generator maps these samples to the image space $(x' = G(z))$. This process enables the generation of both real and synthetic images in the image space, which can then be compared using the discriminator through classification.

In summary, Generative Adversarial Networks (GANs) are designed to learn a probability density model $P(x)$ that can be sampled. The training procedure involves random sampling from a normal distribution, with the generator learning a mapping function to generate images in the image space. The discriminator is employed to compare the real and generated images by classifying them.

DIFFUSION PROBABILISTIC MODELS

The main idea of Diffusion Probabilistic models (DPMs) [75] is to develop an approach that simultaneously achieves both flexibility and tractability. The essential idea, inspired by **non-equilibrium statistical physics**, is to systematically and slowly destroy the structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data [75].

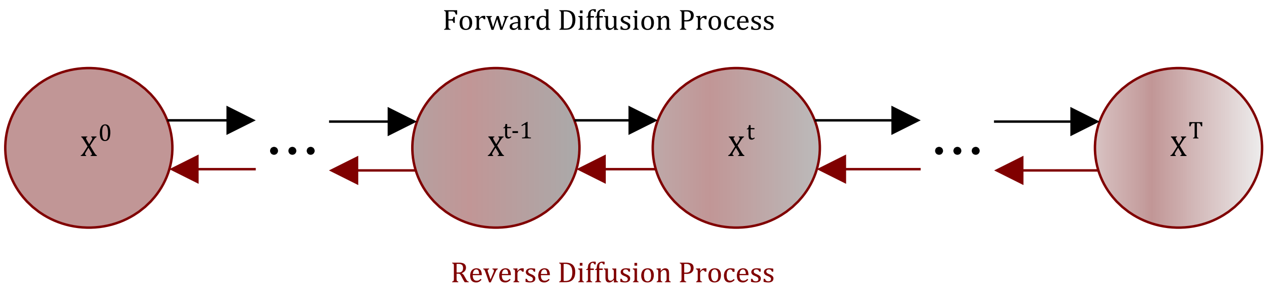

Figure 2.4: Diffusion models architecture

The method proposed in this paper [75] leverages concepts from non-equilibrium statistical physics [81] and sequential Markov Chain Monte Carlo [82]. It introduces a novel approach using a Markov chain to gradually transform one distribution into another.

The central concept introduced by the authors of [75] involves a systematic and gradual destruction of structure within a given data distribution through an iterative forward diffusion process. Subsequently, a reverse diffusion process is learned to restore the structure in the data, resulting in a highly adaptable and computationally manageable generative model.

In this approach, the authors construct a generative Markov chain that converts a known simple distribution (e.g., Gaussian) into the target data distribution through a diffusion process. Instead of using this Markov chain to approximate the evaluation of a pre-defined model, they explicitly define the probabilistic model as the final state of the Markov chain. Since each step in the diffusion chain has a computationally tractable probability, the entire chain can be analytically evaluated.

THE ALGORITHM BEHIND

The objective is to establish a forward diffusion process, referred to as the inference process, that transforms intricate data distributions into simpler and computationally manageable distributions.

Subsequently, a finite-time reversal of this diffusion process is learned, defining the generative model distribution. The authors commence by presenting the details of the forward diffusion process, followed by elucidating the training and utilization of the reverse diffusion process for probability evaluation.

Forward Process: The initial distribution, denoted as $q(x^{(0)})$, represents the data distribution. Through a series of iterative steps using the Markov diffusion kernel $T_\pi(y \mid y'; \beta)$, the data distribution $q(x^{(0)})$ is gradually transformed into a well-behaved distribution $\pi(y)$, which possesses desirable properties and can be analytically evaluated. Here, β represents the diffusion rate, governing the pace of the transformation process.

$$\pi(y) = \int dy'\, T_\pi(y \mid y'; \beta)\pi(y') \tag{2.11}$$

$$q(x^{(t)}|x^{(t-1)}) = T_\pi(x^{(t)} \mid x^{(t-1)}; \beta_t) \tag{2.12}$$

Referring to Equation 2.12, the conditional distribution $q(x^t \mid x^{t-1})$ can be represented as follows: for the Gaussian distribution, $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t; \sqrt{1-\beta_t}\, x^{t-1}, \beta_t I)$, and for the Binomial distribution, $q(x^t \mid x^{t-1}) = \mathcal{B}\left(x^t; (1-\beta_t)\, x^{t-1} + 0.5\beta_t\right)$.
In the case of Gaussian diffusion, the selection of β within the forward trajectory plays a crucial role in determining the efficacy of the trained model. The forward diffusion schedule, denoted as $\beta_{2\ldots T}$, is acquired through gradient ascent on K, ensuring optimized performance. To mitigate the risk of overfitting, a small constant value is assigned to the initial step's variance, denoted as β_1. In the discrete Specifically, the diffusion rate β_t is set to $(T-t+1)^{-1}$, ensuring that a diminishing proportion of the original signal is preserved as the diffusion progresses.

The forward trajectory, which entails initiating from the data distribution and executing T steps of diffusion, can be expressed as follows:

$$q(x^{(0\ldots T)}) = q(x^{(0)}) \prod_{t=1}^{T} q(x^{(t)} \mid x^{(t-1)}) \tag{2.13}$$

In the conducted experiments (see Equation 2.13), the diffusion process $q(x^{(t)} \mid x^{(t-1)})$ corresponds to two scenarios: Gaussian diffusion into a Gaussian distribution with an identity covariance, or binomial diffusion into an independent binomial distribution.
It is important to note that the final state of the diffusion process $\pi(x^{(T)})$ follows a Gaussian

distribution $\pi(x^{(T)}) = \mathcal{N}(x^{(T)}, 0, I)$ for the case of Gaussian diffusion, or a binomial distribution $\pi(x^{(T)}) = \mathcal{B}(x^{(T)}, 0.5)$ for the case of binomial diffusion.

Reverse Process: The generative distribution will be trained to model the reverse trajectory of the forward diffusion process, ensuring consistency between the two

$$p(x^{(T)}) = \pi(x^{(T)}) \tag{2.14}$$

$$p(x^{(0\ldots T)}) = p(x^{(T)}) \prod_{t=1}^{T} p(x^{(t-1)} \mid x^{(t)}) \tag{2.15}$$

$$p(x^{(t-1)} \mid x^{t}) = T(x^{(t-1)}; f_{\mu}(x^{t}, t), f_{\Sigma}(x^{t}, t)) \tag{2.16}$$

Referring to Equation 2.16, the conditional distribution $p(x^{(t-1)} \mid x^{t})$ can be represented as follows: for the Gaussian distribution, $p(x^{(t-1)} \mid x^{t}) = \mathcal{N}(x^{(t-1)}; f_{\mu}(x^{t}, t), f_{\Sigma}(x^{t}, t))$, and for the Binomial distribution, $p(x^{(t-1)} \mid x^{t}) = \mathcal{B}(x^{(t-1)}; f_{b}(x^{t}, t))$.
During the learning process, the estimation of the mean and covariance for a Gaussian diffusion kernel, or the determination of the bit flip probability for a binomial kernel, is required. Specifically, the functions $f_{\mu}(x^{t}, t)$ and $f_{\Sigma}(x^{t}, t)$ define the mean and covariance of the reverse Markov transitions for the Gaussian distribution, while $f_{b}(x^{t}, t)$ represents the function that yields the bit flip probability for the binomial distribution.

Training Objective: The process of training involves maximizing the log-likelihood of the model,

$$L = \int dx^{(0)}\, q(x^{(0)}) \log p(x^{(0)}) \tag{2.17}$$

which has a lower bound provided by Jensen's inequality,

$$L \geq \int dx^{(0\ldots T)} q(x^{(0\ldots T)}). \log p(x^{(T)}) \prod_{t=1}^{T} \frac{p(x^{(t-1)} \mid x^{(t)})}{q(x^{(t)} \mid x^{(t-1)})} \tag{2.18}$$

for the diffusion trajectories, this reduces to,

$$\begin{aligned} &L \geq K \\ &K = -\sum_{t=2}^{T} \int dx^{(0)} dx^{(t)} q(x^{(0)}, x^{(t)}).D_{\mathrm{kL}}(q(x^{(t-1)} \mid x^{(t)}, x^{(0)}) \| p(x^{(t-1)} \mid x^{(t)})) \\ &\quad + C \end{aligned} \tag{2.19}$$

During the training process, the objective is to identify the reverse Markov transitions that optimize the lower bound on the log-likelihood, which will automatically minimize the KL divergence between $(q(x^{(t-1)} \mid x^{(t)}, x^{(0)})$ which is the ground truth (in image in the forward Markov chain) and $p(x^{(t-1)} \mid x^{(t)})$ which is an image generated from the reverse Markov chain.

The training process involves identifying the reverse Markov transitions that maximize the lower bound on the log likelihood

Hence, the estimation of a probability distribution is simplified to the regression task of determining the functions that specify the mean and covariance of a sequence of Gaussian distributions (or the state flip probability for a sequence of Bernoulli trials).

3 Conclusion

In conclusion, this chapter provided a brief overview of data augmentation tools. Data augmentation serves as a widely adopted strategy to address the limited availability of training datasets and enhance the generalization performance of machine learning models. The chapter discussed two primary approaches to data augmentation: traditional (classical) methods and deep-based learning augmentation.

Traditional data augmentation involves manipulating existing training data to introduce variations and expand the size of the dataset. Geometric transformations, such as image flipping, rotation, shearing, cropping, and translation, mimic real-world changes in appearance. Photometric transformations modify pixel values to enhance visual attributes like brightness, contrast, color, and texture. Additionally, the technique of random erasing randomly selects and removes a square region within an image.

Deep-based generative modeling provides an alternative data augmentation approach by generating synthetic data samples that closely resemble real instances. The chapter introduced three generative models: Vanilla Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Basic Diffusion Models (DMs). These models learn the underlying distribution of the training data and generate new samples with desired variations, expanding the dataset and improving the model's adaptability to diverse real-world scenarios.

In summary, data augmentation is a valuable technique for mitigating the scarcity of training data and enhancing the performance of machine learning models. By diversifying and expanding the training dataset, models can achieve better generalization and increased robustness.

In the next chapter, we will present a brief overview of the three categories of diffusion models, which have been refined from diffusion probabilistic models. Additionally, we will examine the manifold applications of Diffusion models in the field of medicine.

CHAPTER 3

Diffusion Models Variants and Their Application in the Medical Field

1 INTRODUCTION

In Chapter 1, it has been observed that the field of Diffusion models has recently gained prominence as a state-of-the-art category of deep generative models [9]. These models have surpassed the long-standing dominance of generative adversarial networks (GANs), particularly in the challenging task of image synthesis [83]. The application of Diffusion models extends across various domains, encompassing computer vision, natural language processing, temporal data modeling, multi-modal modeling, robust machine learning, and interdisciplinary areas such as computational chemistry and medical image reconstruction [83].

This chapter aims to present a brief overview of three notable formulations of diffusion models, along with a succinct overview of relevant research conducted in the medical domain.

2 DIFFUSION MODELS VARIANTS

Three notable formulations of diffusion models have emerged as prominent approaches in the field [80, 83]. These include denoising diffusion probabilistic models (DDPMs) [84], score-based generative models (SGMs) [85], and stochastic differential equations (Score SDEs) [86]. All these formulations share a common fundamental concept with diffusion probabilistic models from [75] (discussed in Chapter 2), they offer distinct perspectives and novel techniques that deviate from traditional diffusion probabilistic models.

2.1 DENOISING DIFFUSION PROBABILISTIC MODELS (DDPMS)

In the denoising diffusion probabilistic model (DDPM) [84], the forward process is meticulously designed to achieve the objective of mapping any given data distribution to a simpler prior distribution, often exemplified by a standard Gaussian distribution. Simultaneously, the reverse process learns transition kernels parameterized by deep neural networks to effectively reverse the operations performed by the forward process.

FORWARD PROCESS

Consider the original data distribution as $q(x_0)$, where x_0 is a data sample drawn from the distribution ($x_0 \sim q(x_0)$). By employing a forward noising process with a transition kernel $q(x_t|x_{t-1})$, a sequence of random variables $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_T$ is generated. The joint distribution of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$

conditioned on x_0, denoted as $q(\mathbf{x}_1, \ldots, \mathbf{x}_T | \mathbf{x}_0)$, can be factorized using the chain rule of probability and the Markov property:

$$q\left(\mathbf{x}_1, \ldots, \mathbf{x}_T \mid \mathbf{x}_0\right)=\prod_{t=1}^T q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) \tag{3.1}$$

In the case of denoising diffusion probabilistic models (DDPMs), the transition kernel $q(x_t|x_{t-1})$ is manually designed to progressively transform the data distribution into a tractable prior distribution $q(x_0)$. A common choice for the transition kernel is Gaussian perturbation, which can be represented as:

$$q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right)$$

Here, t denotes the number of diffusion steps, β represents the variance schedule applied during the diffusion process, I is the identity matrix, and $\mathcal{N}(x; \mu, \sigma)$ refers to the normal distribution with mean $\sqrt{1-\beta_t} x_{t-1}$ and covariance $\beta_t I$.

As previously highlighted in Diffusion probabilistic models(DPMs) [75], the utilization of a Gaussian transition kernel enables the analytical derivation of $q(x_t|x_0)$ for all $t \in 0, 1, \ldots, T$ through the integration of the joint distribution mentioned in Eq. 3.1.

Nevertheless, denoising diffusion probabilistic models (DDPMs) differ from Diffusion Probabilistic modes by obviating the need for an iterative procedure. This is achieved by introducing $\bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_s)$, which allows for the direct sampling of any step of the perturbed latent variable conditioned on the input x_0 using Eq. 3.2.

The values of β_t correspond to a noise schedule designed such that $\bar{\alpha}_t$ converges to zero and $q(\mathbf{x}_T|x_0) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$. The noise schedule controls the diffusion of the data, ensuring that $\bar{\alpha}_t$ approaches zero at the final step, leading $q(\mathbf{x}_T|x_0)$ to approximate a standard normal distribution $\mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$.

$$q\left(x_t \mid x_0\right)=\mathcal{N}\left(x_t ; \sqrt{\bar{\alpha}_t} x_0,\left(1-\bar{\alpha}_t\right) I\right) \tag{3.2}$$

To generate a sample x_t conditioned on x_0 following the normal distribution as expressed in Eq. 3.2, denoising diffusion probabilistic models (DDPMs) make use of the reparametrization trick (Eq.3.3), This technique involves introducing a Gaussian vector ϵ ($\epsilon \sim \mathcal{N}(0, I)$) to enable the transformation of the sample generation process.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{3.3}$$

REVERSE DIFFUSION PROCESS

The reverse process in denoising diffusion probabilistic models (DDPMs) is parametrized by a prior distribution $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ and a learnable transition kernel $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. The choice of the prior distribution as $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ is based on the design of the forward process, ensuring that the base distribution at the end of the forward process approximate to the standard normal distribution ($q(\mathbf{x}_T) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$).

The joint distribution of the full trajectory of the reverse process is defined as follows:

$$p_\theta\left(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T\right) := p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \tag{3.4}$$

Denoising diffusion probabilistic models (DDPMs) defines The learnable transition kernel $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ that sample from x_{t-1} conditioned on x_t as follows:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right) \tag{3.5}$$

Where θ represent the model parameter, whith the mean $\mu_\theta(\mathbf{x}_t, t)$ and variance $\Sigma_\theta(\mathbf{x}_t, t)$ being parametrized by deep neural networks. $\mu_\theta(\mathbf{x}_t, t)$ is trained to predict the mean of the less noisy data x_{t-1} conditioned on the noisy data x_t.

The effectiveness of the sampling process in Denoising Diffusion Probabilistic Models (DDPMs) hinges on the successful training of the reverse process to accurately reverse the forward process in real-time. This training objective is accomplished By adjusting the parameter θ through the optimization process so that the joint distribution of the reverse process approaches and approximates that of the forward process.

During the training phase of the reverse process, where the objective is to learn the true data distribution $q(x_0)$ through $p(x_0)$, we can optimize a variational bound on the negative log-likelihood using the following formulation:

$$\mathbb{E}\left[-\log p_\theta\left(\mathbf{x}_0\right)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta\left(\mathbf{x}_{0:T}\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right]=\mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right)-\sum_{t \geq 1} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right]=: L \quad (3.6)$$

The authors of denoising Diffusion Probabilistic Models (DDPMs) show that loss function 3.6 can be decomposed into several terms as follows:

$$\mathbb{E}_q[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right)}_{L_T}+\sum_{t>1} \underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)}_{L_{t-1}} \underbrace{-\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}_{L_0}] \quad (3.7)$$

The loss in Eq 3.7 is not tractable to estimate. They ignored the fact that the forward process variances β_t are learnable and instead fix them to constants. Thus, in their implementation, the approximate posterior $q(x_T|x_0)$ has no learnable parameters, so L_T is a constant during training and can be ignored. L_0 measure the likelihood of input clean data x_0 even the noisy data x_1 under the denoising process.

In L_{t-1}, where $q(x_{t-1}|x_t, x_0)$ is a tractable posterior distribution (defined in Eq 3.8) predicts the less noisy data x_{t-1} condition on the noisy data x_t and the clean data x_0.

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \tilde{\mu}_t\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t \mathbf{I}\right) \quad (3.8)$$

Where the mean $\tilde{\mu}_t$ is a simple weighted average of the clean data x_0 and the noisy data x_t (defined in Eq 3.9). The $\tilde{\beta}_t$ is determined based on the parameter of the forward process (defined in Eq 3.10).

$$\tilde{\mu}_t\left(\mathbf{x}_t, \mathbf{x}_0\right):=\frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1-\bar{\alpha}_t} \mathbf{x}_0+\frac{\sqrt{1-\beta_t\left(1-\bar{\alpha}_{t-1}\right)}}{1-\bar{\alpha}_t} \mathbf{x}_t \quad (3.9)$$

$$\tilde{\beta}_t:=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t \quad (3.10)$$

Given that both $q(x_{t-1}|x_t, x_0)$ and $p_\theta(x_{t-1}|x_t)$ are Gaussian distributions, the loss function L_{t-1} can be expressed as follows:

$$L_{t-1}=D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)=\mathbb{E}_q\left[\frac{1}{2 \sigma_t^2}\left\|\tilde{\mu}_t\left(\mathbf{x}_t, \mathbf{x}_0\right)-\mu_\theta\left(\mathbf{x}_t, t\right)\right\|^2\right]+C \quad (3.11)$$

The Eq 3.11 represents the square distance between the mean of the tractable posterior of the forward process and the mean of the reverse process. C is a constant that does not depend on any training parameters. The authors of DDPMs observe that the mean of the tractable distribution in Eq 3.9 can be expressed as the following:

$$\tilde{\mu}_t\left(\mathbf{x}_t, \mathbf{x}_0\right)=\frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon\right) \tag{3.12}$$

So they propose to present the mean of the reverse process (as shown in Eq 3.13) using the deep neural network ϵ_θ, parameterized by θ, and it predicts the noise vector ϵ given x_t and t.

$$\mu_\theta\left(\mathbf{x}_t, t\right)=\frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta\left(\mathbf{x}_t, t\right)\right) \tag{3.13}$$

In order to estimate the mean of less noisy data, it should take x_t and subtract it from the neural network ϵ_θ. This neural network ϵ_θ is trained to predict the noise component that was utilized in generating x_t, thereby representing the denoising process. According to the parametrization of Eq 3.13. The L_{t-1} loss function can be reformulated as the following:

$$L_{t-1}=\mathbb{E}_{\mathbf{x}_0 \sim q\left(\mathbf{x}_0\right), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left[\underbrace{\frac{\beta_t^2}{2 \sigma_t^2\left(1-\beta_t\right)\left(1-\bar{\alpha}_t\right)}}_{\lambda_t}\left\|\epsilon-\epsilon_\theta\left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \epsilon, t\right)\right\|^2\right]$$

Such that λ_t is a weighting function that ensures that the training objective is weighted properly for maximum data likelihood training. However authors of DDPMs found that λ_t weight is often very large for small time steps t, so they propose to use L_{simple} loss function (defined in Eq 3.14), where they set λ_t to equal to 1 because according to the experiences, it improves the sample quality.

$$L_{\text{simple}}=\mathbb{E}_{\mathbf{x}_0 \sim q\left(\mathbf{x}_0\right), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(1, T)}\left[\left\|\epsilon-\epsilon_\theta\left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \epsilon, t\right)\right\|^2\right] \tag{3.14}$$

Such that the term $\mathcal{U}(1, T)$ represents a uniform distribution over the set $1, 2, \ldots, T$.

In brief, the authors of DDPMs proposed a technique to enhance the quality of samples within the framework of variational lower bound (VLB) optimization. Their approach involved reweighting various terms within the loss function, resulting in notable performance improvements. Interestingly, they discovered a substantial connection between their modified loss function L_{simple} and the training objective of noise-conditional score networks (NCSNs), a specific category of score-based generative models previously introduced by Song and Ermon [85]. In the next subsection, we will delve into the details of NCSNs and their implications.

2.2 Score-Based Generative Models (SGMs)

Score-based generative models (SGMs) belong to a class of generative models that rely on the direct estimation of the score function [85]. The score function of a probability density function (PDF) $p(x)$ is defined as the gradient of the log probability density, denoted as $\nabla_{\mathbf{x}} \log p(\mathbf{x})$. This gradient captures the derivative of the logarithm of the PDF with respect to the data point x. Following the notations introduced in DDPM section, let $q(x_0)$ denote the data distribution, and $0 < \sigma_1 < \sigma_2 < \cdots < \sigma_t < \cdots < \sigma_T$ be a sequence of noise levels. In order to estimate the score functions, SGMs adopt a similar strategy to that of Diffusion Probabilistic models and DDPMs, which is perturbing a data point x_0 to x_t using the Gaussian noise distribution $q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_t ; \mathbf{x}_0, \sigma_t^2 I\right)$. This yields a sequence of noisy data densities $q\left(\mathrm{x}_1\right), q\left(\mathrm{x}_2\right), \cdots, q\left(\mathrm{x}_T\right)$, where $q\left(\mathbf{x}_t\right):=\int q\left(\mathbf{x}_t\right) q\left(\mathbf{x}_0\right) \mathrm{d} \mathbf{x} 0$.

The objective of SGMs is to estimate the score functions for each perturbed data distribution. This is achieved through training a noise-conditional score network (NCSN) [85]—a deep neural network model specifically designed for this purpose. The NCSN takes into account the noise levels and estimates the corresponding score functions.

Score estimation, involves established techniques such as score matching [87], denoising score matching [88, 89], and sliced score matching [90]. One can directly employ one of these techniques to train the noise-conditional score networks using perturbed data points. For instance, using denoising score matching and similar notations as in Eq. 3.14, the training objective is given by

$$\mathbb{E}_{t \sim \mathcal{U}[1, T], \mathrm{x}_0 \sim q(\mathrm{x}_0), \mathrm{x}_t \sim q(\mathrm{x}_t \mid \mathrm{x}_0)}\left[\lambda(t) \sigma_t^2\left\|\nabla_{\mathrm{x}_t} \log q\left(\mathrm{x}_t\right)-\mathrm{s}_\theta\left(\mathrm{x}_t, t\right)\right\|^2\right] \tag{3.15}$$

$$\stackrel{(i)}{=} \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathrm{x}_0 \sim q(\mathrm{x}_0), \mathrm{x}_t \sim q(\mathrm{x}_t \mid \mathrm{x}_0)}\left[\lambda(t) \sigma_t^2\left\|\nabla_{\mathrm{x}_t} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)-\mathrm{s}_\theta\left(\mathbf{x}_t, t\right)\right\|^2\right]+\mathrm{C} \tag{3.16}$$

$$\stackrel{(ii)}{=} \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathrm{x}_0 \sim q(\mathrm{x}_0), \mathrm{x}_t \sim q(\mathrm{x}_t \mid \mathrm{x}_0)}\left[\lambda(t)\left\|-\frac{\mathbf{x}_t-\mathrm{x}_0}{\sigma_t}-\sigma_t \mathbf{s}_\theta\left(\mathbf{x}_t, t\right)\right\|^2\right]+\mathrm{C} \tag{3.17}$$

$$\overset{(iii)}{=} \mathbb{E}_{t\sim\mathcal{U}[1,T],\mathrm{x}_0\sim q(\mathrm{x}_0),\epsilon\sim\mathcal{N}(0,\mathrm{I})}\left[\lambda(t)\left\|\epsilon+\sigma_t\mathbf{s}_\theta\left(\mathrm{x}_t,t\right)\right\|^2\right]+\mathrm{C} \tag{3.18}$$

The $\nabla_{\mathbf{x}}\log p(\mathbf{x})$ is the marginal score of the diffused data and $s_\theta(x_t,t)$ is a neural network that is trained to predict the marginal score of the diffused data.

The $^{(i)}$ is derived from the work of Vincent et al. (2011) [91], $^{(ii)}$ assumes $q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_t;\mathbf{x}_0,\sigma_t^2 I\right)$, and $^{(iii)}$ follows from the fact that $\mathbf{x}_t=\mathbf{x}_0+\sigma_t\boldsymbol{\epsilon}$. The positive weighting function $\lambda(t)$ and constant C do not depend on the trainable parameter θ.

Comparing Eq. 3.18 with Eq. 3.14, it is evident that the training objectives of DDPMs and SGMs are equivalent when setting $\boldsymbol{\epsilon}_\theta(\mathbf{x},t)=-\sigma_t\mathbf{s}_\theta(\mathbf{x},t)$.

Regarding sample generation, SGMs utilize iterative approaches to produce samples from $s_\theta(x,T), s_\theta(x,T-1),\ldots,s_\theta(x,0)$ sequentially. Due to the decoupling of training and inference in SGMs, various sampling methods can be employed beyond the training stage. One of the initial sampling techniques for SGMs is called Annealed Langevin dynamics (ALD) [85].

There exists a multitude of score-based sampling methods that can be employed for the purpose of sample generation. These methodologies encompass diverse techniques, including Langevin Monte Carlo [85], stochastic differential equations [86,92], ordinary differential equations [93], and their combinations [86].

2.3 STOCHASTIC DIFFERENTIAL EQUATIONS (SCORE SDES)

Denoising Diffusion Probabilistic Models (DDPMs) and Score-Based Generative Models (SGMs) can be extended to handle undefined time steps or noise levels by formulating them as solutions to stochastic differential equations (SDEs). This generalization is referred to as Score SDE [86], which utilizes SDEs for noise perturbation and sample generation while estimating score functions of noisy data distributions to perform the denoising process.

In Score SDE, the perturbation of data with noise is described by the following stochastic differential equation (SDE) [86]:

$$\mathrm{d}\mathbf{x}=\mathbf{f}(\mathbf{x},t)\mathrm{d}t+g(t)\mathrm{d}\mathbf{w} \tag{3.19}$$

Here, $\mathbf{f}(\mathbf{x},t)$ represents the drift function, and $g(t)$ is the diffusion function of the SDE. The term $\mathrm{d}\mathbf{w}$ corresponds to a standard Wiener process, which is a white Gaussian noise. Both DDPMs and SGMs discretize the forward processes based on this SDE. Specifically, for DDPMs, the corresponding SDE is given by:

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}dt + \sqrt{\beta(t)}\mathrm{d}\mathbf{w} \tag{3.20}$$

In this equation, $\beta\left(\frac{t}{T}\right) = T\beta_t$ as T approaches infinity. Similarly, for SGMs, the corresponding SDE is expressed as:

$$\mathrm{d}\mathbf{x} = \sqrt{\frac{\mathrm{d}\left[\sigma(t)^2\right]}{\mathrm{d}t}}\mathrm{d}\mathbf{w} \tag{3.21}$$

Here, $\sigma\left(\frac{t}{T}\right) = \sigma_t$ as T goes to infinity. The distribution of x_t in the diffusion process is denoted as $q_t(x)$. To recover the original data distribution p_0 from a perturbed sample p_T obtained from the forward SDE (Eq. 3.19), a reverse-time SDE is employed. By reversing the diffusion process and starting with a sample from p_T, the reverse-time SDE allows generating samples from the desired data distribution p_0. The reverse-time SDE is given by:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2\nabla_{\mathbf{x}} \log q_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\overline{\mathbf{w}} \tag{3.22}$$

In this equation, $\overline{\mathbf{w}}$ is a standard Wiener process when time flows backward, and dt represents an infinitesimal negative time step. The solution trajectories of the reverse SDE have the same marginal densities as those of the forward SDE, but they evolve in the opposite time direction [86]. These reverse-time SDE solutions gradually convert noise to data. Furthermore, a probability flow ordinary differential equation (ODE) exists, which shares the same marginals as the reverse-time SDE. The probability flow ODE is expressed as:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g(t)^2\nabla_{\mathbf{x}} \log q_t(\mathbf{x})\right] \mathrm{d}t \tag{3.23}$$

Both the reverse-time SDE and the probability flow ODE enable sampling from the desired data distribution since their trajectories possess the same marginals. To estimate the score function at each time step t, denoted as $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, a time-dependent score model $s_\theta(x_t, t)$ is parameterized. The score-matching objective is generalized to continuous time, resulting in the following objective function:

$$\mathbb{E}_{t\sim\mathcal{U}[0,T], \mathbf{x}_0\sim q(\mathbf{x}_0), \mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}0)}\left[\lambda(t)\left|\mathbf{s}_\theta\left(\mathbf{x}_t, t\right) - \nabla_{\mathbf{x}_t} \log q_t\left(\mathbf{x}_t \mid \mathbf{x}_0\right)\right|^2\right] \tag{3.24}$$

Here, $\mathcal{U}[0, T]$ represents the uniform distribution over the interval $[0, T]$, and the remaining symbols follow the score-matching objective as in previous work. This objective is utilized to train the score model within the Score SDE framework. Various numerical techniques such as annealed Langevin dynamics [85], numerical SDE solvers [86], numerical ODE solvers [86], and predictor-corrector methods (combining MCMC and numerical ODE/SDE solvers) [86] can be employed to solve the reverse-time SDE (Eq. 3.22) and the probability flow ODE (Eq. 3.23) for generating samples from the desired data distribution.

3 Diffusion Models tasks in the Medical Field

Over the past decade, generative models have demonstrated significant impact across diverse domains, including images [94, 95], audio [96], and text [97], and there has been a notable increase in research focused on generative models applied to medical image synthesis [16]. Among the cutting-edge generative models that have emerged, Diffusion Models have gained considerable attention and have been extensively applied in various areas such as image generation [14, 16–18], image segmentation [13], image inpainting [11], and image denoising [19, 98, 99], data augmentation [20, 21]. Furthermore, it is anticipated that further investigations will uncover additional tasks in this field.

In the field of medical imaging, there has been a recent surge in the adoption of diffusion-based techniques. Consequently, several survey papers have been published [80, 83], providing comprehensive overviews of deep generative models in medical imaging applications. These surveys encompass different datasets and specifically highlight the utilization and applications of diffusion models. While some of these papers focus on specific tasks, others concentrate on specific types of medical image data. Noteworthy references in this regard include:

3.1 Diffusion models for images generation

- For the specific application of generating synthetic images of lungs in X-Ray and computed tomography (CT) modalities. The study from the paper [16] employed two distinct experiments: In the first experiment, the researchers utilized the OpenAI DALL.E2 API [1] to generate images based on textual input. In the second experiment, three subsets of these generated

[1] https://openai.com/blog/openai-api

images were randomly selected and presented to two expert radiologists who were trained in the interpretation of medical scans.

The radiologists were assigned two primary tasks during the evaluation. Firstly, they were requested to categorize each image as either real, fake, or uncertain based on their professional expertise and perception.

Secondly, the radiologists were asked to provide succinct descriptions outlining potential lung conditions or disease diagnoses that could be inferred from the images, such as normal lungs, significantly damaged lungs, or lungs affected by pneumonia.

Importantly, the radiologists were not provided with any prior information regarding the authenticity of the images. In fact, all images presented to the radiologists were entirely synthetic in nature.

Furthermore, the radiologists worked independently without any knowledge of each other's assessments. It is noteworthy that one of the radiologists possessed prior knowledge of artificial intelligence and generative models, whereas the other radiologist had no prior exposure to deep generative models.

Results: in the table 3.1.

	Real	Fake	Uncertain
Radiologist A	14 X-ray and 3 CT	4 X-ray and 17 CT	2 X-ray
Radiologist B	10 X-ray and 2 CT	10 X-ray and 18 CT	

Table 3.1: presents the evaluation results of the two radiologists, who were independently presented with a common set of 40 images consisting of 20 X-ray images and 20 CT images

Agreement between radiologists: Of the 20 CT images, only three images were labeled as real by both radiologists. Similarly, five X-Ray images were marked as real by both radiologists. There were 2 X-Ray and 2 CT images for which both the radiologists were uncertain. For the second ask, they asked the radiologist to provide a description of what the images may reveal, the radiologists made some interesting observations, some examples are in the table 3.2:

Image modality	Remarks*
X-ray	Left lower lobe effusions Possibility of pneumonia Bilateral infection
CT	Possible effusions Pneumonia

Table 3.2: Samples of remarks from radiologists (no-specific order)

- In another study in [18], researchers aimed to investigate and enhance the representational capacities of large pre-trained foundational models within the context of medical concepts. Specifically, their focus was on harnessing the potential of the Stable Diffusion model to generate domain-specific images encountered in the field of medical imaging. The research delved into the sub-components of the Stable Diffusion pipeline to fine-tune the model for generating medical images.

  For this task, two large CXR datasets were used: The CheXpert dataset contains 224,316 chest radiographs of 65,240 patients treated at Stanford Hospital between October 2002 and July 2017 in both inpatient and outpatient centers [100]. The second dataset, MIMIC-CXR (version 2.0.0), contains a total of 377,110 images from studies performed at the Beth Deaconess Medical Center in Boston, MA, USA under institutional review board approval [101].

  The effectiveness of these endeavors was evaluated through a combination of quantitative image quality metrics and qualitative assessments conducted by expert radiologists, ensuring a comprehensive representation of the clinical content conveyed by conditional text prompts. Notably, the highest-performing model exhibited notable improvements over the stable diffusion baseline, showcasing the ability to conditionally introduce realistic-looking abnormalities into synthetic radiology images while maintaining a classifier accuracy of 95% for detecting said abnormalities.

- The following paper [17] explores the utilization of a latent diffusion model to synthesize a dataset of chest X-ray images that are both anonymous and of high quality. The primary objective of the researchers is to assess the viability of using solely synthetic training datasets for the purpose of learning and identifying thoracic abnormalities in chest radiographs. To achieve this, they employ a Latent Diffusion Model (LDM) [102], which enables the generation of high-quality class-conditional images by sampling from a real data distribution.

  In order to conduct their investigation, the researchers leverage a large-scale dataset known as ChestX-ray14 [103], comprising a total of 112,120 chest radiographs obtained from 30,805 individual patients. To address privacy concerns and prevent the transfer of biometric information during the image generation process, the researchers propose a privacy-enhancing sampling strategy. To assess the quality and potential suitability of the generated images as exclusive training data, the researchers evaluate their performance in a thoracic abnormality classification task. A comparative analysis with a real classifier reveals competitive results, with a performance gap of only 3.5% in terms of the area under the receiver operating characteristic curve. By employing the latent diffusion model and the proposed privacy-enhancing sampling strategy, this study offers insights into the generation of anonymous and high-quality chest radiographs. The findings highlight the feasibility of employing synthetic datasets for training purposes in the field of thoracic abnormality recognition, showcasing the potential of such an approach in medical imaging research.

- The last paper for this task is the following [14] where the researchers employ Latent Diffusion Models (LDM) [102], to generate synthetic images derived from high-resolution 3D brain scans. To accomplish this, they utilize a dataset of 31,740 T1-weighted magnetic resonance imaging (MRI) training images sourced from the UK Biobank [104] for training their models. To effectively handle the challenges associated with applying diffusion models to these high-resolution 3D data, the researchers integrate compression models into their architecture, drawing inspiration from the framework of Latent Diffusion Models (LDM). Moreover, the generation of images is conditioned on various covariables, including age, sex, and brain structure volumes.

  The researchers perform a comparative analysis between the synthetic images produced by their approach and state-of-the-art methods based on Generative Adversarial Networks (GANs). Additionally, they publicly release their synthetic dataset, which consists of 100,000 brain images, to contribute to the scientific community's research efforts in this domain.

  As a result, it showed that LDMs are promising models to be explored in medical image generation.

3.2 DIFFUSION MODELS FOR IMAGE SEGMENTATION

- In this paper [13], the researchers introduce a novel approach named **MedSegDiff**, which is the first Diffusion Probabilistic Model-based model designed for general medical image segmentation tasks. The efficacy of **MedSegDiff** is assessed through its application to three distinct medical segmentation tasks: optic cup segmentation, brain tumor segmentation, and thyroid nodule segmentation.

  To enhance the reconstruction accuracy, the researchers leverage a corrupted current-step mask that dynamically enhances the conditioning features. Additionally, to address the issue of high-frequency noises present in the corrupted mask, they propose the utilization of a feature frequency parser (FF-Parser) that filters the features in the Fourier space.

  The performance of **MedSegDiff** is evaluated across the three aforementioned medical segmentation tasks, which encompass various image modalities. These tasks involve optic cup segmentation utilizing fundus images, brain tumor segmentation employing MRI images, and thyroid nodule segmentation utilizing ultrasound images. Notably, **MedSegDiff** demonstrates state-of-the-art (SOTA) performance across all three medical segmentation tasks, irrespective of the imaging modality.

3.3 Diffusion models for image inpainting

- In the conducted investigation [11], a novel diffusion model is introduced for multitask brain tumor inpainting on multi-sequential brain magnetic resonance imaging (MRI) scans. The proposed model demonstrates the ability to perform various inpainting tasks, including the creation of individual tumoral components, a multi-component tumor, or normal-appearing brain tissue, all within a single inference iteration. Moreover, the model accommodates two distinct types of input regions of interest (ROIs): the free-form ROI and the bounding box ROI. For the free-form ROI, the model precisely inpaints a lesion that conforms to the boundaries of the input ROI, while for the bounding box ROI, it generates a random lesion and its surrounding tissue in a manner that aligns with the given bounding box. Furthermore, the researchers showcase several capabilities of the model, such as generating infinite instances of synthetic images based on a specific input with varying randomization seeds, adjusting the weighting of user-defined ROIs to emphasize their influence on the generated image, and enabling fast inference through the adoption of a DDIM protocol.

3.4 Diffusion models for image denoising

- In a recent study [98], the researchers propose the utilization of a score-based diffusion model to address the denoising task. Specifically, they suggest *hijacking* the generative process of diffusion models by initiating it from the distribution of the noisy images rather than pure Gaussian noise. To preserve fine structures during denoising, a novel low-frequency constraint is introduced, which establishes a natural connection with the recent theory of stochastic contraction in diffusion models [105]. Furthermore, the researchers propose a method to super-resolve the denoised image using the same score function employed for denoising. This innovative approach yields sharper images that retain high-frequency information, an accomplishment not previously reported with widely used self-supervised denoising methods. For training the network, the researchers utilize open-sourced data, including the fast MRI knee dataset [106]. Specifically, they adhere to the recommended training guidelines outlined in [107], employing fully-sampled single coil MRI magnitude images with a resolution of 320×320 pixels. The experimental outcomes of their approach exhibit state-of-the-art performance, surpassing other comparative methods by a substantial margin in terms of signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR).

- In a recent study focusing on retinal optical coherence tomography (OCT) denoising [19], the researchers propose the utilization of a diffusion probabilistic model. The model is trained using 6 volumes of optic nerve head (ONH) scans from the human retina and evaluated on 6 volumes of fovea scans. Each volume consists of 500 b-scans with dimensions of 512 × 500 pixels. To assess the model's performance under varying speckle levels, the data is acquired with three different signal-to-noise ratio levels (92dB, 96dB, 101dB). Since the model aims to learn the speckle pattern rather than the retina's appearance, it is not necessary for the reference image used during training to be the true noise-free image. In this study, the self-fusion method [108, 109] is employed to obtain a clean reference image, and the parameterized Markov chain is trained using variational inference. The algorithm offers flexibility in producing denoised results at different levels by adjusting the number of reverse steps. This adaptability is advantageous as different tasks may require varying degrees of fine detail retention in denoised images.

- In a recent study focusing on Low-Dose computed Tomography (LDCT) denoising [99], the authors present a conditional denoising diffusion probabilistic model (DDPM) to enhance the denoising performance of LDCT images. The proposed approach utilizes a U-Net architecture to learn the reverse diffusion process of a normal-dose CT (NDCT) image conditioned on its LDCT counterpart. By gradually sampling from a normal noise distribution, the DDPM progressively recovers cleaner images while preserving clinically important details and removing structural noise. To improve computational efficiency, a fast ordinary differential equation (ODE) solver [110] is employed, resulting in a significant speed increase of 20 times compared to the original DDPM implementation.

  For this study, the authors selected the 2016 NIH-AAPM Mayo Clinic Low-Dose CT Grand Challenge dataset. The dataset consists of 2,378 paired NDCT and LDCT images with a thickness of 3mm obtained from 10 patients. From this dataset, 1,923 paired images from 8 patients were chosen as the training set, while 455 paired images from the remaining 2 patients were used as the test set. To ensure consistency, the image matrix was resampled to a size of 256x256.

  Overall, the results demonstrate that the ODE Solver has both a high denoising performance and a high sampling efficiency. The use of an ODE solver for DDPM sampling will guide to a great potential for clinical applications.

3.5 Diffusion models for lesion detection

- In this paper [12], the authors propose a novel pixel-wise anomaly detection approach based on Denoising Diffusion Implicit Models (DDIMs). The approach consists of two parts. Firstly, the researchers train a Denoising Diffusion Probabilistic Model (DDPM) and a binary classifier on a dataset of healthy and diseased subjects. Secondly, they encode the anatomical information of an image using the reversed sampling scheme of DDIMs, simulating the noising process. Then, during the denoising process, they employ the deterministic sampling scheme proposed in DDIM with classifier guidance to generate an image of a healthy subject. The final pixel-wise anomaly map is obtained by calculating the difference between the original and synthetic images.

  To evaluate the effectiveness of the proposed approach, the algorithm was applied to two different medical datasets: the BRATS2020 brain tumor challenge and the CheXpert dataset. Comparative analysis was conducted against standard anomaly detection methods such as the Fixed-Point GAN (FP-GAN) [111]and the variational autoencoder (VAE) [112].

  The results indicate that the proposed approach generates realistic-looking images while preserving important details, demonstrating its potential in pixel-wise anomaly detection tasks.

- The proposed paper [15] introduces a novel approach for detecting anomalies in medical images. The approach involves using deep diffusion probabilistic models (DDPMs) to destroy the image and reconstruct a healthy approximation. Instead of using Gaussian noise, simplex noise - a popular method in computer graphics - is employed for anomaly detection. The key contributions of the paper include a partial diffusion strategy where the anomalous image is noised to a parameterized timestep λ and reconstructed from the corruption. Additionally, multi-scale (multi-octave) simplex noise is utilized to allow larger anomalous regions to become reconstructed as healthy regions.

  To evaluate the proposed approach, the Neurofeedback Skull-Stripped (NFBS) repository [113] was used, which contains 125 T1-weighted MRI scans with dimensions 256 × 256 × 192, containing the full skull, skull stripped, and brain mask. Results show that the proposed approach (referred to as AnoDDPM with simplex noise) successfully captures large anomalous regions with stable training that does not require large datasets.

  The use of multi-scale (simplex) noise was found to offer significant improvements in terms of capturing larger anomaly shapes.

3.6 Diffusion models for image translation

- The present research paper [10] focuses on addressing the challenge of unsupervised medical image synthesis through the introduction of a novel adversarial diffusion model, **SynDiff**. The primary objective of **SynDiff** is to enable efficient and high-fidelity modality translation in the context of medical image analysis.

  This paper makes several notable contributions to the field. Firstly, it presents the first adversarial diffusion model specifically designed for high-fidelity medical image synthesis. By utilizing diffusion-based techniques, **SynDiff** offers a unique approach to unsupervised medical image translation, allowing for training on unpaired datasets consisting of different source and target modalities.

  To facilitate efficient image sampling, the authors propose a novel source-conditional adversarial projector. This projector is designed to capture reverse transition probabilities over large step sizes, thereby enhancing the effectiveness of image generation and **SynDiff**.

  To evaluate the performance and capabilities of **SynDiff**, the authors conducted comprehensive demonstrations using two multi-contrast brain MRI datasets (IXI1 and BRATS) as well as a multi-modal pelvic MRI-CT dataset. Through these demonstrations, the authors assessed the efficacy of **SynDiff** in achieving modality translation in different medical imaging scenarios.

  The experimental results presented in the paper clearly indicate the superiority of SynDiff when compared to competing generative adversarial networks (GAN) and diffusion models. **SynDiff** exhibits improved performance and produces more faithful and realistic synthesized medical images, demonstrating its effectiveness in addressing the challenges of unsupervised medical image synthesis.

3.7 Diffusion models for data augmentation

- The present study focuses on enhancing the performance of dermatology classifiers through the utilization of a pipeline that leverages the transformer-based generative model, DALL·E 2 [20]. The primary objective is to produce photorealistic images of skin diseases to augment the training dataset and improve classification accuracy.

  To achieve this, the researchers employed OpenAI's DALL·E2 [95] model to generate photorealistic synthetic images based on seed images from the Fitzpatrick 17k dataset. For each skin condition, a set of sixteen seed images was randomly sampled, comprising eight images from the lightest Fitzpatrick skin types and eight images from the darkest skin types.

To evaluate the impact of the synthetic images on classification performance, image classification models were trained using various train and test splits. These models aimed to predict skin condition labels among the seven skin conditions of interest.

The results of the study demonstrate the efficacy of the targeted generation of synthetic images in improving the performance of dermatological classifiers on a diverse benchmark dataset. Notably, the improvement is particularly notable for underrepresented groups, indicating the potential for the generated synthetic images to address the challenges posed by limited data availability for certain skin conditions. By utilizing the DALL·E 2 model and conducting comprehensive experiments, this study contributes to the advancement of dermatology classifiers. The findings highlight the potential of synthetic image generation as a means to enhance classification accuracy and address data limitations in dermatology, ultimately improving the diagnosis and treatment of skin diseases.

In the present study, the others describe a pipeline for producing photorealistic images of skin disease using the transformer-based generative model DALL·E 2. They generated photorealistic synthetic images from seed images in the Fitzpatrick 17k dataset using OpenAI's DALL·E 2 model. For a given skin condition, we randomly sampled eight images from the lightest and darkest Fitzpatrick skin types (16 total images for one condition) to use as seed images. For the training, they trained image classification models to predict skin condition labels among the seven skin conditions using different train and test splits.

As a result, they show that the targeted generation of synthetic images can be used to improve the performance of dermatological classifiers on a diverse benchmark dataset overall and particularly for underrepresented groups.

- In a recent study conducted by [21], the potential of Diffusion Probabilistic Models (DPMs) for skin disease classification was investigated. The authors focused on fine-tuning DPMs on six distinct disease conditions, namely basal cell carcinoma, melanoma, actinic keratosis, atypical melanocytic nevus, lentigo, and seborrheic keratosis. By conditioning the probabilistic diffusion-based generation on text prompt inputs using stable diffusion model [102], they demonstrated the generation of fine-grained synthetic images that closely resemble real skin disease samples.

  To facilitate their research, a fully synthetic dataset was constructed, consisting of 500 images per skin disease category. In parallel, a set of real images was randomly sampled, including 500 images per class, from a macroscopic skin image dataset. This synthetic dataset was specifically created to evaluate the impact of synthetic images on classification metrics.

  Through their comprehensive classification task involving the six skin diseases, the study

underscored the reliability of synthetic images as valuable data sources. The findings demonstrated the efficacy of synthetic images in skin disease classification, emphasizing their potential benefits in augmenting the limited availability of real-world datasets for this domain.
Overall, this study sheds light on the promising applications of DPMs and synthetic images in skin disease classification tasks, highlighting their value as reliable data sources in the absence of extensive real-world datasets.

4 CONCLUSION

This chapter provides a brief overview of the different variants of diffusion models and presents a comprehensive and detailed literature review of their applications in the medical field. The three prominent variants of diffusion models discussed are Denoising Diffusion Probabilistic Models (DDPMs), Score-Based Generative Models (SGMs), and Stochastic Differential Equations (SDEs). Each formulation offers distinct insights and methodologies for modeling the diffusion process and capturing the underlying dynamics of the data. These models have demonstrated successful utilization across various domains, including computer vision and natural language processing. Furthermore, their application in the medical domain has been extensive, encompassing tasks such as image generation, segmentation, inpainting, and denoising. These models have exhibited their efficacy in generating synthetic medical images, which can be leveraged for training and research endeavors.

Based on previous studies, diffusion models have been widely employed in diverse medical tasks such as image segmentation, image inpainting, image denoising, augmentation, and so on. Primarily, the most used datasets for these tasks are CT and MRI datasets. However, in the realm of data augmentation, dermatology datasets have been predominantly employed. Thus, in light of this, we have chosen to apply data augmentation techniques specifically to the context of diabetic retinopathy grading, an area that has received relatively less attention in this regard. The objective is to present a novel study that offers valuable insights, setting it apart from previously conducted research in the field.
Therefore, the subsequent chapter includes our utilization of the diffusion model within the framework of diabetic retinopathy grading, we will delineate the workflow and methodologies employed for this specific context.

CHAPTER 4

Data Augmentation Using Diffusion Models for Diabetic Retinopathy Diagnosis

1 Introduction

In the previous chapter (Chapter 3), the application of Diffusion models in the field of medicine was explored across various domains. Expanding upon this groundwork, the current chapter delves deeper into the specific topic of data augmentation. The purpose of this study is to evaluate the quality of images generated by Diffusion models.

The primary objective of this chapter is to assess the quality and realism of the images generated by the diffusion model. Specifically, we aim to examine the suitability of these generated images for data augmentation and their potential incorporation into a small-sized training dataset. To evaluate their efficacy, we compare the results obtained from deep-based classifiers trained on the augmented dataset with those obtained using the same deep-based classifiers under identical conditions, but with a different training dataset that underwent traditional data augmentation.

Methodology

The adopted methodology for our task involves employing data augmentation techniques to augment the dataset, followed by fitting the augmented training data to the classifier for evaluation purposes. By applying data augmentation, the aim is to enhance the size and diversity of the dataset, thus enabling the classifier to learn more robust and generalized patterns.

To address this goal, we employ two different types of data augmentation: **traditional data augmentation** and **deep-based data augmentation**. For each augmented dataset, two different deep-based classifiers are trained on the task of Diabetic retinopathy diagnosis.

2 Step1 : Employed Data Augmentation Techniques

In this methodology, we apply multiple data augmentation techniques and conducted a comparative analysis of their respective performances. We use data augmentation techniques that encompassed both traditional methods **geometric** and **Photometric**, as well as two deep-based generative models. The deep generative models employed are Deep convolution Generative Adversarial Networks (DCGAN) [1], a variant of the Generative Adversarial Network (GAN), and a pre-trained DALLE 2 model [95], which falls under the category of diffusion models. The selection of these models is based on several considerations, including resource limitations, specifically the availability of GPUs.

2.1 Traditional Data Augmentation

The initial approach adopted in this study entails the utilization of traditional data augmentation techniques that are widely employed in various image processing tasks. These techniques encompass a range of geometric image transformations, including rotation, translation, scaling, and flipping, among others. For the purpose of our specific task, we choose to apply horizontal and vertical flips as geometric transformations.

Additionally, we adopt photometric image transformations, which involve modifying the brightness or intensity values of the pixels within an image (Image Filters). In our particular case, we decide to augment the images by increasing their brightness. This photometric transformation aims to introduce variations in the intensity levels of the image pixels, potentially enhancing the model's ability to generalize and capture variations in illumination conditions.

By incorporating both geometric and photometric transformations, we aim to create a more diverse and comprehensive dataset that encompasses a wider range of variations in terms of spatial orientation and pixel intensity. This augmented dataset is expected to facilitate improved generalization and performance of the classification model.

2.2 Deep-based Data Augmentation

Deep Convolutional Generative Adversarial Networks (DCGANs)

In the paper [1], the authors introduced Deep Convolutional Generative Adversarial Networks (DCGANs) as a framework for unsupervised representation learning. DCGANs aim to learn meaningful representations from unlabeled data by training a generator network and a discriminator network in an adversarial manner [1].

To facilitate effective image-related tasks, deep convolutional neural networks (CNNs) are employed as the underlying architecture for both the generator and discriminator networks [1]. CNNs are well-suited for capturing spatial dependencies in images, making them suitable for generating high-quality images that closely resemble real images.

The primary objective of DCGANs is to enable unsupervised representation learning. The authors leverage the internal representations learned by the discriminator for this purpose. They observe that the intermediate layers of the discriminator capture meaningful and discriminative features that can be utilized in downstream tasks, such as image classification. The quality of the learned representations is assessed through qualitative analysis of the generated images and quantitative evaluation of image classification tasks, demonstrating their effectiveness [1].

DCGANs offer several advantages as outlined in [1] paper. These advantages include stable training facilitated by architectural constraints, hierarchical feature learning through convolutional and pooling layers, the generation of high-quality images with coherent structures, the ability to learn meaningful representations, enabling fine-grained control over generated outputs, semantic vector arithmetic for manipulating specific attributes, and the potential for transfer learning by utilizing pre-trained discriminator networks as feature extractors. These advantages collectively demonstrate the efficacy of DCGANs in unsupervised representation learning, image generation, and transfer learning tasks.

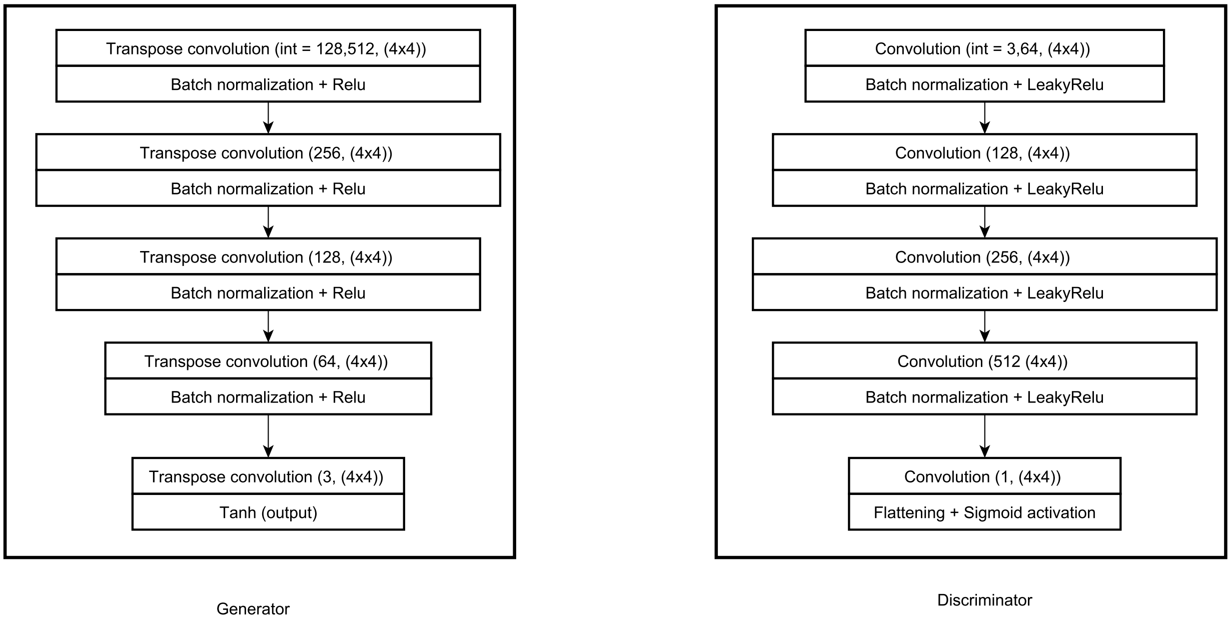
Transpose convolution (int = 128,512, (4x4))
Batch normalization + Relu
Transpose convolution (256, (4x4))
Batch normalization + Relu
Transpose convolution (128, (4x4))
Batch normalization + Relu
Transpose convolution (64, (4x4))
Batch normalization + Relu
Transpose convolution (3, (4x4))
Tanh (output)
Generator
Convolution (int = 3,64, (4x4))
Batch normalization + LeakyRelu
Convolution (128, (4x4))
Batch normalization + LeakyRelu
Convolution (256, (4x4))
Batch normalization + LeakyRelu
Convolution (512 (4x4))
Batch normalization + LeakyRelu
Convolution (1, (4x4))
Flattening + Sigmoid activation
Discriminator

Figure 4.1: The DCGAN architecture [1]

Diffusion models (DALL.E2 model):

A Concise Survey of CLIP

One of the notable advantages of DALL.E2 lies in its integration with CLIP, which refers to Contrastive Language Image Pre-training. CLIP is a robust vision-language model that has been trained on a large-scale dataset encompassing images and their corresponding textual captions extracted from the internet [114].

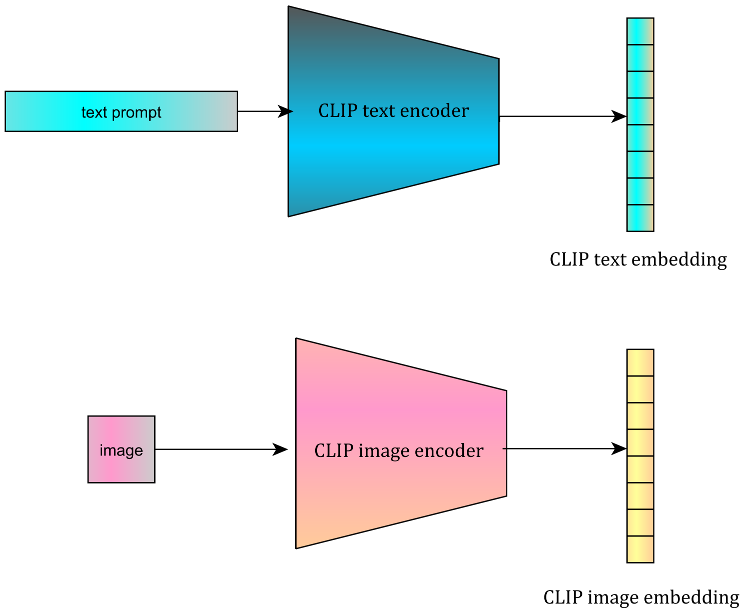

Figure 4.2: The CLIP encoders

The primary functionality of CLIP revolves around matching images with their respective textual captions [114]. It comprises two encoders, as illustrated in Figure4.2: the text encoder, responsible for generating text embeddings denoted as z_t, and the image encoder, responsible for generating image embeddings denoted as z_i. These encoders are trained together to accurately predict the correct pairings of a batch of training examples consisting of image-text pairs [114].

Within the framework of DALL.E2, CLIP is fine-tuned to function as an embedding network tailored to the specific task at hand. For instance, in the case of variation creation, CLIP exclusively generates image embeddings z_i when provided with an input image.

DALL.E2 (unCLIP)

DALL.E2 [95] is an advanced generative model designed to acquire resilient representations of images, effectively capturing both semantic and stylistic aspects.

The text-to-image generation procedure the authors presented in their paper [95]: Initially, a CLIP text embedding is utilized as input and processed by an autoregressive or a diffusion prior. This process yields an image embedding, which subsequently serves as a conditioning factor for a diffusion decoder, ultimately generating the final image output. It is noteworthy that the CLIP model remains static or "frozen" during the training of both the prior and decoder components.

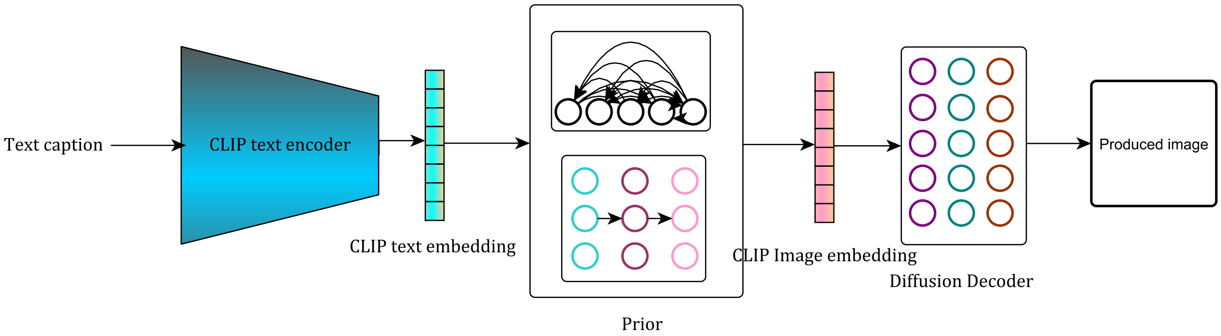

Figure 4.3: The text-to-image generation process

DALL.E2 training dataset consists of pairs (x, y) comprising images x and their corresponding text captions y.

The methodology employed in their scholarly publication [95] encompasses a training approach centered on a dataset composed of image-caption pairs (x, y), where x denotes images and y represents their respective captions. Let z_i and z_t be its CLIP image and text embeddings When provided with an image x, respectively. The architecture of the generative stack, aimed at generating images based on captions, incorporates two key components:

- *A prior* $P(z_i \mid y)$ that produces CLIP image embeddings z_i conditioned on captions y.
- *A decoder* $P(x \mid z_i, y)$ that produces images x conditioned on CLIP image embeddings z_i (and optionally text captions y).

The prior allows us to learn the CLIP image embeddings while the decoder inverts images from their CLIP image embeddings.
Stacking these two components yields a generative model $P(x \mid y)$ of images x given captions y:

$$P(x \mid y) = P\left(x, z_i \mid y\right) = P\left(x \mid z_i, y\right) P\left(z_i \mid y\right)$$

In our study, we employed the pre-trained DALL.E2 framework to utilize the sampled images for the purpose of the diffusion model for data augmentation. The utilization of this framework enabled us to encode a given image, denoted as x, into a bipartite latent representation represented by the tuple (z_i, x_T). This representation proved to be adequate for the decoder to generate a precise reconstruction of the input image. Consequently, our subsequent investigation will primarily concentrate on the decoder component, which plays a crucial role in image manipulation.

THE DIFFUSION DECODER

The authors employed a diffusion model to generate images, conditioned on CLIP image embeddings. This diffusion model, adapted from GLIDE [115], operates as a guided diffusion model that used classifier-free guidance.
conditional diffusion Incorporating conditioning information alongside the timestep information at each iteration provides a natural approach. To transform this into a conditional diffusion model, arbitrary conditioning information y can be straightforwardly included at each transition step [116], resulting in the following formulation:

$$p_\theta(x_{0:T}|y) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y)$$

Classifier-free guidence The proposed method is for providing guidance to diffusion models without the need for a distinct classifier model training. In the context of class-conditional diffusion models, the conventional label y is replaced with an empty sequence (which also refers to as $\emptyset$) during training. Then they guide towards the conventional label y using the modified prediction [115]:

$$\hat{\theta}(x_t|y) = \theta(x_t|\emptyset) + s \cdot (\theta(x_t|y) - \theta(x_t|\emptyset))$$

Here $s \geq 1$ is the guidance scale [115].

GENERATE VARIATION OF AN IMAGE USING DALL.E2

The generation of related images x with shared semantic and stylistic content, known as *image variation*, has been explored in recent research [95]. This involves the fitting of a bipartite representation, denoted as (z_i, x_T), where z_i represents the CLIP image embedding and x_T corresponds to the initial noise utilized in the application of the Decomposition of Inverse Methods (DDIM) technique [1] using the decoder. Specifically, the DDIM approach utilizes the *decoder* to reconstruct the image, while conditioning on the CLIP image embedding z_i and utilizing the initial noise x_T.

[1] the process of recovering or reconstructing an unknown image or signal from its distorted or degraded version

By considering the value of $s \geq 1$ for sampling, the image variations can be examined to gain insights into the information captured within the CLIP image embedding. Increasing the value of s enables a more comprehensive exploration of the variations and provides valuable insights into the content encoded within the CLIP image embedding.

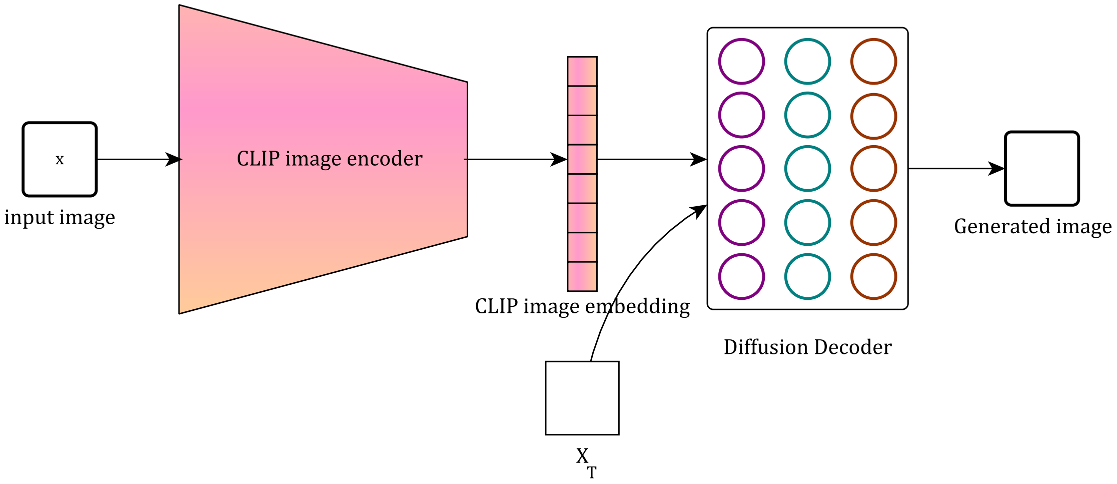

Figure 4.4: Image variation process

3 STEP2: EMPLOYED CLASSIFIERS

Furthermore, to introduce a higher level of challenge, the evaluation will focus on the task of diabetic retinopathy disease grading. This grading task poses significant complexity and requires accurate classification of retinal images based on disease severity. By conducting this evaluation, we seek to determine the effectiveness of the generated images from the diffusion model in improving the performance of deep-based classifiers for diabetic retinopathy disease grading.

Diabetic Retinopathy (DR) is predominantly characterized by retinal vascular alterations, resulting in the manifestation of initial indications such as microaneurysms. These changes arise due to the impact of diabetes on the blood vessels within the retina. Diabetic Retinopathy (DR) is a disease that is closely tied to high blood sugar levels and can result in vision loss in patients with long-term diabetes. This loss of vision can occur due to two main factors: firstly, high blood sugar levels can damage blood vessels in the retina, leading to blood and fluid leakage as well as swelling of the retina; and secondly, abnormal new blood vessels can grow on the retina, increasing pressure within the eye. As such, the examination and analysis of models for DR are of utmost importance in the development of effective screening and diagnosis tools for this disease [117].

3.1 PROPOSED CONVOLUTIONAL NEURAL NETWORK

To address the task of disease grading, we reconstruct from scratch a Convolutional Neural Network (CNN) architecture specifically tailored for this purpose. The design of our CNN model is as follows in figure 4.5:

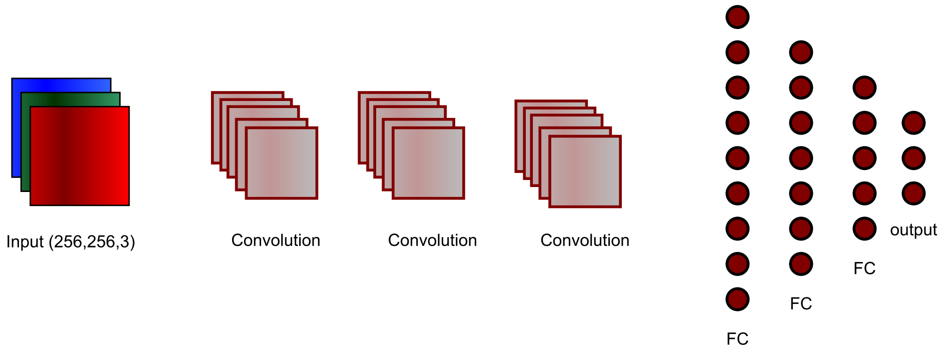

Figure 4.5: The proposed CNN architecture

The Convolutional Neural Network (CNN) architecture employed in this study consists of three convolutional layers. Each convolutional layer is followed by a batch normalization layer to normalize the outputs. The network also includes three fully connected layers, incorporating dropout regularization to mitigate overfitting, and an output layer.
This CNN architecture is developed to effectively capture and learn the intricate features present in the disease images. Moreover, this comprehensive training setup optimizes the model's ability to learn and generalize from the provided dataset, enabling accurate and robust disease grading outcomes.

3.2 FINE-TUNING RESNET50

In addition, we employed a fine-tuning approach by leveraging the ImageNet dataset to address the challenges posed by the relatively limited size of our dataset. Transfer learning enables the utilization of pre-trained models on large-scale datasets such as ImageNet, thereby reducing the extensive training time typically required by deep learning algorithms [118]. The pre-trained model, extensively trained on ImageNet, has been made publicly available, allowing for fine-tuning on other datasets. Due to its adaptability to diverse datasets, the transfer learning methodology can be readily applied to our specific dataset, facilitating efficient learning and inference in the context of the particular problem domain [119].

To fulfill this objective, the ResNet50 model [2] is selected, leveraging its deep architecture for effective feature extraction and representation learning.

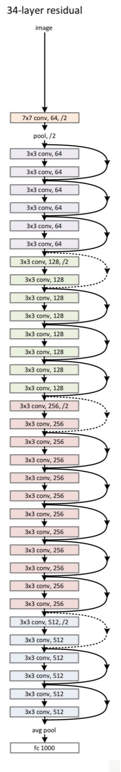

Figure 4.6: The resnet-34 architecture [2]

The resnet50 is to replace each 2-layer block in the 34-layer net with this 3-layer bottleneck block, resulting in a 50-layer ResNet.

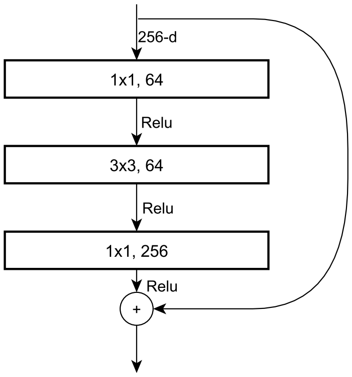

Figure 4.7: The bottleneck block [2]

In our customization process, the final fully connected layer of the ResNet50 model is substituted with two newly introduced layers, tailored to suit the specific requirements of our task. These additional layers are fine-tuned by adjusting their parameters while maintaining the remaining parameters of the ResNet50 model frozen. This approach allows us to capitalize on the pre-trained knowledge embedded within the ResNet50 model while adapting it to the intricacies of our particular classification task.

4 Conclusion

In conclusion, this chapter provided a comprehensive overview of the methodology employed in the study. The primary objective of this chapter is to assess the quality and realism of the images generated by Diffusion models and evaluate their potential for data augmentation in a small-sized training dataset.

To achieve this, we employed two types of data augmentation techniques: traditional data augmentation and two deep-based data augmentation (**DCGANs** and **Diffusion models**).

Furthermore, we evaluate the efficacy of the generated images by using them to train two deep-based classifiers (**Proposed CNN** and **Fine-tunned Resnet50**).

For the task of diabetic retinopathy disease grading. By conducting this evaluation, we aim to determine the effectiveness of the generated images in improving the performance of deep-based classifiers.

In the forthcoming chapter, we will proceed with the practical realization of the outcome through the amalgamation of preceding methodologies. Subsequently, we will present a visual representation of the findings and engage in a comprehensive discourse to scrutinize the results, while also acknowledging and examining the constraints encountered.

CHAPTER 5

Results and Disscusion

1 Introduction

In the previous chapter 4, we delineate the workflow of the chosen task elucidating the essential steps involved in data augmentation through the utilization of both traditional and deep-based techniques. Additionally, we present the deployed classifiers, highlighting their crucial role in the successful attainment of disease diagnosis.

In this chapter, we will present a detailed exploration of the aforementioned techniques applied to the specific IDRID dataset with the objective of diabetic retinopathy grading. We will provide a comprehensive insight into the utilization of these techniques, highlighting their effectiveness and impact in the context of assessing the severity of diabetic retinopathy.

Finally, we will present the final results obtained from our study, accompanied by visualizations and a comprehensive discussion. This analysis will emphasize the implications and significance of our findings within the broader context of our research.

2 The Indian Diabetic Retinopathy Image Dataset (IDRID)

The Indian Diabetic Retinopathy Image Dataset (IDRID) [120] is a meticulously curated collection of images falling under the domain of Biomedical and Health Science. This dataset was assembled using real clinical examinations conducted at an eye clinic located in Nanded, India. Retinal photographs of individuals affected by diabetes were captured with a specific focus on the macula, employing the Kowa XV-10α fundus camera. To ensure optimal image quality, the pupils of all subjects were dilated using a 0.5% concentration of tropicamide prior to image acquisition. The captured images possess a field of view of 50 degrees, a resolution of 4288 × 2848 pixels, and are stored in the JPG format. Notably, the dataset encompasses typical lesions associated with diabetic retinopathy as well as annotations of normal retinal structures at a pixel level. The comprehensive nature of this dataset renders it particularly well-suited for the development and evaluation of image analysis algorithms aimed at the early detection of diabetic retinopathy.

The dataset is organized into three distinct parts, each serving a different purpose:

a. Segmentation: Segmentation of retinal lesions associated with diabetic retinopathy as microaneurysms, hemorrhages, hard exudates, and soft exudates. Comprises original color fundus images that have been divided into a train set and a test set, consisting of 81 images in total in JPG format. In addition to the fundus images, the Segmentation part also includes ground truth images for specific lesions, including microaneurysms, hemorrhages, and hard exudates, all of which are also divided into train and test sets and stored in TIF file format.

Furthermore, this part of the dataset also contains ground truth images for the optic disc, also divided into train and test sets and stored in TIF file format.

b. Disease Grading: Classification of fundus images according to the severity level of diabetic retinopathy and diabetic macular edema. Includes original color fundus images that have been divided into 413 training images and 103 testing images, all stored in JPG format. The Disease Grading part also includes ground truth information for the severity grade of diabetic retinopathy and diabetic macular edema, also divided into train and test sets and stored in CSV file format.

c. Optic Disc and Fovea Center Localization: Automatic localization of optic disc and fovea center coordinates and also segmentation of optic disc. Includes original color fundus images divided into 413 training images and 103 testing images, all stored in JPG format. In addition to the fundus images, the Localization part includes ground truth information for the location of the optic disc center and the fovea center, both of which are divided into train and test sets and stored in CSV file format.

In this study, we choose the Disease Grading dataset. Each image in the dataset is annotated with ground truth information of the severity grade, which ranges from 0 to 4, indicating the severity of diabetic retinopathy. The selection of the Indian Diabetic Retinopathy Image Dataset (IDRID) dataset for diabetic retinopathy grading is primarily motivated by its intrinsic characteristics. The dataset possesses a noteworthy attribute in the form of a relatively limited training dataset, consisting of only 413 images. Additionally, the distribution of these images across the five distinct grade classes is imbalanced, necessitating the implementation of data augmentation techniques to address this inherent imbalance and alleviate its impact on the training process.
By employing data augmentation, our objective is to overcome the challenges posed by the limited training data and imbalanced class distribution, aiming to improve the performance and resilience of the diabetic retinopathy grading task.

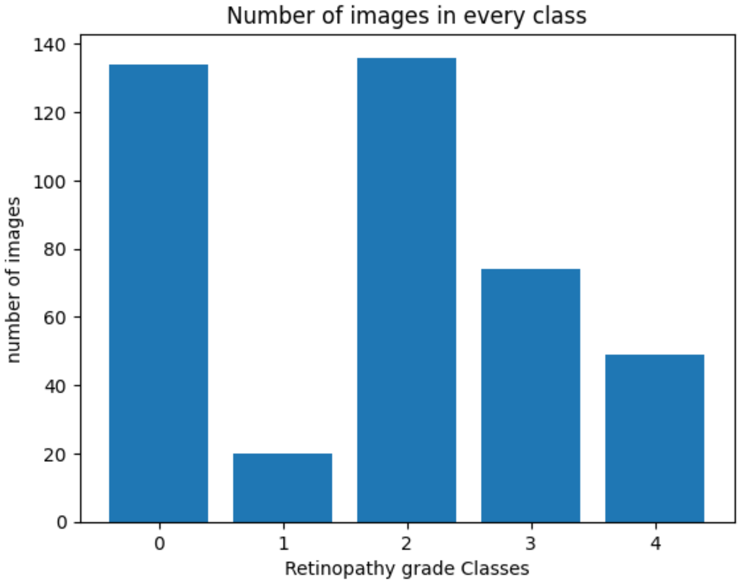

Figure 5.1: images distribution in the five grade classes

A preprocessing step is undertaken prior to the application of data augmentation techniques, involving the removal of the dark areas surrounding the training and testing images. This process entails cropping the images to exclusively retain the region corresponding to the retina. Subsequently, the cropped images are resized to a square shape with dimensions of 256x256 pixels. The purpose of this preprocessing step is to focus solely on the relevant retinal region and reduce the size of the input images, ensuring consistency in their dimensions. The following figures, Figure 5.2, display an original sample from the dataset, while Figure 5.3 shows the same image after applying the preprocessing step of cropping and resizing.

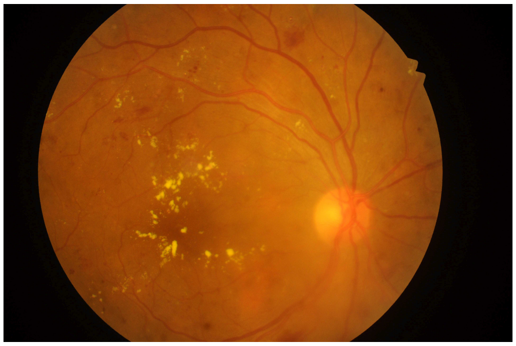

Figure 5.2: Original image sample.

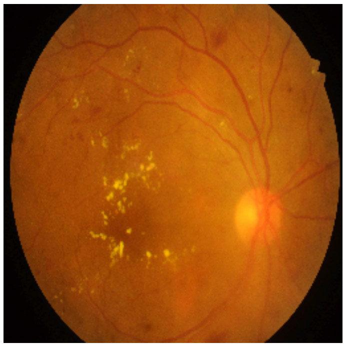

Figure 5.3: Cropped-resized image sample.

3 STEP 1: DATA AUGMENTATION

3.1 TRADITIONAL DATA AUGMENTATION

The predominant justification for adopting horizontal and vertical flips as geometric augmentations reside in the intrinsic circular morphology of the retina, wherein the mirrored rendition of the right side accurately corresponds to the left side. Moreover, the inclusion of horizontal flips is substantiated by its capacity to uphold the fundamental circular configuration of the retina, while concurrently engendering enhancements in brightness via photometric augmentation. Furthermore, the augmentation process is not conducted randomly; rather, it is employed with the objective of addressing the imbalance in image distribution across classes and augmenting the training dataset by increasing the number of available images.

By incorporating these augmentations, the resultant dataset is rendered more comprehensive and varied, thereby fostering augmented robustness and generalizability of the models trained on the enriched data.

3.2 DEEP CONVOLUTION GAN FOR DATA AUGMENTATION

The preference for Deep Convolutional Generative Adversarial Network (DCGAN) over alternative models, including BigGAN, SkyGAN, and ProGAN, stems from various factors, primarily driven by resource limitations such as GPU availability and time constraints.

DCGAN stands out due to its computational efficiency when compared to other models. Its architecture places a significant emphasis on convolutional operations, which leads to streamlined processing and a reduction in computational burden. Consequently, DCGAN becomes more viable for training under conditions of limited GPU availability.

Another critical consideration is the time-consuming nature of training deep generative models. DCGAN's reliance on smaller image sizes allows for faster iterations during the training process. This reduction in image resolution results in a decrease in the number of parameters and computations, thereby mitigating overall training time when compared to models like BigGAN which typically handle higher-resolution images.

The decision to employ 64x64-sized images within the DCGAN framework predominantly arises from the challenges associated with scaling DCGAN to accommodate high-resolution images (as we mentioned in chapter 4). Practical considerations and computational constraints play a significant role in opting for this image size. Training generative models, particularly deep convolutional neural networks, on high-resolution images can demand substantial computational resources and extensive time investment. As image size increases, so does the number of model parameters and computational requirements, which can impede efficient training processes.

Regarding the training specifics, DCGAN was trained on a Tesla T4 GPU, utilizing a learning rate of 0.0002 for 50 epochs. The images produced in the training phase of DCGAN are depicted in Figure 5.4. Additionally, Figure 5.5 illustrates the generated images obtained during the sampling phase. Furthermore, Figure 5.6 presents the relevant training information associated with the DCGAN model.

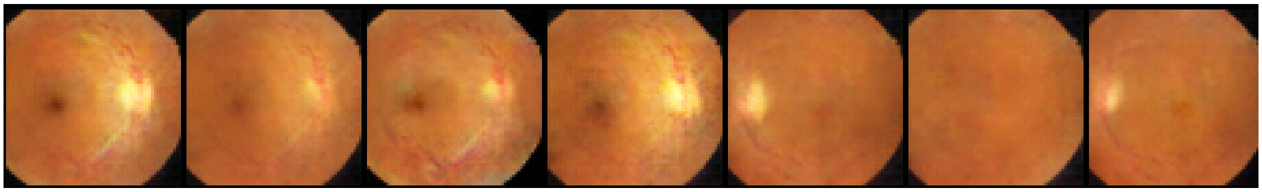

Figure 5.4: DCGAN generated images from training phace

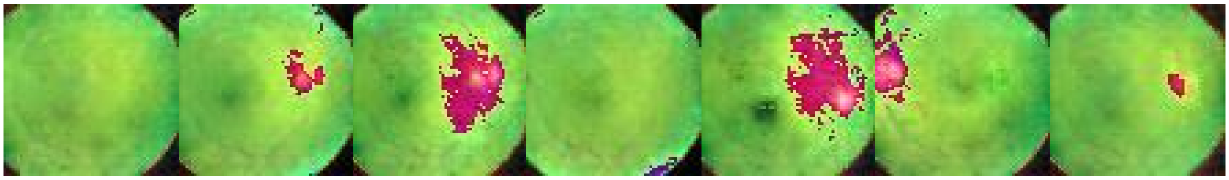

Figure 5.5: DCGAN generated images from sampling phase

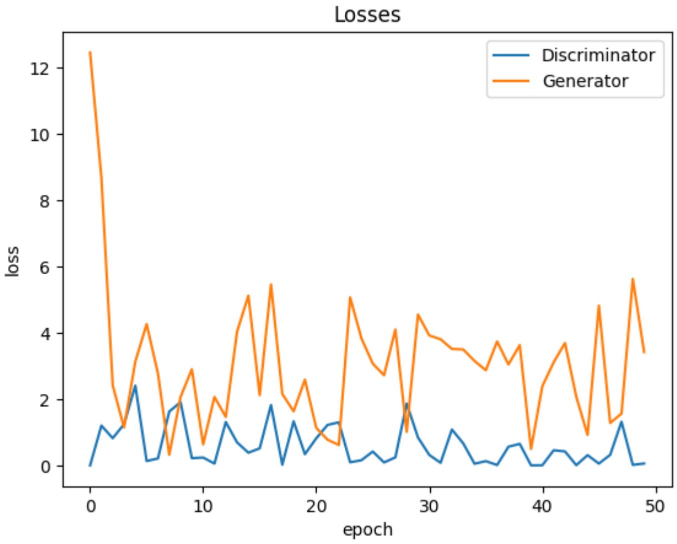

Figure 5.6: DCGAN during Training

Despite the potential of DCGANs in unsupervised representation learning, certain limitations and concerns are acknowledged by the authors of [1]. One common concern with GANs, in general, is their training dynamics, which may suffer from instability issues in their training dynamics such as vanishing or exploding gradients, mode collapse, and non-convergence [79]. In addition, DCGANs exhibit sensitivity to hyperparameters and face challenges in scaling to high-resolution images [1]. We aimed to upscale the images from a resolution of 64x64 pixels to a higher resolution of 256x256 pixels. However, due to time constraints, it is not feasible to complete the image scaling process within the given timeframe.

3.3 DIFFUSION MODELS FOR DATA AUGMENTATION

In these current years, it appears multi-existing trained diffusion models for image generation (mostly text-to-image generation) mentioning the most famous ones as: DALL.E2 [1] and GLIDE [2] from OpenAi, Stable Diffusion [3] from StabilityAi, The premium Midjourney [4], Imagen [5] from Google.

The decision to utilize the DALL.E2 model is made following a comparative analysis of results generated by various freely available models, namely Stable Diffusion and GLIDE, as well as the semi-freely available DALL.E2 model. The objective of the evaluation is to assess the quality of images generated in response to a specific text prompt, which defined the characteristics of an image from the utilized dataset. The prompt specified **photographs of the retina, which is the layer of tissue at the back of the eye that its Retinopathy grade = 1/4 and its Risk of macular edema = 0 for medical purposes**.

[1] https://openai.com/dall-e-2

[2] https://www.glideapps.com/docs/reference/integrations/openai

[3] https://stability.ai/stablediffusion

[4] https://www.midjourney.com/home/

[5] https://imagen.research.google

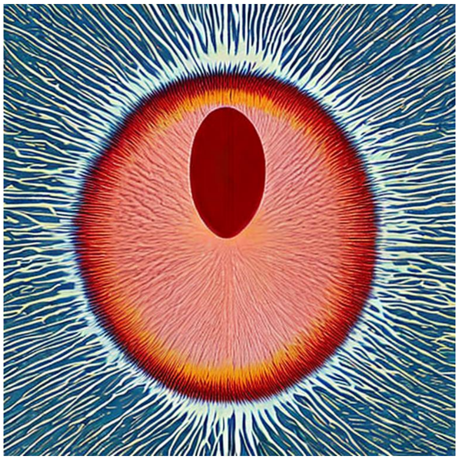

Figure 5.7: sampled from Stable Diffusion

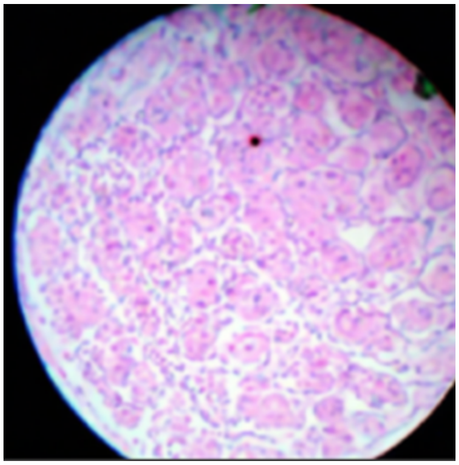

Figure 5.8: Sampled from GLIDE

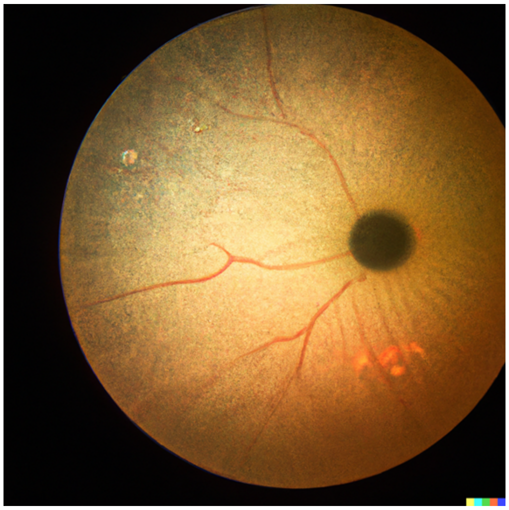

Figure 5.9: Sampled from DALL.E2

Upon analyzing the outputs from the various models used, it becomes evident that the image generated by the DALL.E2 model exhibits the highest resemblance to a retina. The comparison of the samples from different models allowed for a clear distinction in terms of visual fidelity and the ability to capture the specific characteristics of a retina. Among the evaluated models, the DALL.E2 model stood out as it produced an image that closely resembled the desired features and exhibited a high degree of similarity to a real retina (Figure 5.9). This observation highlights the effectiveness of the DALL.E2 model in generating retina-like images, making it a favorable choice for the given task.

How DALL.E2 was used?

We utilize the developer API offered by OpenAI [6] to generate novel images. This API facilitates the creation of diverse variations for each image in our dataset. However, we encounter a challenge during the process as the images have to be submitted individually to the API for generating variations, resulting in a repetitive operation that had to be performed more than 413 times. Subsequently, the generated images were downloaded individually and incorporated back into the dataset. This workflow considerably extended the duration of the operation, spanning several days to complete, while repeating this operation 413 times.
The figure 5.10 resents illustrative outcomes of sampling with the DALL.E2 API across various grades

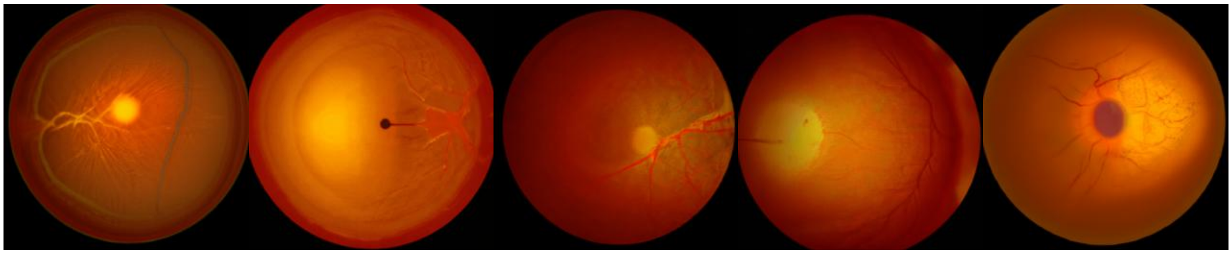

Figure 5.10: Sampling results of different grades (0-4, starting from the left)

4 Step 2: Training the classifiers and obtaining results

Considering the lack of achievement by the current State-Of-The-Art model, it is unjustifiable to employ the sampled images generated by the DCGAN generator. Hence, the subsequent results will entail a comparative analysis between the outcomes obtained by employing traditional data augmentation techniques and the outcomes obtained through augmentation using the Diffusion model.

Upon the successful completion of the data augmentation process, our training sets are enriched with a total of 1773 images, meticulously curated to ensure a well-balanced distribution across the five distinct grade classes. We conducted a comparative analysis between traditional and diffusion models for augmentation techniques to evaluate their effectiveness in medical image augmentation.

[6] https://platform.openai.com/docs/guides/images/introduction

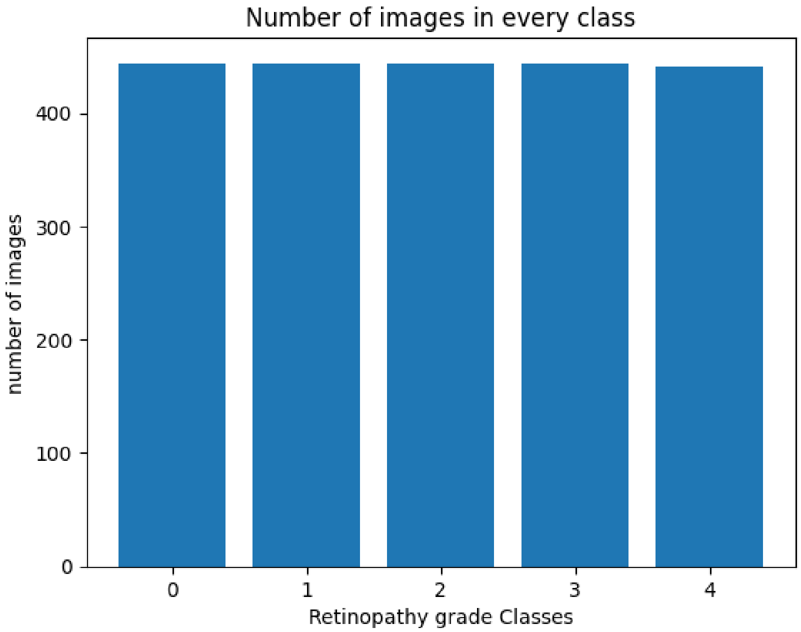

Figure 5.11: images distribution in the five grade classes after the augmentation process

The primary objective of our study is to comprehensively assess the realistic quality and validate the utilization of generated images from diffusion models as a data augmentation technique, specifically within the context of medical images which are known for their sensitive and accurate information. This objective is accomplished by training classifiers using augmented data obtained from both traditional techniques and deep-based techniques, specifically employing deep-based diffusion models, while ensuring the use of consistent parameters and conditions.
Through rigorous evaluation using classification reports, we seek to uncover any advantages or limitations associated with each technique, paving the way for improved diagnostic and prognostic capabilities in medical decision-making.

Note: Before feeding the datasets into the network, a preprocessing step of z-score normalization is applied. This technique involved standardizing the pixel values to have a mean of zero and a variance of one. Z-score normalization is a common practice in data preprocessing that ensures the data's distribution is centered and rescaled, facilitating more effective model training and performance evaluation in a standardized metric.

4.1 Classification results of our CNN model

In our study, we present a comparative analysis of traditional data augmentation and diffusion model augmentation techniques using the same training set which is augmented with an equal number of images with the same distribution across grade classes. Both techniques were evaluated using a convolutional neural network (CNN). In this process, the CNN is trained during 80 epochs and using a batch size of 25 for both datasets. Additionally, The model's weights are initialized using a specific seed value (0). The optimization is performed using the Adam optimizer, a widely-used optimization algorithm that combines adaptive learning rates with momentum, effectively accelerating the convergence process. The model's performance is evaluated using the cross-entropy loss, a common loss function utilized for classification tasks.

results

When employing traditional data augmentation, the grading results achieve an overall accuracy of 33.01% with a corresponding loss of 3.56. However, it should be noted that these results are obtained under the condition of network overfitting. This implies that the model performed well on the training data but may not generalize effectively to unseen testing data. On the other hand, when the CNN is trained using the augmented data with diffusion model, the grading results demonstrate an improved overall accuracy of 40.78% and a reduced loss of 2.82. Similarly, these results were attained under the condition of network overfitting.

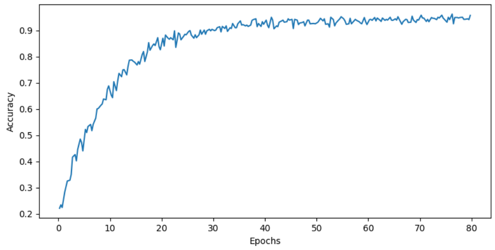

Figure 5.12: Accuracy and Loss behavior during the training in the traditional data augmentation

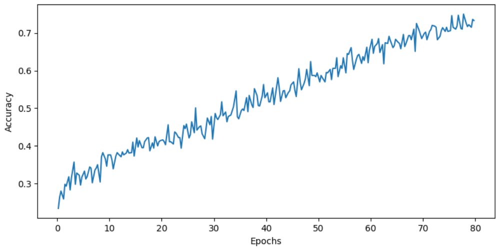

Figure 5.13: Accuracy and Loss behavior during the training in diffusion model data augmentation

The confusion matrix of the classifier for both data sets is as the following:

	G-0	G-1	G-2	G-3	G-4
G-0	1	1	15	4	3
G-1	2	0	2	0	1
G-2	3	1	17	3	8
G-3	3	0	6	5	5
G-4	4	0	4	4	1

Table 5.1: Confusion matrix of the CNN trained with augmented data with traditional technique

	G-0	G-1	G-2	G-3	G-4
G-0	18	0	14	1	1
G-1	2	0	2	1	0
G-2	6	1	19	5	1
G-3	3	2	5	5	4
G-4	1	2	9	1	0

Table 5.2: Confusion matrix of the CNN trained with augmented data using diffusion models

One notable finding is that the training process with traditional data augmentation exhibits a faster tendency to overfit the data compared to the diffusion model augmentation. This implies that the traditional approach is more susceptible to overfitting, whereby the model performed exceptionally well on the training data but struggled to generalize effectively to test data.

Conversely, the diffusion model augmentation demonstrates superior performance in terms of accuracy and loss and the confusion matrix, which provides additional information about the classification performance. The improvements observed in these metrics indicate that the diffusion model augmentation technique facilitates enhanced learning and generalization capabilities, contributing to a more effective and robust model compared with traditional augmentation.

Nevertheless, it is important to exercise caution when interpreting the improvements in accuracy and loss, as network overfitting was still present even with the diffusion model augmentation. Overfitting remains a challenge that needs to be addressed to ensure the model's ability to generalize well to unseen data.

4.2 Classification results of fine-tuned Resnet50

This research study involves a comparative analysis of traditional data augmentation techniques and augmentation using diffusion models. Two training sets are created, with one set augmented using geometric techniques (horizontal and vertical flipping) and photometric techniques (increased brightness), while the other set is augmented using diffusion models. Both augmented datasets are balanced in the number of included images and in terms of image distribution across different grade classes. We Fine-Tune a pre-trained ResNet50 model [2] by replacing the last fully connected layer with two new layers. The Resnet model is fine-tuned with a batch size of 25 over the course of 25 epochs. The newly added layer weights are initialized following the same distribution as Xavier's normal distribution. The optimization process employs the Adam optimizer, and the model's performance is assessed using the cross-entropy loss, a prevalent loss function commonly employed for classification tasks.

Our objective is to evaluate the impact of these augmentation strategies on classification accuracy and loss metrics. Through systematic evaluation, we aim to determine the effectiveness of each technique when integrate with the Fine-Tuned ResNet50 model.

RESULTS

When the pre-trained ResNet50 model is Fine-Tuned using the training set augmented with traditional techniques, the training achieves an overall accuracy of 41.74% with a corresponding loss of 2,11. In contrast, when the pre-trained ResNet50 model is Fine-Tuned using the training set augmented with diffusion techniques, the training achieves an overall accuracy of 53.40% with a corresponding loss of 1,36.

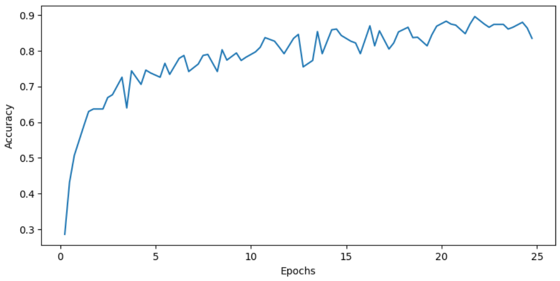

Figure 5.14: Accuracy and Loss behavior during the fine-tuning of the training augmented using traditional augmentation

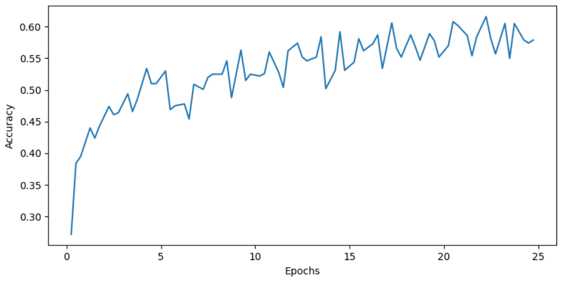

Figure 5.15: Accuracy and Loss behavior during the fine-tuning of the training augmented using diffusion model

With confusion matrices as the following:

	G-0	G-1	G-2	G-3	G-4
G-0	13	0	19	2	0
G-1	3	0	1	1	0
G-2	4	1	16	11	0
G-3	0	1	7	11	0
G-4	1	0	7	2	3

Table 5.3: Confusion matrix of the fine-tuned Resnet-50 with augmented data with traditional technique

	G-0	G-1	G-2	G-3	G-4
G-0	22	1	11	0	0
G-1	4	0	1	0	0
G-2	7	0	22	0	3
G-3	1	0	10	7	1
G-4	2	1	6	0	4

Table 5.4: Confusion matrix of the fine-tuned Resnet-50 with augmented data using diffusion models

It is crucial to highlight that the results obtained from the augmented training using traditional augmentation techniques are affected by the presence of network overfitting. Conversely, in the case of fine-tuning the ResNet50 model with the training augmented using the diffusion model, the network demonstrates resilience against overfitting to the training data.

These observations emphasize the contrasting effects of traditional and diffusion model augmentation on the generalization capability of the network. While traditional augmentation leads to overfitting, the diffusion model approach effectively mitigates this issue, allowing the network to generalize well to unseen data.

This finding underscores the potential of the diffusion model augmentation technique in enhancing the generalization performance of the fine-tuned ResNet50 model. It suggests that diffusion model augmentation can contribute to improved model robustness and performance in scenarios where overfitting is a concern.

5 Conclusion

In conclusion, due to the limited performance of our DCGAN generator, it was not suitable for utilization as an augmentation technique in the classification task. Consequently, we did not incorporate the generator as part of our experimentation process, and therefore, no results were obtained using this approach.

Thus, our study focused on evaluating the effectiveness of traditional data augmentation versus diffusion model augmentation in the context of medical image classification. Through rigorous analysis and comparison, we aimed to assess the impact of these augmentation strategies on classification outcomes, model generalization, and overfitting tendencies.

Our findings demonstrate that the diffusion model augmentation technique offers notable advantages over traditional data augmentation approaches. When applied to both a convolutional neural network (CNN) model and a fine-tuned ResNet50 model, the diffusion model augmentation consistently led to higher accuracy and lower loss values. This suggests that the diffusion model augmentation facilitated enhanced learning and improved generalization capabilities, contributing to more robust and effective models.

Importantly, our results highlight the challenge of overfitting in medical image classification tasks. While both augmentation techniques exhibited some degree of overfitting, the traditional data augmentation approach displayed a faster tendency to overfit compared to the diffusion model augmentation. This emphasizes the potential of diffusion models in mitigating overfitting and improving the generalization performance of the models.

Overall, the findings from this study emphasize the value of diffusion model augmentation in the realm of medical image classification. By enhancing accuracy, reducing overfitting tendencies, and improving generalization capabilities, diffusion models offer promising avenues for improving diagnostic and prognostic capabilities in medical decision-making.

General Conclusion

The findings of this study demonstrate that the utilization of a diffusion model for data augmentation presents notable advantages over traditional data augmentation approaches. Specifically, when applied to a deep-based classifier, the diffusion model augmentation exhibited superior performance. These results highlight the challenge of overfitting commonly encountered in medical image classification tasks. While both augmentation techniques displayed some degree of overfitting, the traditional data augmentation approach exhibited a faster tendency to overfit in comparison to the diffusion model augmentation. These findings underscore the significance of incorporating diffusion model augmentation in the realm of medical image classification.

The research objective of this study was to evaluate the quality of images generated by Diffusion models. This objective has been successfully addressed in this thesis through a comparative analysis of the performance exhibited by deep-based classifiers trained on two distinct datasets. One dataset was augmented using diffusion models, while the other dataset underwent traditional data augmentation techniques.

The findings of this study bear significance in the field of medical image analysis, particularly in the context of medical classification tasks. These findings effectively tackle the prevailing challenges associated with the scarcity of available medical datasets and the fatigued process of collecting and annotating medical images.

While this study has yielded valuable insights, it is crucial to acknowledge its limitations. One limitation pertains to the restricted number of generated images due to the time-consuming nature of the image generation process using the pretrained DALLE2 model. Additionally, the achieved accuracy of 53% indicates that further improvements are necessary to enhance the effectiveness of the classifiers. Furthermore, the low resolution of the images generated by the DCGAN model prevented their inclusion in the comparative analysis as they could not be fed to the classifier.

These limitations highlight areas for future research and emphasize the need for several improvements. These improvements encompass the generation of a larger number of images for the training set through the utilization of DALLE2, thereby aiming to achieve improved classification accuracy. Additionally, it is imperative to improve the DCGAN model to generate higher quality images. This enhancement will enable more advanced comparative analyses and facilitate a comprehensive assessment of image quality. Moreover, in order to provide a comprehensive evaluation, we will incorporate other data augmentation techniques alongside the aforementioned methods, allowing for a thorough comparison with diffusion models.

BIBLIOGRAPHY

[1] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[3] D. J. Park, M. W. Park, H. Lee, Y.-J. Kim, Y. Kim, and Y. H. Park, "Development of machine learning model for diagnostic disease prediction based on laboratory tests," *Scientific reports*, vol. 11, no. 1, p. 7567, 2021.

[4] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[5] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ digital medicine*, vol. 1, no. 1, p. 6, 2018.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang, *et al.*, "Annotation-efficient deep learning for automatic medical image segmentation," *Nature communications*, vol. 12, no. 1, p. 5915, 2021.

[8] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, p. 100258, 2022.

[9] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[10] M. Özbey, S. U. Dar, H. A. Bedel, O. Dalmaz, Ş. Özturk, A. Güngör, and T. Çukur, “Unsupervised medical image translation with adversarial diffusion models,” *arXiv preprint arXiv:2207.08208*, 2022.

[11] P. Rouzrokh, B. Khosravi, S. Faghani, M. Moassefi, S. Vahdati, and B. J. Erickson, “Multitask brain tumor inpainting with diffusion models: A methodological report,” *arXiv preprint arXiv:2210.12113*, 2022.

[12] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, “Diffusion models for medical anomaly detection,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pp. 35–45, Springer, 2022.

[13] J. Wu, H. Fang, Y. Zhang, Y. Yang, and Y. Xu, “Medsegdiff: Medical image segmentation with diffusion probabilistic model,” *arXiv preprint arXiv:2211.00611*, 2022.

[14] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, “Brain imaging generation with latent diffusion models,” in *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pp. 117–126, Springer, 2022.

[15] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, “Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 650–656, 2022.

[16] H. Ali, S. Murad, and Z. Shah, “Spot the fake lungs: Generating synthetic medical images using neural diffusion models,” in *Artificial Intelligence and Cognitive Science: 30th Irish Conference, AICS 2022, Munster, Ireland, December 8–9, 2022, Revised Selected Papers*, pp. 32–39, Springer, 2023.

[17] K. Packhäuser, L. Folle, F. Thamm, and A. Maier, “Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems,” *arXiv preprint arXiv:2211.01323*, 2022.

[18] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting pretrained vision-language foundational models to medical imaging domains," *arXiv preprint arXiv:2210.04133*, 2022.

[19] D. Hu, Y. K. Tao, and I. Oguz, "Unsupervised denoising of retinal oct with diffusion probabilistic model," in *Medical Imaging 2022: Image Processing*, vol. 12032, pp. 25–34, SPIE, 2022.

[20] L. W. Sagers, J. A. Diao, M. Groh, P. Rajpurkar, A. S. Adamson, and A. K. Manrai, "Improving dermatology classifiers across populations using images generated by large diffusion models," *arXiv preprint arXiv:2211.13352*, 2022.

[21] M. Akrout, B. Gyepesi, P. Holló, A. Poór, B. Kincső, S. Solis, K. Cirone, J. Kawahara, D. Slade, L. Abid, *et al.*, "Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images," *arXiv preprint arXiv:2301.04802*, 2023.

[22] Y. Chen and D. Zhang, "Integration of knowledge and data in machine learning," *arXiv preprint arXiv:2202.10337*, 2022.

[23] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[24] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics–challenges in topic discovery, data collection, and data preparation," *International journal of information management*, vol. 39, pp. 156–168, 2018.

[25] H. Chen, H. Takamura, and H. Nakayama, "Scixgen: A scientific paper dataset for context-aware text generation," *arXiv preprint arXiv:2110.10774*, 2021.

[26] E. A. Buchanan and E. E. Hvizdak, "Online survey tools: Ethical and methodological concerns of human research ethics committees," *Journal of empirical research on human research ethics*, vol. 4, no. 2, pp. 37–48, 2009.

[27] C. M. Cheung and D. R. Thadani, "The impact of electronic word-of-mouth communication: A literature analysis and integrative model," *Decision support systems*, vol. 54, no. 1, pp. 461–470, 2012.

[28] J. Driedger and M. Müller, “Extracting singing voice from music recordings by cascading audio decomposition techniques,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 126–130, 2015.

[29] R. H. Toczydlowski, “An efficient workflow for collecting, entering, and proofing field data: harnessing voice recording and dictation software,” *Bulletin of the Ecological Society of America*, vol. 98, no. 4, pp. 291–297, 2017.

[30] A. Kornilova, M. Faizullin, K. Pakulev, A. Sadkov, D. Kukushkin, A. Akhmetyanov, T. Akhtyamov, H. Taherinejad, and G. Ferrer, “Smartportraits: Depth powered handheld smartphone dataset of human portraits for state estimation, reconstruction and synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21318–21329, 2022.

[31] S. A. Jebur, K. A. Hussein, H. K. Hoomod, L. Alzubaidi, and J. Santamaría, “Review on deep learning approaches for anomaly event detection in video surveillance,” *Electronics*, vol. 12, no. 1, p. 29, 2022.

[32] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “Youtube-8m: A large-scale video classification benchmark,” *arXiv preprint arXiv:1609.08675*, 2016.

[33] L. Chan, C.-H. Hsieh, Y.-L. Chen, S. Yang, D.-Y. Huang, R.-H. Liang, and B.-Y. Chen, “Cyclops: Wearable and single-piece full-body gesture input devices,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3001–3009, 2015.

[34] K. Prabhakar, V. Vinod, N. Sahoo, and V. B. Radhakrishnan, “Few-shot domain adaptation for low light raw image enhancement,” *British Machine Vision Conference*, 2021.

[35] F. Cai, W. Lu, W. Shi, and S. He, “A mobile device-based imaging spectrometer for environmental monitoring by attaching a lightweight small module to a commercial digital camera,” *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[36] E. Sheehan, B. Uzkent, C. Meng, Z. Tang, M. Burke, D. Lobell, and S. Ermon, “Learning to interpret satellite images using wikipedia,” *arXiv preprint arXiv:1809.10236*, 2018.

[37] A. Datta, Z. Zhong, and S. Motakef, “A new generation of direct x-ray detectors for medical and synchrotron imaging applications,” *Scientific reports*, vol. 10, no. 1, p. 20097, 2020.

[38] S. C. Deoni, P. Medeiros, A. T. Deoni, P. Burton, J. Beauchemin, V. D'Sa, E. Boskamp, S. By, C. McNulty, W. Mileski, *et al.*, "Development of a mobile low-field mri scanner," *Scientific reports*, vol. 12, no. 1, p. 5690, 2022.

[39] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, p. 5, 2021.

[40] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, pp. 1–13, 2018.

[41] J. Li, G. Zhu, C. Hua, M. Feng, P. Li, X. Lu, J. Song, P. Shen, X. Xu, L. Mei, *et al.*, "A systematic collection of medical image datasets for deep learning," *arXiv preprint arXiv:2106.12864*, 2021.

[42] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[43] H. Harvey and B. Glocker, "A standardised approach for preparing imaging data for machine learning tasks in radiology," *Artificial intelligence in medical imaging: opportunities, applications and risks*, pp. 61–72, 2019.

[44] A. Yala, T. Schuster, R. Miles, R. Barzilay, and C. Lehman, "A deep learning model to triage screening mammograms: a simulation study," *Radiology*, vol. 293, no. 1, pp. 38–46, 2019.

[45] P. Lee, H. Wei, A. N. Pouliopoulos, B. T. Forsyth, Y. Yang, C. Zhang, A. F. Laine, E. E. Konofagou, C. Wu, and J. Guo, "Deep learning enables reduced gadolinium dose for contrast-enhanced blood-brain barrier opening," *arXiv preprint arXiv:2301.07248*, 2023.

[46] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose ct," *IEEE transactions on medical imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[47] E. J. Hwang, S. Park, K.-N. Jin, J. Im Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, J. G. Cohen, *et al.*, "Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs," *JAMA network open*, vol. 2, no. 3, pp. e191095–e191095, 2019.

[48] M. R. Arbabshirani, B. K. Fornwalt, G. J. Mongelluzzo, J. D. Suever, B. D. Geise, A. A. Patel, and G. J. Moore, "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration," *NPJ digital medicine*, vol. 1, no. 1, p. 9, 2018.

[49] H. Fu, Y. Xu, D. W. K. Wong, and J. Liu, "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pp. 698–701, IEEE, 2016.

[50] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative ophthalmology & visual science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[51] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. Mari Aparici, *et al.*, "A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.

[52] A. Parakh, H. Lee, J. H. Lee, B. H. Eisner, D. V. Sahani, and S. Do, "Urinary stone detection on ct images using deep convolutional neural networks: evaluation of model performance and generalization," *Radiology: Artificial Intelligence*, vol. 1, no. 4, p. e180066, 2019.

[53] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, 2018.

[54] B. D. de Vos, J. M. Wolterink, T. Leiner, P. A. de Jong, N. Lessmann, and I. Išgum, "Direct automatic coronary calcium scoring in cardiac and chest ct," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2127–2138, 2019.

[55] P. Schelb, S. Kohl, J. P. Radtke, M. Wiesenfarth, P. Kickingereder, S. Bickelhaupt, T. A. Kuder, A. Stenzinger, M. Hohenfellner, H.-P. Schlemmer, *et al.*, "Classification of cancer at prostate mri: deep learning versus clinical pi-rads assessment," *Radiology*, vol. 293, no. 3, pp. 607–617, 2019.

[56] T. A. Ngo, *Medical Image Segmentation Combining Level Set Method and Deep Belief Networks*. PhD thesis, 2015.

[57] I. Badash, K. Burtt, C. A. Solorzano, and J. N. Carey, “Innovations in surgery simulation: a review of past, current and future techniques,” *Annals of translational medicine*, vol. 4, no. 23, 2016.

[58] P.-H. C. Chen, Y. Liu, and L. Peng, “How to develop machine learning models for healthcare,” *Nature materials*, vol. 18, no. 5, pp. 410–414, 2019.

[59] M. Abedi, L. Hempel, S. Sadeghi, and T. Kirsten, “Gan-based approaches for generating structured data in the medical domain,” *Applied Sciences*, vol. 12, no. 14, p. 7075, 2022.

[60] E. Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Youth Opportunities Initiative*. European Commission Brussels, Belgium, 2011.

[61] M. Iima, R. Sakamoto, T. Kakigi, A. Yamamoto, B. Otsuki, Y. Nakamoto, J. Toguchida, and S. Matsuda, “The efficacy of ct temporal subtraction images for fibrodysplasia ossificans progressiva,” *Tomography*, vol. 9, no. 2, pp. 768–775, 2023.

[62] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, “A review of medical image data augmentation techniques for deep learning applications,” *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[63] M. Vallières, C. R. Freeman, S. R. Skamene, and I. E. Naqa, “A radiomics model from joint fdg-pet and mri texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities,” *Physics in Medicine & Biology*, vol. 60, p. 5471, jun 2015.

[64] S. H. Hasanpour, M. Rouhani, M. Fayyaz, and M. Sabokrou, “Lets keep it simple, using simple architectures to outperform deeper and more complex architectures,” *arXiv preprint arXiv:1608.06037*, 2016.

[65] N. E. Khalifa, M. Loey, and S. Mirjalili, “A comprehensive survey of recent trends in deep learning for digital images augmentation,” *Artificial Intelligence Review*, pp. 1–27, 2022.

[66] Q. Wang, F. Meng, and T. P. Breckon, “Data augmentation with norm-vae for unsupervised domain adaptation,” *arXiv preprint arXiv:2012.00848*, 2020.

[67] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, “Effective data augmentation with diffusion models,” *arXiv preprint arXiv:2302.07944*, 2023.

[68] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, “Bagan: Data augmentation with balancing gan,” *arXiv preprint arXiv:1803.09655*, 2018.

[69] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[70] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” *arXiv preprint arXiv:1712.04621*, 2017.

[71] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

[72] P. Y. Simard, D. Steinkraus, J. C. Platt, *et al.*, “Best practices for convolutional neural networks applied to visual document analysis.,” in *Icdar*, vol. 3, Edinburgh, 2003.

[73] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[74] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *In Advances in Neural Information Processing Systems*, no. 11, pp. 1–9, 2014.

[75] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International Conference on Machine Learning*, pp. 2256–2265, PMLR, 2015.

[76] Z. Wu, S. Wang, Y. Qian, and K. Yu, “Data augmentation using variational autoencoder for embedding based speaker verification.,” in *INTERSPEECH*, pp. 1163–1167, 2019.

[77] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, “Towards visually explaining variational autoencoders,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8642–8651, 2020.

[78] H. Shao, S. Yao, D. Sun, A. Zhang, S. Liu, D. Liu, J. Wang, and T. Abdelzaher, “Controlvae: Controllable variational autoencoder,” in *International Conference on Machine Learning*, pp. 8655–8664, PMLR, 2020.

[79] M. Wiatrak, S. V. Albrecht, and A. Nystrom, “Stabilizing generative adversarial networks: A survey,” *arXiv preprint arXiv:1910.00927*, 2019.

[80] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models for medical image analysis: A comprehensive survey," *arXiv preprint arXiv:2211.07804*, 2022.

[81] C. Jarzynski, "Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach," *Physical Review E*, vol. 56, no. 5, p. 5018, 1997.

[82] R. M. Neal, "Annealed importance sampling," *Statistics and computing*, vol. 11, pp. 125–139, 2001.

[83] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[84] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[85] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[86] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[87] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching.," *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[88] M. Raphan and E. Simoncelli, "Learning to be bayesian without supervision," *Advances in neural information processing systems*, vol. 19, 2006.

[89] M. Raphan and E. P. Simoncelli, "Least squares estimation without priors or supervision," *Neural computation*, vol. 23, no. 2, pp. 374–420, 2011.

[90] Y. Song, S. Garg, J. Shi, and S. Ermon, "Sliced score matching: A scalable approach to density and score estimation," in *Uncertainty in Artificial Intelligence*, pp. 574–584, PMLR, 2020.

[91] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[92] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman, and I. Mitliagkas, "Gotta go fast when generating data with score-based models," *arXiv preprint arXiv:2105.14080*, 2021.

[93] T. Karras, M. Aittala, T. Aila, and S. Laine, “Elucidating the design space of diffusion-based generative models,” *arXiv preprint arXiv:2206.00364*, 2022.

[94] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, “Cvae-gan: fine-grained image generation through asymmetric training,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2745–2754, 2017.

[95] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, 2022.

[96] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” *arXiv preprint arXiv:2009.09761*, 2020.

[97] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, “Diffusion-lm improves controllable text generation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 4328–4343, 2022.

[98] H. Chung, E. S. Lee, and J. C. Ye, “Mr image denoising and super-resolution using regularized reverse diffusion,” *IEEE Transactions on Medical Imaging*, 2022.

[99] W. Xia, Q. Lyu, and G. Wang, “Low-dose ct using denoising diffusion probabilistic model for 20xtimes speedup,” *arXiv preprint arXiv:2209.15136*, 2022.

[100] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 590–597, 2019.

[101] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, “Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports,” *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[102] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[103] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

[104] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.

[105] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12413–12422, 2022.

[106] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, *et al.*, "fastmri: An open dataset and benchmarks for accelerated mri," *arXiv preprint arXiv:1811.08839*, 2018.

[107] H. Chung and J. C. Ye, "Score-based diffusion models for accelerated mri," *Medical Image Analysis*, vol. 80, p. 102479, 2022.

[108] I. Oguz, J. D. Malone, Y. Atay, and Y. K. Tao, "Self-fusion for oct noise reduction," in *Medical Imaging 2020: Image Processing*, vol. 11313, pp. 45–50, SPIE, 2020.

[109] D. Hu, C. Cui, H. Li, K. E. Larson, Y. K. Tao, and I. Oguz, "Life: a generalizable autodidactic pipeline for 3d oct-a vessel segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pp. 514–524, Springer, 2021.

[110] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *arXiv preprint arXiv:2206.00927*, 2022.

[111] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang, "Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 191–200, 2019.

[112] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders," *arXiv preprint arXiv:1806.04972*, 2018.

[113] B. Puccio, J. P. Pooley, J. S. Pellman, E. C. Taverna, and R. C. Craddock, "The preprocessed connectomes project repository of manually corrected skull-stripped t1-weighted anatomical mri data," *Gigascience*, vol. 5, no. 1, pp. s13742–016, 2016.

[114] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[115] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[116] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022.

[117] B. Harangi, J. Toth, A. Baran, and A. Hajdu, "Automatic screening of fundus images using a combination of convolutional neural network and hand-crafted features," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2699–2702, IEEE, 2019.

[118] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[119] D. Yang, C. Martinez, L. Visuña, H. Khandhar, C. Bhatt, and J. Carretero, "Detection and analysis of covid-19 in medical images using deep learning techniques," *Scientific Reports*, vol. 11, no. 1, p. 19638, 2021.

[120] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, 2018.

République Algérienne Démocratique Et Populaire

Université Kasdi Merbah- Ouargla
Faculté des Nouvelles Technologies
de l'Information et de la Communication
Département d' Informatique et technologie de l'information

AUTORISATION DE SOUTENANCE Master II
Année universitaire : 2022/2023

Encadreur :
Nom : Bouanane
Prénom : Khadra

Candidats :
Nom /Prénom : Lakas Badia Wissem
Nom / Prénom : Meddour Bouthayna

Spécialité : Intelligence Artificielle et Sciences de données.

Titre du mémoire :
Diffusion Models for Data Augmentation of Medical Images.

Ouargla le : 13/06/2023
Signature