الجمهورية الجزَائِرية الديمقرَاطية الشعبية
PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

وزَارة التعليم العَالي و البحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

جَامعة عمَار ثليجي بِالأَغوَاط
AMAR TELIDJI UNIVERSITY OF LAGHOUAT

كلية العلوم
FACULTY OF SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

Master's Thesis

Field : Mathematics and Computer Science
Department : Computer Sciences
Option : Data Sience and Artificial Intelligence

By: Ikram ELHOUITI & Amina HABIB

TOPIC

Exploring GANs and Diffusing Models for Medical Image Data Augmentation

Defended publicly on July 9th, 2025, before a jury composed of:

Pr. Hadda CHERROUN	Prof	President
Dr. Atika SAHLAOUI	M.A.A	Examiner
Dr. Saida Sarra BOUDOUH	M.A.B	Examiner
Dr. Younes GUELLOUMA	M.C.A	Supervisor

Thesis No. - Academic Year 2024/2025

Acknowledgments

All glory is due to Almighty Allah, upon whom we rely for strength and endurance. We thank Him for granting us the health, patience, and willpower to undertake and accomplish this work.

In the foremost place, our heartfelt thanks and gratitude go to our supervisor, ***Prof. Younes GUELLOUMA****, for his continuous support, valuable guidance, and insightful feedback throughout this research. From the very beginning, his mentorship has been instrumental in shaping our academic path. His commitment to excellence and generosity in sharing his knowledge greatly enriched our learning and contributed significantly to the scientific quality of this work.*

We are also honored to thank the respected members of the thesis committee for their presence and for evaluating our work with care and attention.

A special word of appreciation goes to ***Prof. Chaker Abdelaziz KERRACHE****, for his availability, encouragement, and kindness, which had a meaningful impact on both our progress and our experience as students.*

Our sincere thanks go as well to the Department of Computer Science for the academic support and resources provided during our research.

We would like to express our deepest appreciation to all those who supported and contributed to the completion of this Master's thesis. This research journey has been a profoundly enriching and rewarding experience, marked by valuable challenges, growth, and learning.

Dedications

First and foremost, I thank Allah, the Almighty, for it is by His grace and guidance that I was able to accomplish this work. None of this would have been possible without His blessings. Every step, every moment, and every achievement happened by His will.

To my dear ***mother****, whose unconditional love, tireless sacrifices, and endless prayers have always been my greatest support. Your strength, patience, and belief in me have carried me through the hardest moments. I owe you everything.*

To my beloved ***father****, whose wisdom, encouragement, and constant presence have shaped my determination and discipline. Thank you for always pushing me to be my best and for standing proudly by my side.*

To my brothers ***Abdelkader****,* ***Mohamed Mahdi*** *and younger sister* ***Nour El Houda****, whose smiles, support, and love brought light to even my most stressful days. You are all my strength and my joy.*

To my cousin ***Doha Merad****, who is like a twin sister to me, thank you for always motivating me, being there for me, and encouraging me more than words can express.*

To ***Ikram Elhouiti****, my sister, my bestie, my study partner, and my rock throughout this entire thesis. You have been by my side in the most difficult situations, always ready to help without hesitation. Your words lifted me, your belief in me strengthened me, and your support never wavered. Thank you for everything from the long nights of hard work to the moments of laughter that made it all bearable.*

To the memory of my dear grandparents ***Mohamed Merad****,* ***Mubaraka Fatmi****,* ***Abdelkader Habib****,* ***Fatima Fatmi****, all of whom have passed away but continue to live in my heart. Their love, values, and legacy have deeply influenced the person I have become. I carry their memory with pride and gratitude.*

To my entire family, ***Habib*** *and* ***Merad****, thank you for your constant love, support, and belief in me.*

And finally, I dedicate this work to myself, to the version of me that never gave up, even when things were overwhelming. To the sleepless nights, the quiet perseverance, the self-doubt, and the strength I found within. This journey has shaped me, tested me, and taught me and I am proud of the person I have become.

Amina HABIB

Dedications

To my loving parents ***Tahar EL HOUITI*** *and* ***OumElkheir EL HOUITI****, who have supported me from day one until this very moment, your love, care and prayers have always been my pillars of strength, my comfort, and my greatest source of motivation. No matter what, your smiles warm my heart, and I am truly thankful and honored to have you as my parents. Alhamdulillah.*

To my beloved siblings ***Yasmine*** *and* ***Anes****, my constant companions and confidants thank you for always cheering me up, inspiring me, and sharing life's joys and challenges with me.*

To my dear grandparents ***Bebah****,* ***Ma****,* ***Papa Allah yarehmo*** *and* ***Ma Allah yarhamha*** *you mean so much to me. I am deeply grateful for your du'a, love, and wise guidance.*

To my life friend and sister ***Messaouda KECIBA****, through life's ups and downs, you have been a steady source of support, laughter, and inspiration. I am truly lucky to have someone like you.*

To my sister and dear friend ***Amina HABIB****, who prepared this work with me, thank you for being my companion through every sleepless night and every moment of doubt. Your presence has made this journey bearable and memorable.*

To all my ***uncles****,* ***aunts****, and* ***cousins*** *especially* ***Marouane****, thank you for your endless support, love, and encouragement.*

And finally, to the person I was, the person I am, and the person I'm becoming, to myself, for showing resilience, courage, and consistency in the face of every challenge. To the quiet nights, the countless revisions, and the tired eyes: you have come a long way ***Ikram****, Alhamdulillah for everything and always.*

Ikram EL HOUITI

Abstract

The scarcity of medical images remains a major challenge in the development of accurate and reliable deep learning models for medical image analysis.
This thesis explores the use of generative models specifically Generative Adversarial Networks (GANs) and Diffusion Models to synthesize realistic chest X-ray images for data augmentation. Different generative models were implemented and evaluated such as DCGAN, WGAN-GP, ProGAN, StyleGAN3, DDPM, and DDIM. All models were trained to generate grayscale chest X-ray images at a resolution of 64×64 pixels. The quality of the generated images was assessed using both quantitative metrics, such as the Fréchet Inception Distance (FID), and qualitative evaluation through a Visual Turing Test (VTT) conducted by a radiologist.
Results showed that diffusion models, particularly DDIM and DDPM, produced the most realistic images, in the normal and pneumonia classes. Interestingly, DCGAN showed unexpectedly strong performance when generating Normal chest X-ray images.
The findings demonstrate that generative models, especially diffusion-based models can effectively contribute to expanding medical datasets and improving deep learning performance, although computational limitations and resolution constraints remain challenges.

Keywords: Medical Imaging, GANs, Diffusion Models, Data Augmentation, Image Analysis, Artificial Intelligence, Generative models, Deep Learning, Image synthesis, Chest X-Ray.

الملخص

لا تزال ندرة الصور الطبية تمثل تحديا كبيرا في تطوير نماذج التعلم العميق الدقيقة والموثوقة لتحليل الصور الطبية. تستعرض هذه المذكرة استخدام نماذج التوليد، وبشكل خاص الشبكات التوليدية التنافسية (GANs) ونماذج الانتشار (Diffusion Models)، لتوليد صور أشعة سينية للصدر واقعية تُستخدم في تعزيز البيانات. تم تنفيذ وتقييم عدة نماذج توليدية مثل DCGAN، WGAN-GP، ProGAN، StyleGAN3، DDPM، DDIM. تم تدريب جميع النماذج على توليد صور أشعة سينية بدرجة وضوح 64×64 بكسل. تم تقييم جودة الصور المُولَّدة باستخدام مقاييس كمية مثل (FID)، بالإضافة إلى تقييم نوعي من خلال اختبار بصري (VTT) الذي أُجري بواسطة طبيب أشعة. أظهرت النتائج أن نماذج الانتشار، وخصوصاً DDIM وDDPM، أنتجت صوراً أكثر واقعية في صنفي الصور العادية والمصابة بالالتهاب الرئوي. ومن المثير للاهتمام أن نموذج DCGAN أظهر أداءً قوياً وغير متوقع عند توليد صور الأشعة السينية العادية. توضح هذه النتائج أن نماذج التوليد، خاصةً تلك المعتمدة على الانتشار، يمكن أن تساهم بشكل فعال في توسيع مجموعات البيانات الطبية وتحسين أداء نماذج التعلم العميق، على الرغم من وجود بعض القيود المتعلقة بالقدرات الحاسوبية ودقة الصور.

الكلمات المفتاحية: التصوير الطبي، الشبكات التوليدية التنافسية (GANs)، نماذج الانتشار، تعزيز البيانات، تحليل الصور، الذكاء الاصطناعي، النماذج التوليدية، التعلم العميق، توليد الصور، أشعة الصدر السينية.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AI Artificial Intelligence. 1, 3, 4, 5, 40

ANNs Artificial Neural Networks. 6

CGAN Conditional Generative Adversarial Networks. 14, 15

CNNs Convolutional Neural Networks. 1, 6, 14, 19, 41

CT Computed Tomography. 1, 5, 15, 36, 41

CXR Chest X-Ray. 3, 4, 13, 15, 18, 19

DCGAN Deep Convolutional Generative Adversarial Networks. 1, 2, 14, 15, 16, 17, 27, 29, 30, 32, 33, 34, 35, 36, 40

DDIMs Denoising Diffusion Implicit Models. 21, 23, 24, 27, 29, 30, 32

DDPMs Denoising Diffusion Probabilistic Models. 21, 22, 23, 27, 29, 30, 31, 32

DL Deep Learning. 3, 6

DMs Diffusion Models. 3, 10, 21, 29

FID Fréchet Inception Distance. 1, 15, 16, 17, 19, 27, 29, 30, 35, 36, 37, 40, 41

GANs Generative Adversarial Networks. 1, 3, 5, 8, 9, 10, 13, 14, 15, 16, 17, 19, 27, 29, 33, 36, 40, 41

LeakyReLU Leaky Rectified Linear Unit. 15, 30

ML Machine Learning. 3, 5

MRI Magnetic Resonance Imaging. 5, 41

NLP Natural Language Processing. 6

PET Positron Emission Tomography. 5

ProGAN Progressive Growing Generative Adversarial Networks. 1, 14, 17, 18, 23, 27, 29, 30, 31, 38, 40, 41

ReLU Rectified Linear Unit. 7, 15

StyleGAN2 Style-Based Generative Adversarial Network 2. 19, 20

StyleGAN3 Style-Based Generative Adversarial Network 3. 1, 19, 20, 27, 29, 30, 31, 38, 40, 41

StyleGAN3-R StyleGAN3-Rotation-equivariant. 19

StyleGAN3-T StyleGAN3-Translation-equivariant. 19

VAEs Variational Autoencoders. 9

VTT Visual Turing Test. 1, 15, 16, 19, 27, 29, 30, 37, 38, 40, 41

WGAN Wasserstein Generative Adversarial Network. 15, 16, 31

WGAN-GP Wasserstein GAN Gradient Penalty. 1, 14, 16, 17, 18, 27, 29, 30, 31, 32, 33, 35, 40

CHAPTER 1

INTRODUCTION

Contents

1.1 Problematic and motivation

Medical Image Analysis is the process of applying computer vision, artificial intelligence (AI), and image processing techniques to interpret medical images such as X-rays, MRIs, CT scans, and ultrasounds. This field plays a critical role in supporting diagnosis, treatment planning, and medical research. However, one of the primary challenges in medical image analysis is the limited availability of datasets. This scarcity is often due to patient privacy regulations, the high cost of expert labeling, and the rarity of certain medical conditions, making it difficult to collect large, diverse datasets necessary for training deep learning models.
Deep learning approaches, particularly convolutional neural networks (CNNs), have shown impressive performance in various medical imaging tasks. Yet, these models typically require large volumes of high-quality data to achieve robust and generalizable results. To overcome this limitation, data augmentation techniques are widely used. Traditional augmentation methods, such as flipping, rotation, cropping, and brightness adjustment help improve dataset diversity but are inherently limited in their ability to produce truly novel samples.
In response, more advanced augmentation methods using Generative Adversarial Networks (GANs) and Diffusion Models have emerged. These generative models are capable of synthesizing realistic medical images, thereby effectively expanding dataset size and variability. This study explores the use of such models to generate synthetic chest X-ray images, aiming to address data scarcity in a clinically meaningful way.
In this thesis, we implemented and compared several generative models, including DCGAN, WGAN-GP, ProGAN, StyleGAN3, DDPM, and DDIM, to evaluate their ability to generate high-quality chest X-ray images at a resolution of 64×64 pixels. We assessed the results using both quantitative metrics, such as the Fréchet Inception Distance (FID), and qualitative evaluation through the Visual Turing Test (VTT), which was conducted by a radiologist. Despite the resolution limitations and computational constraints, models like DDPM, DDIM and

DCGAN demonstrated competitive visual performance, with DCGAN surprisingly matching diffusion-based models in some Normal class cases.
The findings in this thesis contribute to the growing field of synthetic data for medical imaging and highlight promising directions for future work involving more complex models, higher resolutions, and expanded evaluation frameworks.

1.2 Organization of the thesis

The remainder of this thesis is organized as follows:

- **Chapter 2: Background**
  Introduces the fundamental concepts needed to understand the methodologies explored in this thesis.

- **Chapter 3: Related Work**
  Reviews existing literature on medical image synthesis, with special emphasis on GAN- and diffusion-based methods for chest X-ray generation. We begin with foundational GAN variants (DCGAN, WGAN-GP, ProGAN, and StyleGAN3), then we will explore the foundational variants of diffusion models (DDPM, DDIM).

- **Chapter 4: Our Contribution**
  Details of our contribution, we describe network architectures, training hyperparameters, our evaluation protocol, used datasets, preprocessing steps (resizing, normalization), and implementation details (TensorFlow versions, hardware), followed by results and discussions.

- **Chapter 5: Conclusion And Future Perspectives**
  Summarizes our key findings, discuss limitations, and propose directions for future research.

CHAPTER 2

BACKGROUND

Contents

2.1 Introduction

This chapter provides a comprehensive overview of the foundational concepts essential for understanding the methodologies explored in this thesis.
We begin by discussing medical images and their critical role in clinical diagnostics, with a specific emphasis on chest X-ray (CXR) images, which are the primary domain of our study. This leads into an overview of medical image analysis and exploring the concept of synthetic data, which is increasingly used to overcome challenges related to data privacy and scarcity in medical imaging. Building on this, we introduce the broader field of Artificial Intelligence (AI), narrowing our focus to Machine Learning (ML) and Deep Learning (DL). We then discuss data augmentation, emphasizing its role in enhancing training data diversity, particularly in medical applications. This sets the stage for our focus on Generative Artificial Intelligence, which aims to create realistic data samples. Finally, we delve into the two main types of generative models used in this thesis: Generative Adversarial Networks (GANs) and Diffusion Models (DMs). These advanced techniques form the backbone of our approach to generating high-quality synthetic chest radiograph images for research and diagnostic support.

2.2 Terminology and Concepts

This section defines the key concepts and terminologies that form the theoretical basis of this thesis. By establishing clear and precise definitions, we aim to ensure a consistent understanding of the technical language used throughout the work.

2.2.1 Medical Images

A medical image is a visual representation that captures anatomical or functional information of the human body in vivo, including organs, tissues, or cells, as well as data derived from these visuals. These images are acquired using appropriate imaging technologies to address specific clinical questions or medical needs identified by healthcare professionals [8]. The use of medical images is essential because they offer detailed insights into the anatomical and physiological state of patients, guiding clinical decisions and personalized treatment plans. For example, chest X-rays are widely used to detect lung diseases like tuberculosis, while MRI scans provide high-resolution images of soft tissues for neurological or musculoskeletal assessments [9]. However, training deep learning models on medical images often faces challenges due to limited data availability, class imbalance, and privacy concerns. To address these issues, data augmentation techniques are employed to artificially increase the size and diversity of medical image datasets [10]. In summary medical images are crucial for effective clinical care, and augmenting them helps build stronger, more accurate AI systems in the medical field. It exist some well-known chest X-Ray datasets widely used in medical imaging research such as NIH Chest X-ray Dataset [11], and Chest X-Ray Images (Pneumonia) dataset [12].

2.2.1.1 Chest X-Ray

The discovery of X-Rays began in 1895 by Wilhelm Roentgen [13]. Chest X-Ray (CXR) is a type of X-Ray image of the chest area, which provides images of the lungs, heart, airways, bones, and blood vessels, as it is shown in Fig 2.1. Within the same image, we can observe examples of a normal chest X-ray and one affected by pneumonia. CXR helps in checking for fractures, fluids or air around or in the lungs, infections, enabling precise diagnostics [14].

a. Normal Chest X-Ray

b. Pneumonia Chest X-Ray

Figure 2.1: Chest X-Ray Images.

2.2.1.2 Pneumonia

Pneumonia is a common and serious respiratory disease, typically caused by an infection from bacteria, viruses, or fungi. Symptoms often include cough, fever, chest pain, and difficulty breathing. While it can be mild in some cases, it may become very serious or even life-threatening, especially in infants, older adults, or individuals with weakened immune systems.
Pneumonia on a chest X-ray usually shows up as white or cloudy areas in the lungs. These areas form when the infection leads the tiny air sacs in the lungs (called alveoli) to fill with

fluids or pus rather than air, making it harder to breathe and indicating that the lungs are inflamed. The white patches can be in one part of the lung or spread out, depending on how serious the infection is [15, 16, 17].

2.2.1.3 Medical Imaging Analysis

Medical imaging analysis is the process of interpreting medical images to aid in diagnosis, disease detection, treatment planning, etc., using computational techniques, methods, and tools. The images can be in 2D or 3D forms, such as X-ray, MRI, CT, ultrasound, and PET scans. It combines medical knowledge with technological innovation, facilitating decision-making to optimize treatment strategies and elevating healthcare quality [18, 19].

2.2.1.4 Synthetic Data

Synthetic data refers to the data that is deliberately created using computational operations, algorithmic modeling, data simulation, or advanced machine learning techniques employing Generative techniques such as Generative Adversarial Networks (GANs) GANs, specifically to replicate the statistical properties and structural characteristics of real datasets without relying on real-world observations or individually identifiable information [20].

2.2.2 Artificial Intelligence

Artificial Intelligence (AI), is a field of computer science, the science and engineering of creating systems or machines capable of performing tasks that normally require human intelligence. It was originally introduced by John McCarthy in 1956. AI encompasses a wide variety of technologies, including machine learning, deep learning, natural language processing, robotics, as shown in Fig. 2.2 [21, 22, 23].

2.2.2.1 Machine Learning

Machine Learning (ML) is a subfield of artificial intelligence that is centered on creating algorithms and models that enable computers to learn from the data they process without the need for explicit programming. It involves developing algorithms that can identify patterns, make decisions, and predict outcomes based on data [24]. It includes different types, such as supervised learning, where the model is trained on labeled data, with each input paired with its corresponding output (label). The model learns by making predictions and refining itself based on the difference between its predicted outputs and the actual labels. It is frequently applied to tasks such as image classification, spam detection, and regression analysis [25]. Another approach is unsupervised learning, where the model is trained on unlabeled data, where it is left to discover patterns, structures, or relationships in the data on its own, it does not require labeled output data. It is commonly applied to tasks such as clustering, anomaly detection, and dimensionality reduction [25]. Also, semi-supervised learning is another machine learning type that combines both labeled and unlabeled data during training. Typically, a small portion of the data is labeled, while the majority is unlabeled [24]. Additionally, reinforcement learning, where the models learn through trial and error with feedback [24].

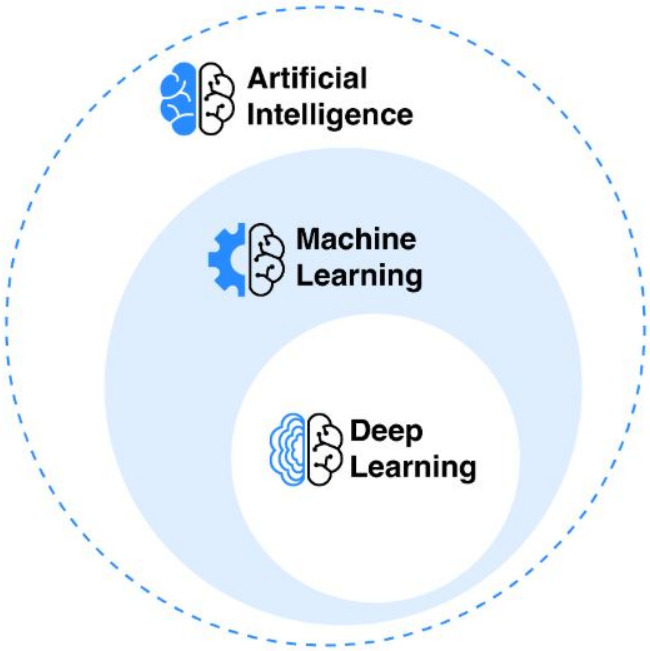


Figure 2.2: How Artificial Intelligence, Machine Learning, and Deep Learning Fit Together [1].

2.2.2.2 Deep Learning

Deep Learning (DL) is a branch of machine learning that uses deep neural networks of three or more layers (Input layer, Hidden layer, Output layer). It has transformed the field of artificial intelligence by empowering models, especially Artificial Neural Networks (ANNs) [26, 27], to learn from massive datasets and achieve cutting-edge performance across diverse domains such as computer vision, Natural Language Processing (NLP), and medical imaging [28].

2.2.2.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a specialized type of Artificial Neural Network (ANNs), primarily designed for processing data with a grid-like structure, such as images. CNNs have revolutionized computer vision tasks like image classification, object detection, and segmentation due to their ability to automatically learn hierarchical feature representations [29, 30].
CNNs, as shown in Fig. 2.3, operate using a series of specialized layers that extract and learn visual features from input images:

- **Input Layer:** Receives raw image data, usually represented as a matrix with dimensions (height × width × channels), no computations are done here, this layer only passes the data.
- **Convolutional Layers:** Apply small, learnable filters (kernels / Feature Detector / Feature Extractor) to the image, every single kernel will use a specific feature of the image. These filters scan the image in a sliding window fashion, detecting features like edges, textures, and shapes, producing feature maps. The convolution operation is mathematically

defined as:

$$O(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i+m, j+n) \cdot K(m,n) \tag{2.1}$$

where I is the input image, K is the kernel, and $O(i,j)$ is the output feature map.

- **Activation Function:** A non-linear function is applied to introduce non-linearity to a layer. This ensures that the model can learn complex patterns and then improve generalization capabilities. One of the most used activation functions is ReLU. Indeed, it replaces all the negative values of a layer's with zeros. The ReLU function is given for a layer's linear output z by:

$$ReLU(z) = \max(0, z) \tag{2.2}$$

- **Pooling Layers (Downsampling):** Reduce the spatial size of the feature maps, making the model faster and more efficient. Common pooling methods include Max Pooling (selecting the largest value) and Average Pooling (taking the average value).

- **Fully Connected Layers (Dense Layers):** Flatten the feature maps into a 1D vector and connect them to one or more dense layers. This part acts as a traditional neural network, which is expressed as:

$$y = \sigma(Wx + b) \tag{2.3}$$

  for each one of the defined dense layers, where x is the flattened input, W is the weight matrix, b is the bias, and σ is the activation function.

- **Output Layer:** Produces the final class probabilities (for classification) or a value (for regression), depending on the task. [29, 30].

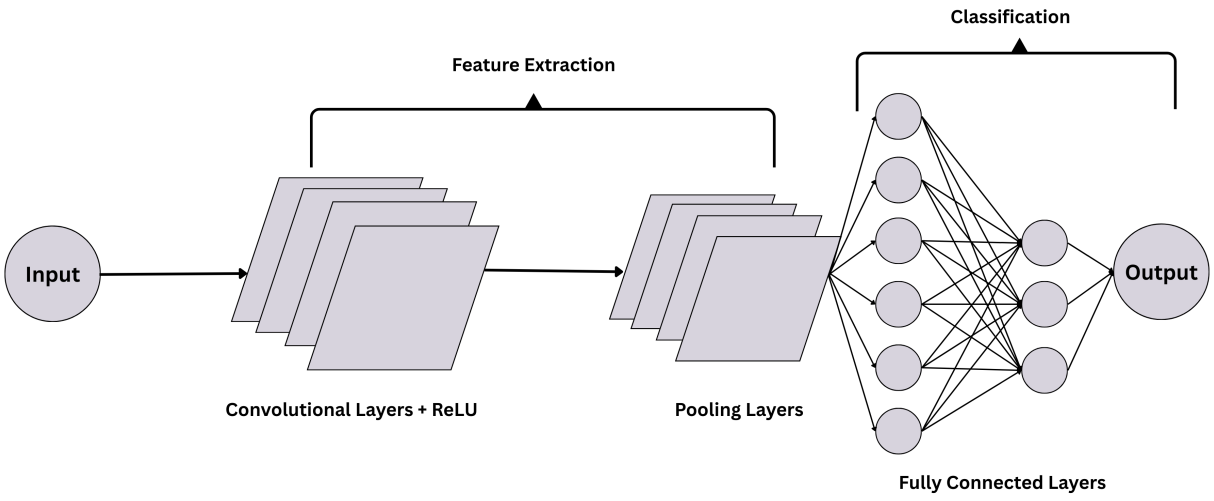


Figure 2.3: Convolutional Neural Networks.

2.2.2.4 Data Augmentation

Data augmentation is a technique that helps increase the size, diversity and quality of the training dataset by creating a set of modified and related data from a limited dataset to be used in machine learning and deep learning so that models can be trained with them. Traditional data augmentation techniques for images, audios, and text augmentation include geometric transformations such as rotation, flipping, scaling, cropping, zooming, and translation, which introduce spatial variations. Color transformations like brightness adjustment and contrast alteration help models become invariant to lighting conditions. Noise injection by adding gaussian or random noise and blurring can improve robustness by simulating real-world noise. Additionally, mixing images and random erasing or mask regions to create more diverse samples, also Word or sentence shuffling by randomly changing the position of a word or sentence. However, these basic augmentation methods may not sufficiently capture complex variations in the data. To address this, advanced techniques leverage generative methods such as GANs and Diffusion Models to synthesize entirely new data instances that resemble the original distribution [31, 32]. Data augmentation has its advantages and disadvantages, such as:
Advantages:

- Reduces overfitting by generating more data from limited samples and increases dataset size.
- Enhances model generalization by increasing its resilience to variations and noise.
- Enhances performance which leads to better accuracy and reliability, especially in image classification tasks.
- Cost-effective, it reduces the need for collecting and labeling large datasets [33].

Disadvantages:

- Increases computational cost, as models must train on a larger number of augmented images.
- Excessive or unrealistic augmentation can degrade model performance.
- Some augmentations may not generalize well across different tasks or domains, for example, applying color adjustments to grayscale images (like black-and-white sketches) may result in unrealistic images that do not match the target domain.
- Data augmentation, especially with advanced generative methods like GANs, can produce unrealistic or misleading images (hallucinations), requiring specialized expertise to carefully validate the generated samples for accuracy and reliability [33, 31].

2.2.3 Generative Artificial Intelligence

Generative Artificial Intelligence or Gen AI, is an artificial intelligence branch designed to generate new data samples that resemble the training data by learning the underlying data distribution, patterns, and styles of a given dataset. These models can create entirely new data instances, making them useful and ideal for different tasks such as image synthesis, text generation, audio, data augmentation, and even complex 3D designs.
Generative AI typically functions in three main phases:

- Training: Builds a foundational model that serves as a starting point for a wide range of generative AI applications.
- Tuning: Adapts the foundational model to suit a specific generative AI task or application.
- Generation, Evaluation, and Retuning: Produces outputs, assesses their quality, and continually refines the model for improved accuracy and performance.

Generative AI has gained significant attention due to its transformative capabilities across various domains. In computer vision, it is used to create high-quality images, enhance image resolution, or perform style transfer. In natural language processing, it powers text generation, language translation, and conversational agents. Recent advancements have also enabled generative AI to design new molecules for drug discovery, simulate complex physical phenomena, and generate synthetic medical images for training machine learning models.
The core of generative AI lies in its use of advanced generative models, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models as shown in Fig 2.4, each leveraging different techniques to learn and generate data [34, 35, 36].

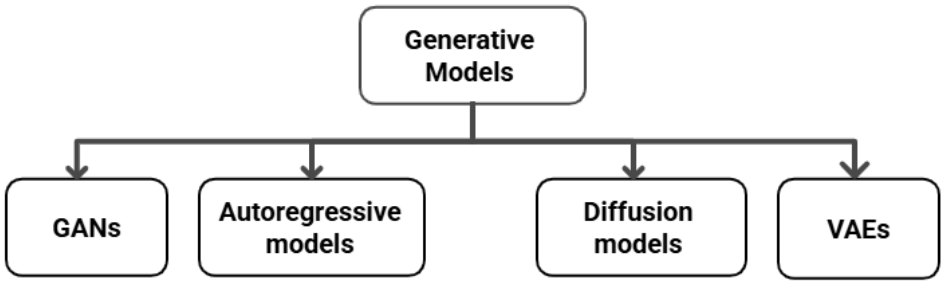


Figure 2.4: Generative Models Types.

2.2.3.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs), are unsupervised generative models introduced by Goodfellow et al. the concept is to train two neural networks named generator and discriminator as shown in Fig 2.5, where the generator receives a random noise vector, typically drawn from a normal or uniform distribution, which acts as a seed or starting point for generating data. The generator creates data as its output, such as an image in the case of a GANs designed for image generation. This generated data is then passed to the discriminator, which classifies it as either real or fake. The generator updates its weights in response to feedback from the discriminator. If the discriminator accurately recognizes the generated data as fake, the generator modifies its weights to create more realistic data in the next iteration.
The discriminator gets the real and the generated data as inputs and tries to distinguish the real from the fake ones. It produces a scalar value between 0 and 1, where a value close to 1 indicates the sample is probably real, while a value near 0 indicates it is probably fake. Both the generator and discriminator are trained using binary cross-entropy loss.
The whole goal is that the generator have to produce convincing fake data to fool the discriminator while the last have to differential between the real and fake no matter how good the fake

data looks like, it can be viewed as a minimax game which is described in the following value function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]. \tag{2.4}$$

Where D and G represent, respectively, the discriminator and the generator, $p_{\text{data}}(x)$ is the true data distribution, $p_z(z)$ is Gaussian or uniform noise from which generator inputs are sampled. The $\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]$ is the average of $\log D(x)$, where x comes from the real data distribution, and $\mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$ is the average of $\log(1 - D(G(z)))$ where z is random noise, and $G(z)$ is the generated (fake) image.
$\log D(x)$ assign high probabilities to real images, and $\log(1 - D(G(z)))$ assign low probabilities to fake images.
GANs are widely used in applications such as image synthesis, style transfer, data augmentation, super-resolution, and text-to-image generation.
Despite their remarkable success in generating realistic data, GANs face several limitations, including training instability, where the generator and discriminator may fail to reach balance, leading to mode collapse (producing limited variations). They also demand substantial computational resources for training due to their high complexity [37, 38].

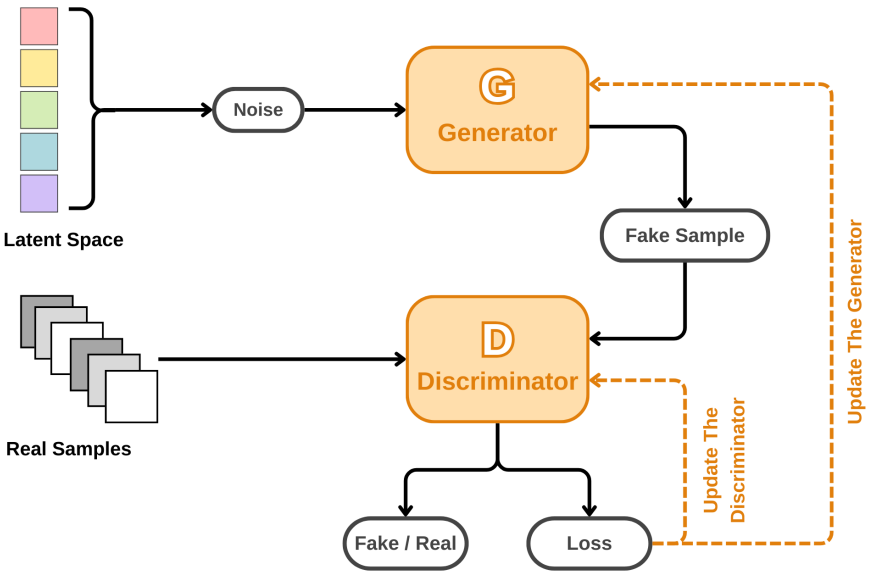


Figure 2.5: GAN Architecture.

2.2.3.2 Diffusion Models

Diffusion Models (DMs) are a class of generative models that operate through a two-step process, the Forward Diffusion Process and the Reverse Diffusion Process. These models are inspired by physical diffusion processes, where particles naturally disperse from crowded areas to less crowded areas. In diffusion models, this idea is used to gradually turn clear data into

noisy data and then learn how to reverse this process to generate new, realistic data.
In the forward diffusion phase, the model begins with a clean data sample (such as an image) and gradually adds Gaussian noise to it which is a type of statistical noise characterized by a probability distribution known as the Gaussian (or normal) distribution, the noise is random, but most values are close to 0, over a series of steps. Each step follows a Markov chain, meaning the noise added at any step depends only on the state of the data from the previous step. This is presented in the equation below, where $\beta_t \in (0, 1)$ is noise schedule for step t, $\mathcal{N}(.)$ is a normal (Gaussian) distribution, and x_t is a noisy version of the data at step t:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}) \tag{2.5}$$

The key idea is to progressively degrade the data in a controlled manner, making it almost indistinguishable from random noise at the final step.
The reverse denoising phase is where the true power of diffusion models lies. During this phase, the model learns to reverse the noise addition process step by step, using another Markov chain. It starts with pure noise and gradually denoises it, recovering the original data distribution from $x_T \to x_{T-1} \to \cdots \to x_0$. This denoising is guided by a series of learned noise estimation networks, allowing the model to generate clear and realistic data. As shown in the equation below, where $\mu_\theta(x_t, t)$ is the predicted mean of the denoised image, Σ_θ is the predicted or fixed variance, and θ is the parameters (the weights and biases) of the neural network.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2.6}$$

This dual-phase design as shown in Fig 2.6, it describes the probability of what the image looks like at step t, based on what it looked like at the previous step t-1 during the forward diffusion process, makes diffusion models highly effective at generating detailed and accurate data across various domains, including image synthesis, text generation, and even molecular design [39, 40].

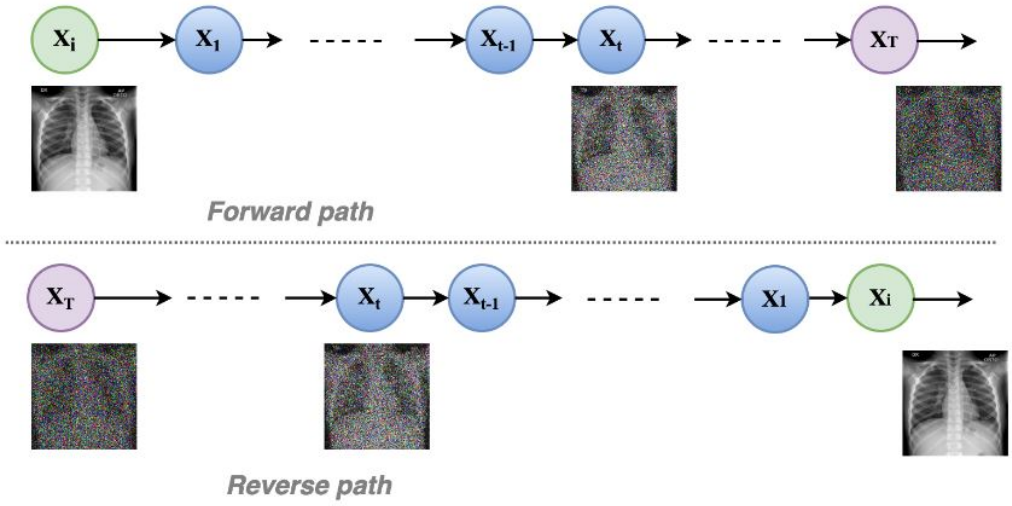


Figure 2.6: Forward pass and reverse pass in diffusion model training [2].

2.3 Conclusion

This chapter provided the foundational concepts necessary for understanding the methodologies explored in this thesis, covering key principles of artificial intelligence, machine learning,

and generative models. This theoretical background established, which underpins the methodologies explored in the subsequent chapters on medical image synthesis.

CHAPTER 3

RELATED WORK

Contents

3.1 Introduction

The rapid advancement of deep generative models has opened new horizons for synthesizing high-fidelity medical images, mitigating data scarcity caused by various factors [41]. This section surveys recent and current research in this field, with a particular emphasis on GANs and diffusion models approaches for generating CXR images. We organize these studies along a chronological timeline, offering a clear picture of the field's evolution, pinpointing current limitations, and identifying the specific gaps that our work will address.

3.2 Generative Adversarial Networks

Generative Adversarial Networks, consist of two competing neural networks, a generator that creates synthetic samples and a discriminator that tries to tell real data apart from those fakes. Through this adversarial process, the generator gradually learns to produce increasingly realistic outputs that can "fool" the discriminator, while the discriminator becomes ever better at spotting the fakes as a minimax two player game [37], it is a data augmentation technique to augment the training dataset [31]. Since the first introduction of GANs a wide variety of GANs variants have been proposed to address different challenges such as training instability, mode

collapse, and conditional generation like DCGAN [3], CGAN [42], CycleGAN [43], StyleGAN [44] and ProGAN [4]. The evolution of a few popular GANs can be seen in Fig. 3.1. In this part we will be talking about DCGAN [3], WGAN-GP [45], ProGAN [4], and StyleGAN3 [5].

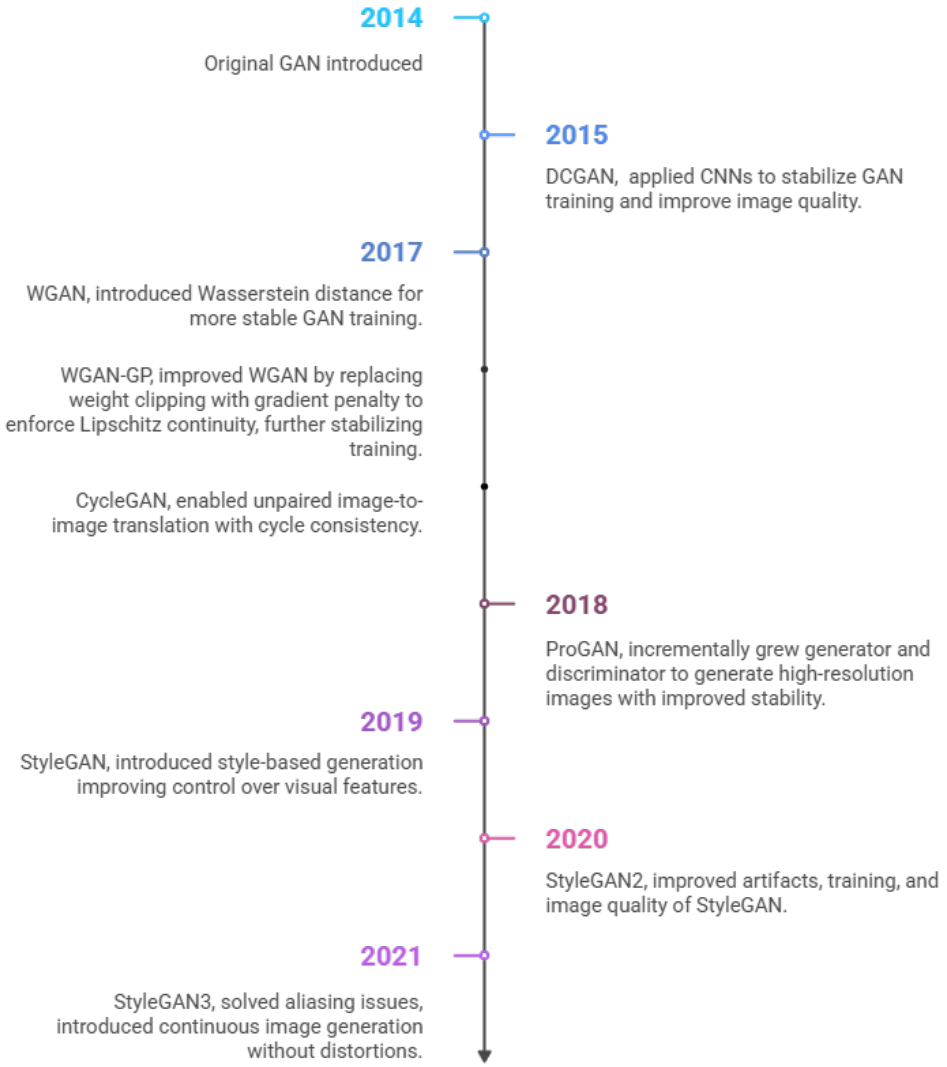


Figure 3.1: GANs Variants Chronological appearance.

3.2.1 Deep Convolutional Generative Adversarial Networks (DCGAN)

Deep Convolutional Generative Adversarial Networks (DCGAN), is a type of generative adversarial network architecture that applies convolutional neural networks (CNNs) to both the generator and the discriminator models to improve image generation quality. Proposed by Radford et al. in 2015, DCGAN replaces the fully connected layers of the original GANs [37] with convolutional layers, enabling the networks to learn spatial hierarchies and features more

effectively.
It replaces traditional pooling layers with strided convolutions in the discriminator for downsampling and fractional-strided (transposed) convolutions in the generator for upsampling, allowing the network to learn spatial transformations automatically. To stabilize training and accelerate convergence, batch normalization is applied to most layers in both the generator and discriminator, except at the generator's output and the discriminator's input. For activation functions, the generator uses ReLU activations in all layers except the output, where Tanh is used to scale pixel values between -1 and 1. The discriminator, on the other hand, employs LeakyReLU activations to maintain gradient flow even when neurons are inactive, improving the learning process.
As shown in Fig 3.2 the generator takes a low-dimensional latent vector (e.g., random noise) and upsamples it through transposed convolutions, each followed by batch normalization and ReLU, progressively upsampling the vector to produce realistic images, while the discriminator downscales input images through strided convolutions with batch normalization and LeakyReLU activations, extracting features and to classify the images as real or fake. This design enables DCGAN to generate higher-quality and more stable image samples compared to the original GANs architecture [37, 3].

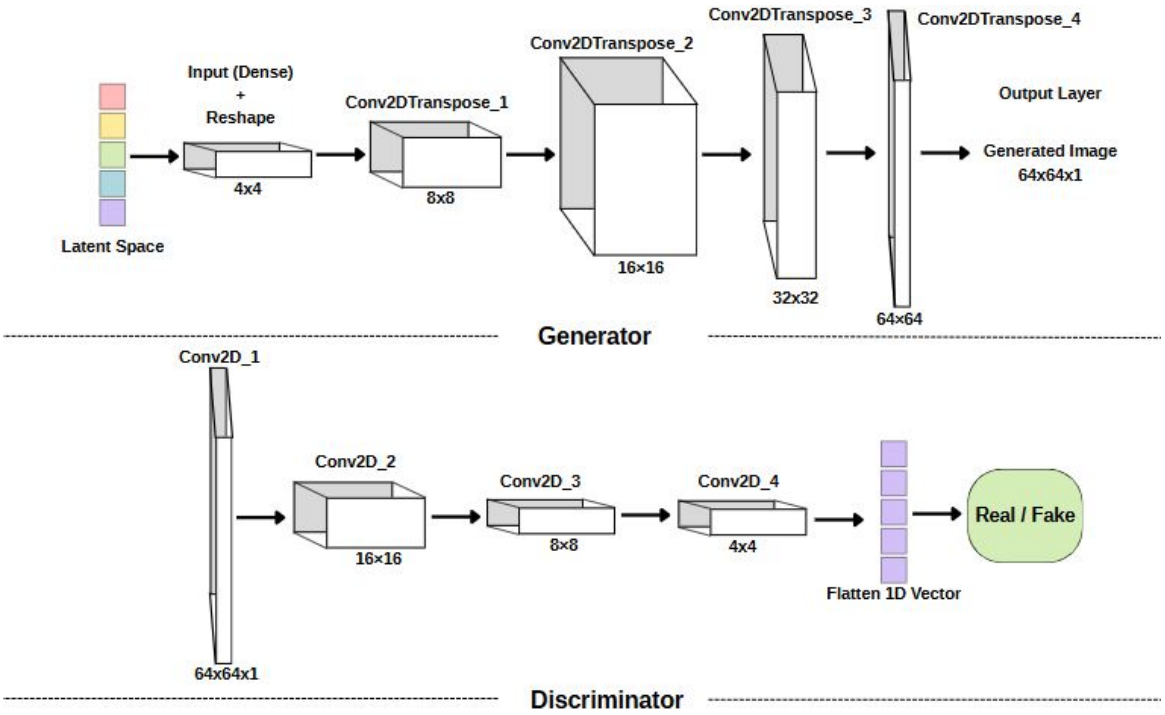


Figure 3.2: DCGAN Generator And Discriminator [3].

The synthesis of realistic medical images via generative models has been the subject of extensive investigation in recent years. Ubale Kiru et al. (2022), carried out a comparative study of five GAN architectures: CGAN [42], DCGAN [3], f-GAN [46], WGAN [47], and CycleGAN [43] on two COVID-19 imaging datasets CT scans and CXR, they used only 50 pictures of each datasets and all the images were of 28x28 pixels and a greyscale value assigned to each pixel, trained each GAN for 60 epochs using identical hyperparameters such as Adam optimizer, learning rate 1e-5, batch size 64. They evaluated images realism qualitatively using VTT by a medical expert, and quantitatively using FID. In the CXR experiments, DCGAN

[3] achieved 72% accuracy in the VTT, and an FID score of 17.45 [48].

In their 2023 study, Mai Feng Ng and Carol Anne Hargreaves [49], applied DCGAN to generate synthetic COVID-19 chest X-ray images as part of a strategy to augment limited annotated medical datasets. The authors trained DCGAN on 1,000 CLAHE (Contrast Limited Adaptive Histogram Equalization) preprocessed images resized to 64×64 pixels, running the training for 800 epochs.
Image quality was assessed using the FID, where DCGAN achieved a score of 1.399, outperforming WGAN-GP in this metric. The synthetic images were used to augment classifier training, and the resulting model achieved a classification accuracy of 0.98, an improvement over the 0.96 accuracy obtained using only real data. The study also noted that training the model for longer periods (beyond 800 epochs) led to a decline in image realism [49].

3.2.2 WGAN-GP

Wasserstein GAN Gradient Penalty (WGAN-GP), is a robust variant of GANs introduced by Gulrajani et al. (2017) to address the instability and convergence issues commonly observed in traditional GANs. It builds on the Wasserstein Generative Adversarial Network (WGAN) framework [47], which replaces the original GAN loss with the Wasserstein distance (also known as Earth Mover's Distance), which is defined as:

$$W(P_r, P_g) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim P_r}[f(x)] - \mathbb{E}_{\tilde{x} \sim P_g}[f(\tilde{x})] \tag{3.1}$$

to provide a more reliable and stable way to estimate the difference between real and generated data distributions, where:

- P_r is the real data distribution, and P_g is the generated data distribution.
- sup is the least upper bound, and and f is a 1-Lipschitz function.
- $\mathbb{E}_{x \sim P_r}$ is expectation over real data samples, and $\mathbb{E}_{\tilde{x} \sim P_g}$ is the expectation over generated data samples.

Instead of using weight clipping (which forces the critic's parameters to stay within a small fixed range, such as -0.01 to 0.01), WGAN-GP introduces a gradient penalty to enforce the 1-Lipschitz constraint, which ensures the critic's output does not change too abruptly with small input changes.
The architecture as shown in Fig 3.3, consists of a generator, which creates synthetic data from random noise, and a critic (rather than a discriminator), which evaluates how realistic the samples are without explicitly classifying them as real or fake, it assign higher scores to real images and lower scores to fake ones.
To ensure the critic satisfies the required 1-Lipschitz constraint (which means that the critic should not change its output too rapidly, small changes in input shouldn't cause big jumps in output), WGAN-GP introduces a gradient penalty instead of weight clipping, this penalty enforces smoothness by penalizing the model when the gradient norm deviates from 1 [45].
This leads to more reliable convergence, reduces mode collapse, and results in higher-quality outputs across tasks such as image synthesis and medical image generation.

In the same 2023 study by Mai Feng Ng and Carol Anne Hargreaves [49], they also explored the use of WGAN-GP for generating synthetic COVID-19 X-ray data. The goal was to improve the performance of downstream classifiers through data augmentation. Like DCGAN, WGAN-GP was trained on 1,000 CLAHE preprocessed chest X-ray images resized to 64×64 pixels for 800 epochs.
While WGAN-GP achieved a slightly higher FID score (1.583) compared to DCGAN, it still significantly contributed to enhancing classification performance, with augmented data pushing InceptionV3 classifier accuracy to 0.98 versus 0.96 when trained on real images alone.
The study further reported that prolonged training beyond 800 epochs adversely affected image quality for both models [49].

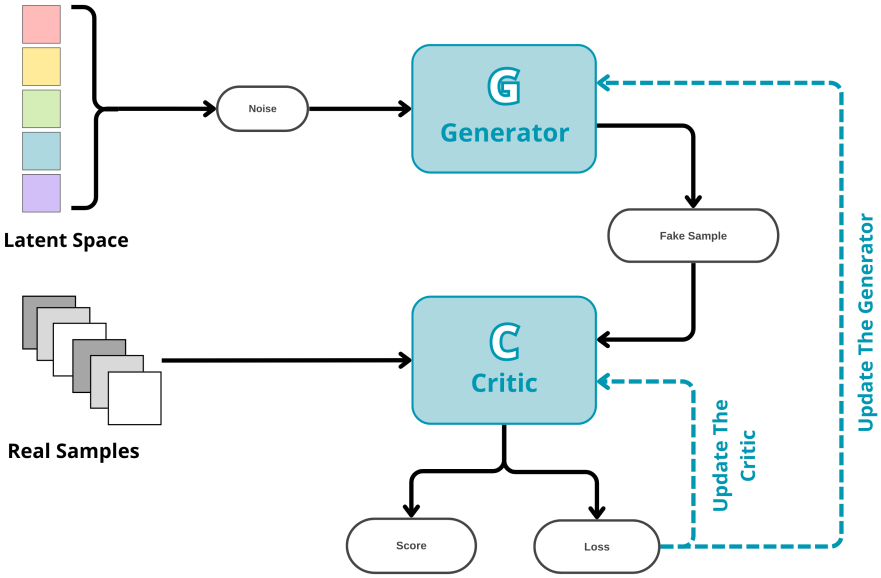


Figure 3.3: WGAN-GP Architecture.

3.2.3 Progressive Growing GAN (ProGAN)

Progressive Growing Generative Adversarial Networks (ProGAN), is a type of generative adversarial network architecture, introduced by Karras et al. in 2017, designed to improve the quality, stability, and variation of generative adversarial networks, especially for high-resolution image synthesis, where traditional GANs often struggle with generating high-resolution images due to training instability and mode collapse, ProGAN addresses these challenges by gradually increasing the resolution of generated images during training.
The core idea of ProGAN, as shown in Fig 3.4 is to start training at a low resolution images (e.g., 4×4) and progressively expand both the generator and discriminator networks by adding layers that increase the resolution (e.g., to 8×8, 16×16, ..., up to 1024×1024). This progressive growing allows the networks to first learn coarse features and then refine finer details, making the training more stable and efficient.
Both the generator and discriminator grow symmetrically. The generator begins from a latent noise vector and produces a 4×4 image, then new layers are incrementally added to upsample the output. Conversely, the discriminator receives high-resolution images and progressively downsamples them through added convolutional layers.
When new layers are introduced during training, a fade-in mechanism as shown in Fig 3.5 is

used to avoid abrupt transitions. This is achieved by a parameter α, which linearly interpolates between the outputs of the old (lower-resolution) and new (higher-resolution) layers. As α increases from 0 to 1, the network gradually shifts its focus to the newly added layers. Additional techniques employed in ProGAN include a mini-batch standard deviation layer in the discriminator to promote variation in outputs and the use of Wasserstein loss with gradient penalty (WGAN-GP) for more stable convergence [4].

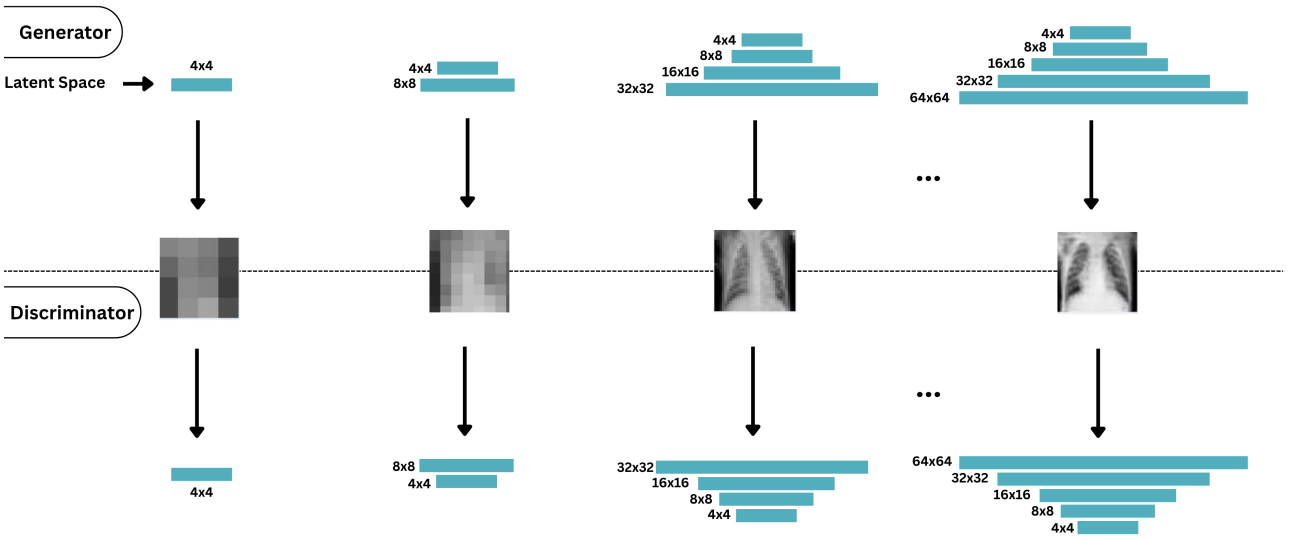


Figure 3.4: ProGAN Architecture.

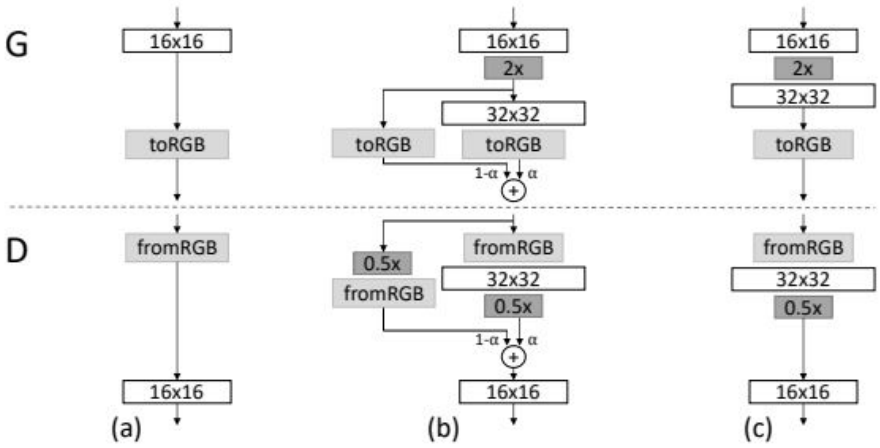


Figure 3.5: Fade-in Mechanism [4].

ProGAN is best known for generating ultra-realistic human faces, Segal et al. (2021) [50], in their work aimed to generate high-resolution CXR images using a modified Progressive Growing GAN (PGGAN) and assess their clinical realism and utility in diagnostic tasks.
The model was trained on the ChestX-ray14 dataset [51] using enhancements such as WGAN-GP loss, minibatch discrimination, and pixel normalization, the training took 6 days and was conducted up to a resolution of 1024×1024 on Amazon Web Services (AWS) infrastructure.

The results were evaluated using the FID with a score of 32.05 for the Pneumonia class, and with an expert review by radiologists whom misclassified synthetic images as real 61% of the time, and real images were correctly identified 73% of the time.
A classifier trained uniquely on synthetic images achieved an accuracy higher than the classifier trained uniquely with real images [50].

Miso Jang et al. (2023) [52], tried to generate high-resolution CXR images using Progressive Growing GAN (PGGAN) and evaluate their realism and utility for training deep learning classifiers.
The dataset comprised 72,958 normal and 91,163 abnormal CXR sourced from Asan Medical Center in South Korea, with DICOM images converted to 1024×1024 PNG format. Two models were trained: ABN-PGGAN using all 91,163 abnormal images, and NOR-PGGAN using only the 72,958 normal images.
The evaluation involved visual Turing tests (VTT) with six radiologists, including both experts and residents, who were asked to distinguish between real and synthetic CXR. In addition, convolutional neural networks (CNNs) were trained and tested using real images, as well as a combination of real and synthetic images, to assess the utility of synthetic data in classification tasks.
The results demonstrated that radiologists frequently struggled to differentiate synthetic images from real ones, with mean classification accuracies of 67.4% for ABN-PGGAN and 69.9% for NOR-PGGAN images. In terms of classification performance, CNNs trained exclusively on real data achieved an accuracy of 91.0% while those trained on both real and synthetic data performed slightly better, achieving an accuracy of 93.3% [52].

3.2.4 StyleGAN3

Style-Based Generative Adversarial Network 3 (StyleGAN3) is a state-of-the-art generative adversarial network (GAN) architecture proposed by Karras et al. (2021) [53], designed to synthesize high-resolution, high-fidelity images with significantly improved spatial consistency and temporal coherence. As the successor to StyleGAN2, it introduces a novel alias-free synthesis approach grounded in signal processing principles [53]. This approach addresses a key limitation of earlier GANs: aliasing artifacts, which result from inconsistent spatial frequency handling during upsampling, downsampling, and nonlinear transformations [53]. Unlike its predecessors, StyleGAN3 models feature maps as continuous signals rather than discrete grids and strictly enforce band-limited representations through the use of carefully designed low-pass (FIR) filters [53]. This design ensures translation equivariance, the output image shifts smoothly and predictably when the input is translated and in some variants, approximate rotation equivariance [53].
StyleGAN3 comes in two main variants:
StyleGAN3-T: is optimized for translation equivariance, making it particularly well-suited for video synthesis and producing temporally stable results [53, 54].
StyleGAN3-R: which additionally supports rotation equivariance, making it effective for symmetry-sensitive domains like medical imaging or scientific visualization [53, 54].
By eliminating flickering artifacts, spatial jitter, and other high-frequency inconsistencies, StyleGAN3 sets a new benchmark for alias-free image generation, and is widely adopted in domains requiring precise spatial transformations or realistic motion continuity [53, 54].

StyleGAN3 has shown considerable potential in the field of medical imaging, especially for synthesizing realistic chest X-Ray images that can support tasks such as data augmentation, anomaly detection, and deep learning model training. Its alias-free design ensures spatial accuracy and removes visual defects like flickering or pattern distortion, which is crucial for applications requiring fine-grained anatomical detail [53]. In chest radiography, StyleGAN3 helps address challenges like class imbalance and insufficient data by creating synthetic images that closely mimic real diagnostic content. According to McNulty et al. 2024 [55], using GAN generated images to supplement training datasets led to improved performance in classifying rare conditions in chest X-Rays. Earlier versions like StyleGAN2 struggled with spatial artifacts that undermined clinical reliability, whereas StyleGAN3's ability to maintain consistency during image translations and rotations has made it more appropriate for tasks such as disease tracking and anatomical change detection over time [53]. Additionally, its use has extended to unsupervised anomaly detection and semi-supervised learning, enabling models to learn effectively from limited labeled data [56]. Overall, StyleGAN3 supports more reliable and scalable approaches to medical imaging, especially in settings where acquiring large and well-balanced datasets is a challenge.

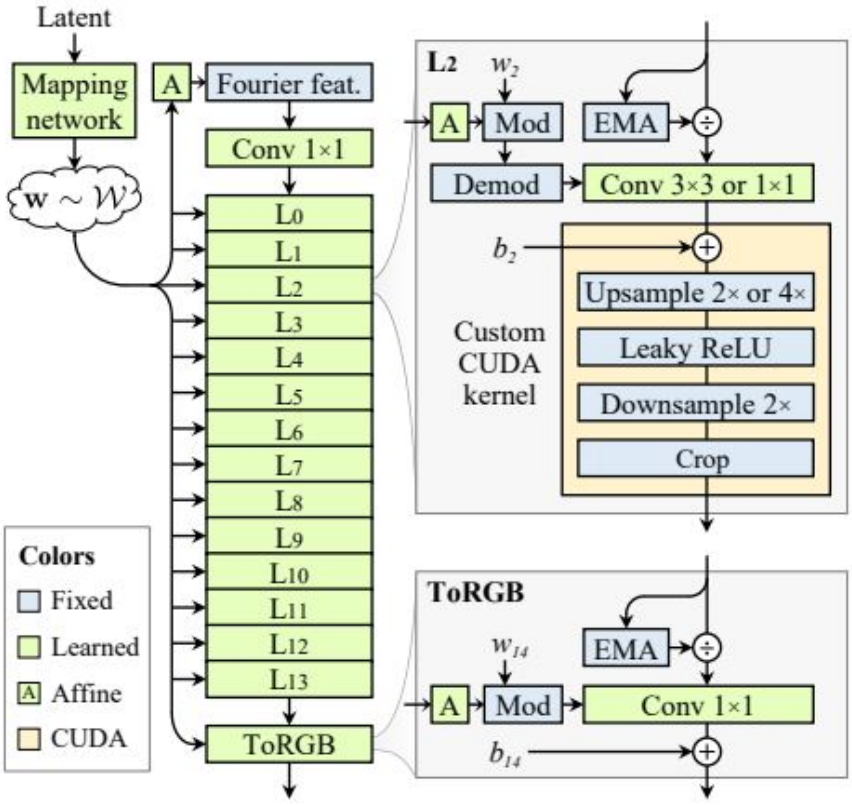


Figure 3.6: Alias-Free StyleGAN3 generator architecture [5].

StyleGAN3 improves on StyleGAN and StyleGAN2 by making image generation alias-free, meaning it avoids visual artifacts like flickering or texture shifting when images are moved or scaled. It treats images as smooth, continuous signals instead of pixel grids, which helps maintain spatial consistency.
The generation starts with a mapping network that turns a random vector z into a style vector w, which controls how the image is generated at each layer. Unlike earlier versions, StyleGAN3

removes positional encodings and noise injection. Instead, it uses modulated convolutions combined with low-pass FIR filters before and after each layer to prevent aliasing.
The generator increases resolution step by step (e.g., $4 \times 4 \rightarrow 8 \times 8 \rightarrow ... \rightarrow 1024 \times 1024$), using smooth filter-based upsampling instead of transposed convolutions. At different scales, feature maps are converted into RGB images and combined to form the final output.
The discriminator is also updated with FIR-based downsampling to match the generator's alias-free structure, improving training stability and image quality.
These improvements are especially useful in medical imaging (like chest X-rays), where it's important to keep anatomical details clear and stable [5].

3.3 Diffusion Models

Diffusion Models (DMs), are generative models that synthesize data through two phases: a forward diffusion process that gradually adds Gaussian noise to clean data, and a reverse process that denoises it step by step to regenerate realistic samples. Inspired by physical diffusion and based on Markov chains, these models learn to recover original data from noise using trained denoising networks. Their two-phase design makes them powerful for generating high-quality images, text, and more.
Diffusion models can generally be divided into two categories: text-to-image diffusion models, which condition the generation process on textual prompts, and standard diffusion models, which generate images without any external input [57]. The historical development of diffusion models is illustrated in Fig 3.7. While many models are shown, we will focus on discussing only those most pertinent to our objectives such as DDPMs and DDIMs.

3.3.1 Standard Diffusion Models

In this work, we deliberately focus on standard diffusion models with the goal of building autonomous image generators that operate without the need for external prompts or conditioning inputs.
Our objective is to develop models capable of learning the underlying data distribution of chest X-ray images and synthesizing realistic samples directly from noise, without relying on descriptive text or any human-provided guidance. Standard diffusion models are specifically designed to learn the underlying distribution of images and generate new samples without the need for additional inputs, such as text prompts. In the context of chest X-ray datasets, detailed textual descriptions accompanying each image are generally not available, making text-to-image approaches unsuitable. Text-to-image diffusion models, like Stable Diffusion and DALL·E, rely on paired text-image data to translate descriptive prompts into images, which is rare in the medical imaging field. In contrast, standard diffusion models require only image data to learn the generation process. Moreover, text-to-image diffusion is part of a separate research domain that focuses on processing and interpreting language to direct the image generation process. Our work, however, concentrates on the generation of chest X-ray images to augment datasets and support AI training in medical imaging, rather than addressing the challenges of language-based image generation.
In this context, we focus on two prominent types of diffusion models which are Denoising Diffusion Probabilistic Models (DDPMs) [6] and Denoising Diffusion Implicit Models (DDIMs) [7],

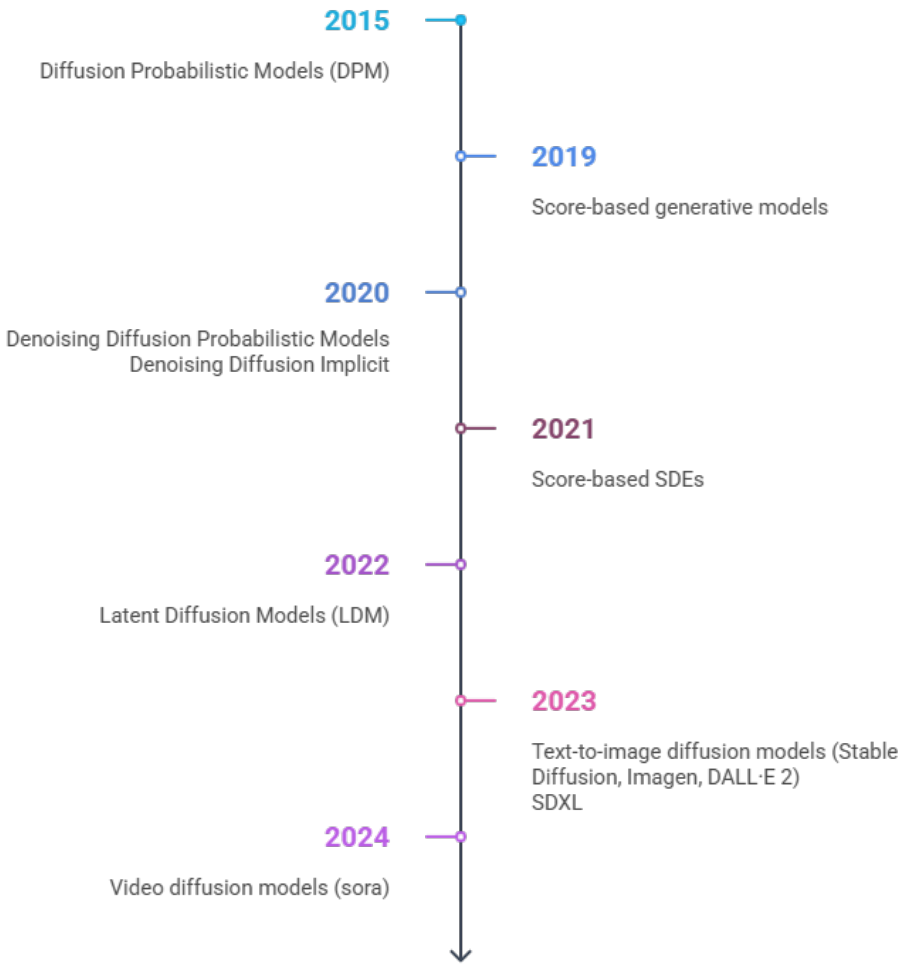


Figure 3.7: Diffusion Models Timeline.

both of which are well-suited for unconditional image synthesis tasks.

3.3.1.1 Denoising diffusion probabilistic models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs), are a class of generative models that produce high-quality synthetic data through a two-step process inspired by thermodynamic diffusion. Initially proposed by Ho et al. (2020), that generate data by learning to reverse a gradual noising process.
The forward process is formulated as a Markov chain that gradually degrades input data such as images, by introducing Gaussian noise step by step over multiple timesteps. Starting from a clean image, each step introduces a small amount of noise according to a predefined schedule, eventually transforming the data into nearly pure noise, this is modeled by the distribution $q(x_t \mid x_{t-1})$, shown as the dashed arrow in Fig 3.8.
The reverse process involves training a neural network commonly a U-Net architecture, to predict the noise added at each timestep, allowing it to iteratively denoise the sample, this process is learned and modeled by the neural network via $p_\theta(x_{t-1} \mid x_t)$.
During training, the model minimizes the discrepancy between the predicted and actual noise using a mean squared error (MSE) loss. This incremental denoising framework makes the

generative task more manageable than direct single-step generation [6, 58].

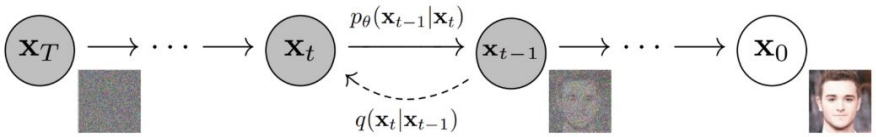


Figure 3.8: How DDPMs Works [6].

In [59], Iman Khazrak et al. proposed a comprehensive framework to evaluate how two generative models, DDPMs and ProGAN can improve classification performance on small and imbalanced chest X-ray datasets. Used chest X-ray data (1,802 normal and 1,800 pneumonia images) from Kaggle, and trained a DDPM to generate 2,000 synthetic chest X-ray images per class (normal and pneumonia). These DDPM generated images were then added to small and imbalanced datasets, and the performance of four different classifiers: a custom CNN, untrained VGG16, pretrained VGG16, and pretrained ResNet50 was evaluated, on small and imbalanced datasets with Greedy K sampling, DDPM achieved up to 0.97 accuracy. The results showed that DDPM consistently outperformed PGGAN in both image quality and classification effectiveness and well-suited for augmenting small, imbalanced medical datasets to improve AI diagnostic tools [59].
To address the data scarcity of pneumonia-positive chest X-ray images, in the article [60], Thulana Abeywardane et al. proposed a DDPM with a U-Net architecture to generate realistic chest X-ray images of pneumonia cases. They trained their model on over 3,000 pneumonia X-ray images from a Kaggle dataset, preprocessing them to grayscale and resizing to 256×256 pixels. The model was trained on Google Colab for 50 epochs, achieving its best performance at epoch 31. The evaluation was done using Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM), with results showing that synthetic pneumonia images were more similar to real pneumonia images (MSE: 105.65, SSIM: 0.43) [60].

3.3.1.2 Denoising diffusion Implicit models (DDIMs)

Denoising Diffusion Implicit Models (DDIMs), introduced by Jiaming Song, Chenlin Meng, and Stefano Ermon in 2020 [7], are a class of generative models designed to address the slow sampling speed inherent in DDPMs [6].
While DDIMs shares the same forward diffusion process as DDPMs, where a clean image is progressively degraded by Gaussian noise across a series of timesteps, the key innovation lies in its deterministic and non-Markovian reverse process.
Unlike DDPMs, which follow a step-by-step Markov chain, DDIMs allows the model to deterministically sample states such as x_3 from x_2 and x_0, or x_2 from x_1 and x_0, as shown in the Fig 3.9. This approach leverages the predicted noise from the model to compute previous states without random sampling, thereby allowing high-quality generation in as few as 10 to 50 steps. DDIMs retains the same U-Net-based neural network architecture used in DDPMs, incorporating timestep embeddings and skip connections to predict either the added noise or the original image.
This design shows a more flexible process that can deterministically reconstruct images by leveraging the original data at each step and ensures that DDIMs can be seamlessly integrated

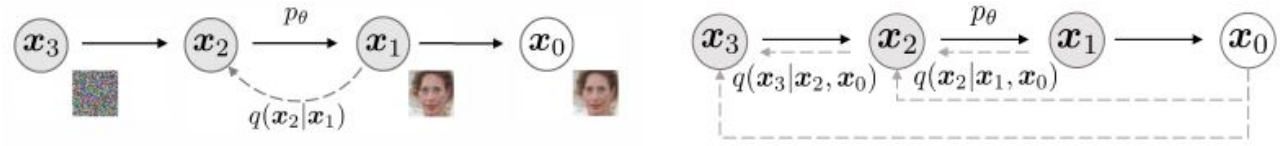


Figure 3.9: Graphical models for diffusion (left) and non-Markovian (right) inference models [7].

into existing diffusion pipelines while offering significant improvements in sampling speed and efficiency [7, 61].
While diffusion models have been increasingly explored in medical imaging, to the best of our knowledge, no existing studies have applied DDIMs specifically for the generation of chest X-ray images. Although several studies have explored diffusion-based approaches in medical imaging, for instance, the paper [62] introduces a method called Dif-fuse, which combines DDPM and DDIM to generate healthy-looking versions of brain scans. In this approach, DDPM is used to modify the pathological regions, while DDIM reconstructs the unaffected areas, resulting in realistic and coherent counterfactual images.

3.3.2 Text-To-Image Diffusion Models

Text-To-Image diffusion models are an advanced class of generative models designed to produce high quality images from natural language descriptions. These models operate through a two step process: in the forward process, an image is progressively noised until it becomes nearly indistinguishable from random noise.
In the reverse process, a neural network typically a U-Net, learns to gradually remove the noise step by step to reconstruct a meaningful image [6]. A central component in this process is the text prompt, which acts as the semantic guide for image synthesis. When a user inputs a prompt, such as "a futuristic cityscape at sunset with flying cars", the text is first passed through a text encoder like CLIP or T5. This encoder converts the prompt into a high-dimensional embedding vector, which captures the semantic meaning of the sentence. This embedding is then injected into the diffusion model usually through cross attention layers or concatenated as conditioning information at each denoising step. As the model removes noise over time, it uses the text embedding to ensure that the evolving image matches the content described in the prompt. This mechanism allows the model to align visual features (such as objects, colors, styles, and settings) with the structure and meaning of the input text. As a result, Text-To-Image diffusion models can generate highly diverse and semantically accurate images from a wide range of user provided descriptions (Rombach et al., 2022) [63].
Several influential models have been developed based on the diffusion framework, each offering unique architectural innovations and practical advantages. The following sections present a brief overview of three notable text-to-image diffusion models.

3.3.2.1 Stable Diffusion

Stable Diffusion has been widely explored in medical imaging for generating synthetic data with high realism and low computational cost. It uses a latent space diffusion process, which makes it suitable for tasks like simulating rare conditions or augmenting small datasets. Recent studies have applied Stable Diffusion to chest radiographs, showing promise in improving

classification tasks and preserving diagnostic features in synthetic samples [63, 64, 65].

3.3.2.2 DALL-E

DALL-E's ability to generate creative and coherent images from text prompts has led to applications in medical education and illustration. Though not trained on clinical data, it has been evaluated for generating concept level medical images and assisting with visual training datasets. For instance, it has been used to depict anatomical structures or disease scenes based on textual input, with encouraging results in educational contexts [66, 67].

3.3.2.3 Imagen

Imagen, though originally developed for general-purpose image generation, has been adapted into Surgical Imagen a model that transforms textual action descriptions (e.g., "clipper clip cystic duct") into photorealistic surgical images. This adaptation uses the same cascaded diffusion framework and T5-based text encoder as the original Imagen, demonstrating high alignment and realism suitable for surgical training and simulation purposes [68, 69].

3.4 Conclusion

In summary, the existing body of research demonstrates a growing interest in applying generative models particularly GANs and diffusion models to medical imaging tasks such as data augmentation, and image synthesis as shown in the Table 3.1. In the next chapter, we present the methodology of the approaches, also describe the datasets used, preprocessing steps, models architecture, training strategies, and evaluation metrics.

Table 3.1: Summary of The Articles Used in The Related Work.

Generative Models	Types	Articles
GANs	DCGAN	Comparative analysis of some selected generative adversarial network models for image augmentation: a case study of covid-19 x-ray and ct images [48]. Generative adversarial networks for the synthesis of chest x-ray images [70].
	WGAN-GP	Generative adversarial networks for the synthesis of chest x-ray images [70].
	ProGAN	Evaluating the clinical realism of synthetic chest x-rays generated using progressively growing gans [50]. Image turing test and its applications on synthetic chest radiographs by using the progressive growing generative adversarial network [52].
	StyleGAN3	Synthetic medical imaging generation with generative adversarial networks for plain radiographs [55]. Analyzing and improving the image quality of stylegan [56].
Diffusion Models	DDPM	Addressing small and imbalanced medical image datasets using generative models: A comparative study of ddpm and pggans with random and greedy k sampling [59]. Ddpm-based x-ray image synthesizer [60].
	DDIM	Diffusion models for counterfactual generation and anomaly detection in brain images [62].

CHAPTER 4

OUR CONTRIBUTION

Contents

4.1 Introduction

This chapter, outlines the key contributions of our work, which focus on the application and evaluation of advanced generative models namely GANs and Diffusion Models for the generation and analysis of chest X-ray images. Our objective is to explore the potential of these models in chest X-ray images.

We begin by detailing our experimental setup, including the dataset used, data preprocessing techniques, training strategies, model architectures, and evaluation metrics. The generative models explored in this study include DCGAN, WGAN-GP, ProGAN, and StyleGAN3 from the GAN family, as well as DDPMs and DDIMs from the class of diffusion models.

We present and analyze the results for Normal and Pneumonia chest X-rays using both FID and VTT metrics, and compare the performance of different generative models, highlighting their strengths and weaknesses in medical imaging.

4.2 Experimental Setup

This section outlines the key components of the experimental workflow used to evaluate various generative models for chest X-ray image synthesis.

4.2.1 Dataset

In this study, we collected chest X-ray images from five publicly available medical imaging datasets, focusing exclusively on two categories: normal and pneumonia cases, to build a diverse and comprehensive dataset for training and evaluating generative models. The dataset include:

- Chest X-Ray Images (Pneumonia) [71].
- NIH Chest X ray 14 (224x224 resized) [72].
- Pneumonia and Tuberculose with Normal and Non-X-ray [73].
- Tuberculosis (TB) chest x-ray database [74].
- Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification (chestxray2017) [75].

Each dataset contains different types of medical cases and X-ray images, which helps make the training more reliable and applicable to a wider range of situations. Table 4.1 summarizes the number of images extracted from each dataset. The combined dataset was used to train and evaluate six generative models, as described in the following sections.

Table 4.1: Number of chest X-ray images collected from each dataset.

Dataset	Normal	Pneumonia
Chest X-Ray Images (Pneumonia)	1,583	4,273
NIH Chest X ray 14 (224x224 resized)	-	120
Pneumonia and Tuberculose with Normal and Non-X-ray	5,488	4,273
Tuberculosis (TB) chest x-ray database	3,500	-
Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classifica- tion (chestxray2017)	1,583	4274
Total	**12,154**	**12,939**

4.2.2 Data Preprocessing

To ensure consistency across multiple datasets and enable fair comparison across models, several preprocessing steps were applied to the collected chest X-ray images. The datasets

used in this work were combined into two datasets, Normal Dataset with 12,154 images and Pneumonia Dataset with 12,939 images. Because generative models such as GANs and DMs are computationally intensive to train on large image sizes and are time-intensive, all models were trained on and generated images with a resolution of 64×64. As chest X-ray images are grayscale by nature, their pixel values were normalized to the range [-1,1] for all models, except for the DDIM model, which was normalized to the range [0,1] in accordance with the original implementation and its diffusion process assumptions.

4.2.3 Training Strategy

To make the comparison between the different generative models fair and consistent, all models were trained using the same process, with adjustments made for each model's specific design and needs. The batch size was dynamically adjusted depending on the model architecture and the available GPU memory, ranging from 32 to 128 samples per batch. All models were trained using the Adam optimizer, but with a learning rate that depends on each model in the range of 1×10^{-3} to 2×10^{-4}. All models were trained for a fixed number of epochs which is 500 epochs depending on convergence behavior. Intermediate checkpoints were saved periodically to allow visual monitoring of generated samples and recovery in case of training interruptions. Considering the loss function in GANs, DCGAN and ProGAN used the original GAN loss with binary cross-entropy, for the WGAN-GP the Wasserstein loss with a gradient penalty was used, StyleGAN3 used the path length regularization and non-saturating GAN loss as described in the original implementation. For the Diffusion Models DDPMs was trained using the mean squared error (MSE) loss, and DDIMs was trained using the mean absolute error (MAE). All models were trained using two NVIDIA RTX A5000 GPUs 24 VRAM for each, in a controlled environment, with either TensorFlow 2.18.0 or PyTorch 2.5.1 frameworks. The experiments were conducted on Ubuntu 24.04 LTS with CUDA 12.1.

4.2.4 Evaluation Metrics

To assess the quality of generated chest X-ray images, two evaluation metrics were employed: the Fréchet Inception Distance (FID) [76] and the Visual Turing Test (VTT). These metrics provide both quantitative and qualitative assessments of the realism and fidelity of the synthetic images produced by each model.

4.2.4.1 Fréchet Inception Distance (FID)

Fréchet Inception Distance (FID), is a widely used metric for evaluating the quality of images generated by generative models such as GANs or DMs. It was introduced in 2017 by Heusel, M., et al, [76], it compares real and generated images by looking at the patterns in their features, instead of just comparing pixels. These features are taken from a pretrained Inception-v3 model, which helps focus on how the images look overall, rather than checking each pixel exactly.
FID computes the Fréchet distance (also known as the Wasserstein-2 distance) between the multivariate Gaussian distributions of features from real and generated images. Mathematically, it is defined as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\frac{1}{2}}\right)$$

where μ_r, Σ_r are the mean and covariance of real image features, and μ_g, Σ_g are those for generated images, and Tr is the trace of a matrix (sum of the diagonal elements). As a rule of thumb, a lower FID score indicates better performance of a GAN or a Diffusion Model, as it signifies that the generated images more closely resemble the training images [76, 77].
In this study, FID was computed between the real chest X-ray test set and the generated images for each model. All images were resized to 299×299 and normalized to match the input requirements of the Inception-v3 model.

4.2.4.2 Visual Turing Test (VTT)

The Visual Turing Test (VTT), is a subjective evaluation metric where expert observers (e.g., radiologists or medically informed viewers) are asked to distinguish between real and generated images. If the synthetic images are indistinguishable from real ones to the human eye, the model is considered to have passed the test.
In this work, a set of real and generated chest X-ray images was shown in random order to a radiologist, who was asked to classify each image as either real or synthetic. The performance was measured by the percentage of generated images that were mistakenly classified as real. A higher score indicates greater visual realism of the generated images.

4.2.5 Models Architectures

This section presents the architectures of the generative models used in this study for chest X-ray image synthesis. We evaluate four GAN-based models (DCGAN, WGAN-GP, ProGAN, and StyleGAN3) and two diffusion-based models (DDPMs and DDIMs). Each model is implemented based on its original design, with minimal modifications to accommodate grayscale input and a fixed output resolution of 64×64 pixels. The following subsections describe the key components, design choices, and architectural differences of each model.

4.2.5.1 DCGAN

The DCGAN architecture is based on the original design proposed by [3], as illustrated in the Fig 3.2. In this work, it was adapted to generate grayscale chest X-ray images at a resolution of 64×64 pixels with a batch size equal to 64 and the images were normalized to the range of [-1,1]. The generator takes a random latent vector $z \in \mathbb{R}^n$ as input and produces a synthetic image through a series of upsampling blocks, typically implemented with transposed convolutional layers with LeakyReLU as an activation function. The discriminator is composed of downsampling convolutional layers also with LeakyReLU activation that assess the realism of the generated image, and the model is trained using the Binary Cross-Entropy loss function.

4.2.5.2 WGAN-GP

The WGAN-GP architecture builds upon the standard DCGAN framework by addressing its training instability through the use of the Wasserstein loss with a gradient penalty, as proposed by Gulrajani et al. (2017) [45].
In this model, the generator retains a similar structure to DCGAN, using transposed convolutional layers to progressively upsample a latent vector into a synthetic image. The discriminator,

referred to as the critic in WGAN has major changes. Specifically, batch normalization is removed from the critic to avoid introducing correlations between samples, and the final output layer is linear. The model is trained with a batch size of 128 with a learning rate of $1 \times 10^{-}4$. In this study, WGAN-GP is adapted for grayscale chest X-ray image generation at 64×64 resolution, with input normalization to the range [-1,1].

4.2.5.3 ProGAN

The ProGAN architecture is designed to improve the stability and quality of generative models when working with higher-resolution images. Instead of training the full-resolution generator and discriminator from the beginning, ProGAN starts with generating low-resolution images and gradually adds layers to both networks during training [4].
In this work, as illustrated in Fig 3.4, ProGAN is adapted for grayscale chest X-ray generation at a fixed resolution of 64×64 pixels, with input normalization to the range [-1,1]. The batch size varies depending on the target image resolution; in this setup, it is initialized using the list [32 , 32 , 32 , 16 , 16] corresponding to different stages of growth. Both the generator and the discriminator are built using convolutional layers and employ a fade-in mechanism to smoothly transition between resolutions during training, and with a learning rate $1 \times 10^{-}4$.

4.2.5.4 StyleGAN3

StyleGAN3 is a generative adversarial network designed to produce high-quality alias-free images, Its architecture is composed of three main parts.
First, the mapping network takes a random latent vector z and passes it through several fully connected layers to produce an intermediate latent vector w, which controls the style of the image.
Second, the synthesis network starts from a learned constant input and progressively builds the image using modulated convolutional layers,with carefully designed filters to build the image while avoiding aliasing artifacts.
Finally, the discriminator is a convolutional neural network that evaluates whether the image is real or generated, guiding the training process through adversarial learning [5].
In this study, pretrained StyleGAN3 models were adapted to generate grayscale chest X-ray images at a resolution of 64×64 pixels, with inputs normalized to the range [-1,1]. Minor modifications were made to support single-channel image output while preserving the core structure and functionality of the original model.

4.2.5.5 DDPMs

The DDPMs, is a generative model that synthesizes images through a gradual denoising process. It is based on the idea of reversing a forward diffusion process, in which Gaussian noise is progressively added to an image over a fixed number of time steps. The model learns to reverse this process by predicting the noise added at each step, effectively denoising a sample back to a clean image. The model is built around a U-Net architecture with skip connections that pass feature maps from the encoder (downsampling path) to the decoder (upsampling path). It also incorporates residual blocks for more stable training and better gradient flow. Additionally, self-attention blocks are used to capture long-range dependencies within the image. The network includes a fade-in concatenation mechanism that gradually merges encoder features into the decoder, helping reconstruct more detailed and accurate outputs [6]. The model is trained

using Mean Squared Error (MSE) between the predicted and actual noise. DDPM was adapted to generate grayscale chest X-ray images at a resolution of 64×64 pixels, with input images normalized to the range [-1,1], with a learning rate $2 \times 10^{-}4$, Adam optimizer and a batch size of 32.

4.2.5.6 DDIMs

DDIMs, shares the same network architecture as DDPMs, including the U-Net backbone with skip connections, residual blocks, self-attention, etc. While DDPMs uses a stochastic reverse process that introduces random noise at each step, DDIMs introduces a deterministic reverse process that removes this randomness, which allows the model to generate images in fewer steps [7]. The same U-Net model used for DDPMs is applied for DDIMs sampling, enabling efficient and high-quality generation of grayscale chest X-ray images at a resolution of 64×64 pixels. The input images are normalized to the range [0,1], and the model is trained using the Adam optimizer with a learning rate of $1 \times 10^{-}3$ and a batch size of 64.

4.3 Results and Discussions

This section presents and analyzes the results obtained from training and evaluating the proposed generative models. We compare the performance of GAN-based and diffusion-based models in generating realistic Chest X-ray images using quantitative metrics and qualitative observations.

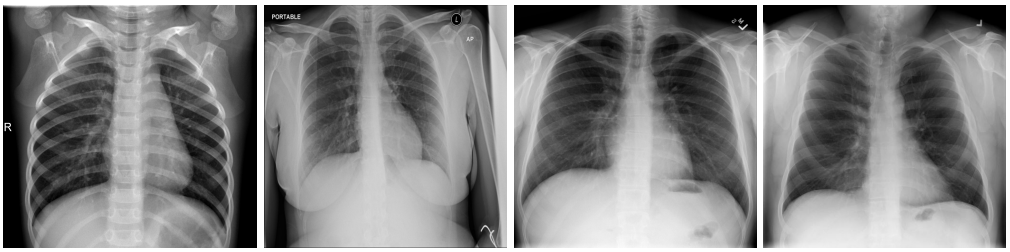
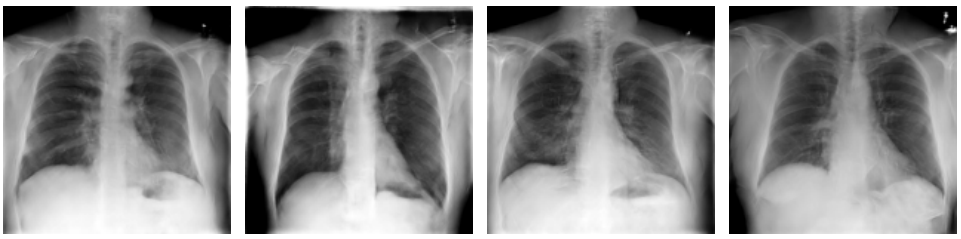
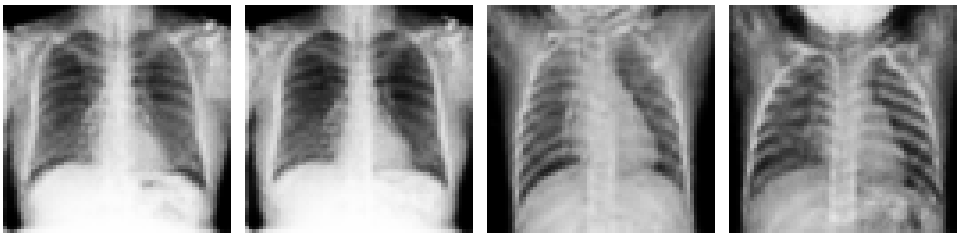
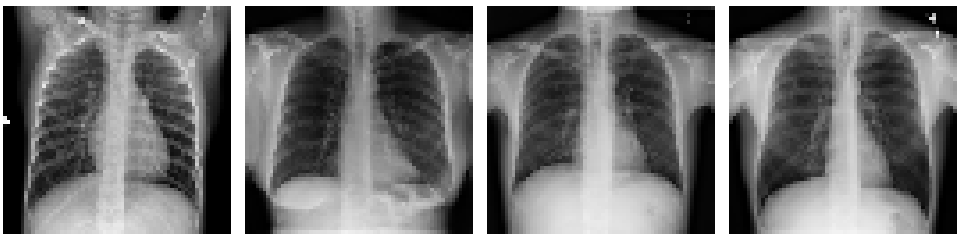
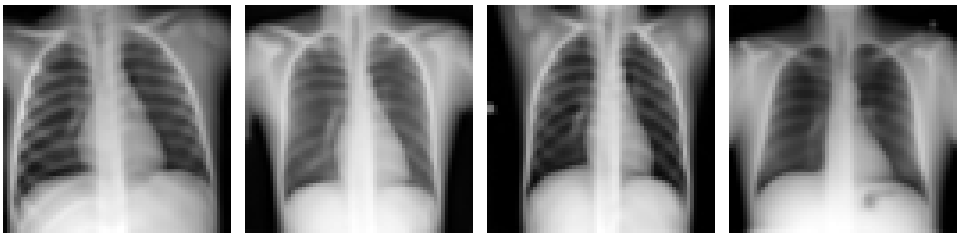
4.3.1 Normal Chest X-Ray Results

The Table 4.2 displays sample outputs from all generative models for Normal Chest X-Ray images.
Models like DCGAN, DDPM and DDIM produced visually convincing results, with well-defined lung fields and smooth transitions in grayscale intensity. Their outputs show a high level of structural consistency with the real normal images, particularly in terms of chest symmetry and rib cage boundaries.
WGAN-GP also performed relatively well at generating realistic chest X-ray images, showing clear and consistent body structures. However, some small problems were noticed at times, like slightly bent or misshaped bones or areas that looked a bit blurry or unnatural.
DCGAN generated images that were visually as good as those produced by DDPM and DDIM in generating normal Chest X-Ray. This means that, even though DCGAN is an older and simpler model, it was still able to produce realistic-looking chest X-ray images similar in quality to those from more advanced diffusion models.

4.3.2 Pneumonia Chest X-Ray Results

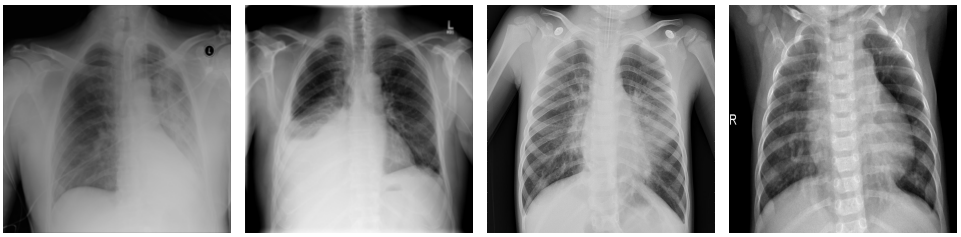
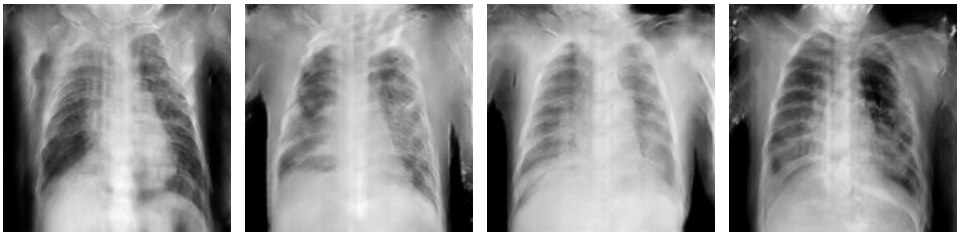
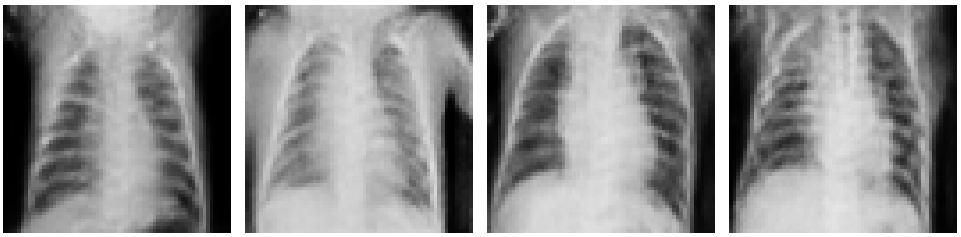
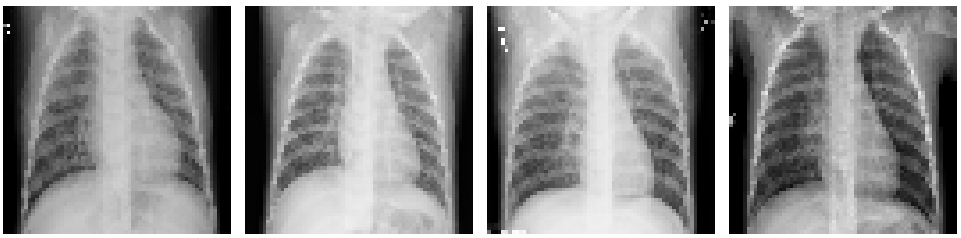
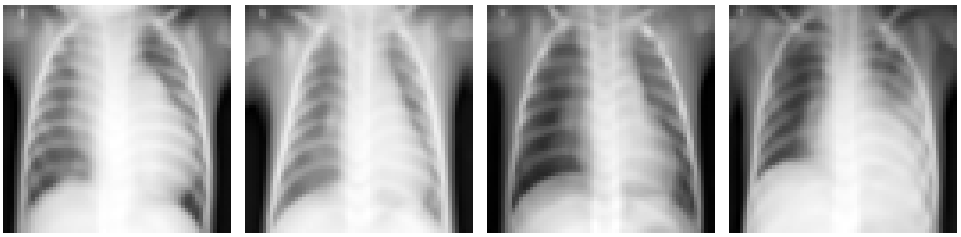
The Table 4.3 displays sample outputs from all generative models for Pneumonia Chest X-Ray images.
When evaluating the generated Pneumonia chest X-ray images, notable differences emerged between the models in terms of visual realism and structure. Based on visual inspection, DDIM produced the most realistic pneumonia images, capturing the complex patterns and

Table 4.2: Generated Normal Chest X-ray samples by different models.

Model	Generated Images
Real images	
DCGAN	
WGAN-GP	
DDPM	
DDIM	

asymmetrical whiteness typically found in infected lungs. Its deterministic reverse process appears effective in preserving fine-grained pathological details, leading to a high-quality output. DDPM followed closely behind, also producing visually convincing images with realistic texture and structure, though some outputs lacked the sharper contrast seen in DDIM's results. Both diffusion models outperformed the GAN-based models, particularly in their ability to reproduce abnormal lung features of Pneumonia cases. Among the GANs, WGAN-GP performed better than expected, generating pneumonia-like textures and maintaining lung structure more consistently than DCGAN. However, occasional blurring and less defined pathological regions

Table 4.3: Generated Pneumonia Chest X-ray samples by different models.

Model	Generated Images
Real images	
DCGAN	
WGAN-GP	
DDPM	
DDIM	

were observed. DCGAN made the least realistic pneumonia images compared to the other models. The images often looked too smooth, with missing details and fake-looking textures that didn't match how real pneumonia usually appears on chest X-rays. These results highlight the increased difficulty of generating pneumonia images compared to normal ones, likely due to the greater variability and complexity of the disease presentation.

4.3.3 FID Results

In this study, we evaluated the generative performance of four models DCGAN, WGAN-GP, DDPM, and DDIM on the combined dataset of chest X-ray images categorized into normal and pneumonia classes. The evaluation metric of choice was the Fréchet Inception Distance (FID), which has become a standard quantitative measure in generative image modeling [78].

The FID metric measures how far apart the feature distributions of real and artificial images are within a high-dimensional embedding space derived from a pre-trained InceptionV3 network. A smaller FID score reflects a closer resemblance between the generated and real images, implying better perceptual quality and variation in the produced samples.

As shown in Table 4.4, the generative models were trained to produce images at a resolution of 64×64 pixels, which is considered relatively low for medical image analysis but is often used as a starting point for evaluating generative models due to its computational efficiency.

Table 4.4: FID Scores of Generated Chest X-ray samples by different models.

Model	**FID Score (Normal)**	**FID Score (Pneumonia)**
DCGAN	163.38	261.64
WGAN-GP	260.76	277.28
DDPM	173.30	183.02
DDIM	152.64	181.97

For both classes, DDIM consistently achieved the lowest FID scores, indicating superior performance in generating realistic chest X-ray images. DCGAN performed moderately well on normal cases but exhibited significantly higher error on pneumonia images. DDPM also demonstrated balanced performance, slightly lagging behind DDIM. WGAN-GP, although theoretically more stable due to the gradient penalty mechanism, resulted in the highest FID values, suggesting difficulties in modeling the underlying data distribution, possibly due to the model architecture or training instability at low resolution.

The main aspect of this evaluation is the influence of image resolution on the reliability and interpretation of FID scores. The InceptionV3 model used for computing FID was originally trained on the ImageNet dataset, where input images are expected to be of size 299×299 pixels. Since our generated images were significantly smaller (64×64), they had to be scaled up to the required input size before FID computation.

This preprocessing step encounters several issues. First, a **Loss of fine-grained features** and diagnostic structures is observed, particularly in X-ray images where subtle texture and edge information are clinically relevant. **Blurring and distortion artifacts** caused by interpolation methods are encountered, which can mislead the feature extractor into interpreting the image as less realistic. Finally, we can confirm that the InceptionV3 network which was not trained on medical images undermines further its sensitivity to mdeical relevant

features.

To validate the effect of resolution on FID, we conducted an additional experiment using DCGAN trained at 224×224 resolution. The resulting FID scores were 130.16 for the normal class and 218.07 for pneumonia, showing a noticeable improvement over the 64×64 setting. These results support the hypothesis that higher resolution leads to more reliable and lower FID scores, despite the generated images sometimes exhibiting visual artifacts or training instabilities at this scale.

This highlights how image resolution plays a critical role in both the perceived quality of generated images and the reliability of quantitative evaluation metrics like FID.

Another important observation is the consistent trend of higher FID scores for pneumonia images across all models. This disparity may be attributed to the increased complexity and variability of pathological patterns in pneumonia cases, which -in our opinion- are harder for generative models to learn, especially with limited data or low resolution. Normal chest X-rays are relatively more homogeneous and therefore easier to model. This suggests that future improvements in pneumonia image synthesis may require class-conditional training, conventional data augmentation, or domain-specific architectural enhancements.

The results goes in the direction the consensus in the literature that diffusion-based models (DDPM and DDIM) tend to outperform traditional GANs in terms of sample quality and training stability, particularly in domains like medical imaging that demand fine detail preservation. DDIM, in particular, benefits from a deterministic sampling procedure that accelerates generation while preserving fidelity.

While GANs are known for sharp image generation, they often suffer from mode collapse (which is not applicable in our case since our solutions are Goal-Specifi), training instability, and sensitivity to hyperparameters, especially in low-data or high-resolution regimes. Diffusion models, on the other hand, incrementally denoise Gaussian noise through a learned reverse diffusion process, allowing them to generate more diverse and realistic samples at the cost of slower inference. However, techniques like DDIM and accelerated samplers have mitigated these drawbacks significantly.

FID, despite being an accepted metric for evaluating generative models, has limitations when applied to medical images. Firstly, the InceptionV3 network, was originally trained on the ImageNet dataset, consisting of diverse natural images. This means its learned features are optimized for recognizing everyday objects and scenes, not for the specific and often subtle patterns found in medical images like X-rays, MRIs, or CT scans. Consequently, InceptionV3 might overlook or misinterpret clinically relevant details, such as subtle lesions, tissue abnormalities, or other diagnostic indicators that a model designed for medical domains would prioritize.

Secondly, FID is a statistical distance that primarily quantifies the similarity between the overall distributions of real and generated image sets. While it reflects perceptual quality and diversity to some extent, it does not directly assess semantic correctness or diagnostic relevance. In medical imaging, it is crucial that generated images are not only visually realistic but also medically accurate and useful for diagnosis. FID does not inherently tell us if a generated tumor resembles a real tumor pathologically, or if the anatomical structures are correctly represented in a clinically meaningful way. This gap means a good FID score doesn't necessarily guarantee clinical utility.

Finally, the FID metric is known to be sensitive to various image processing factors and dataset characteristics. For instance, its value can be significantly affected by changes in image resolution; upscaling generated images to match a target resolution can introduce artifacts that

artificially inflate or deflate the FID score. Furthermore, the metric is sensitive to the statistical properties of the dataset, including potential class imbalances, which are common in medical datasets where certain pathologies are rare. Such sensitivities can lead to misleading FID scores that do not accurately reflect the true quality or clinical utility of the generated medical images.

4.3.4 VTT Results

While FID gave us a crucial quantitative measure, it was worth important to bring in a human perspective. That's why we conducted a Visual Turing Test (VTT) to truly gauge the perceptual realism and clinical plausibility of our generated chest X-ray images. This qualitative approach is a recognized standard in the evaluation of generative models, particularly in the sensitive domain of medical imaging, where visual fidelity and domain-specific insights are absolutely critical.

To carry out this essential evaluation, we requested the expertise of a licensed medical doctor who possessed good professional experience, having been specifically trained to interpret chest X-rays during the intense period of the COVID-19 pandemic. During that time, chest radiography was a widely used tool for diagnosis.

Our expert was then presented with a randomized selection of chest X-ray images. Crucially, they had no idea whether any given image was real or synthetic, ensuring an unbiased assessment. Their task was straightforward but accurate:

- Could these images pass as authentic chest X-rays?
- Did they show anatomical consistency, meaning proper lung boundaries, convincing rib structures, and accurate diaphragm outlines?
- Were there any obvious flaws or unrealistic patterns that would immediately undermine their diagnostic value?

Despite the admittedly low resolution, the doctor reported that many of the generated images were generally acceptable (although the resolution is far from sufficient to make a good diagnosis). Many managed to preserve the overall anatomical layout, showing good lung symmetry, appropriate contrast between thoracic regions, and even recognizable structures like clavicles and the subtle shadows of the heart.

However, the assessment wasn't without its caveats. The expert also highlighted the presence of unrealistic artifacts in some samples. These included:

- Rib cage outlines that were irregular or incomplete,
- Lung textures that appeared blurry and lacked the fine pathological detail seen in real cases,
- Anatomical distortions, such as a misplaced mediastinum or noticeable asymmetry in the lungs,
- And in a few instances, "melted" textures, a common tell-tale sign of GAN failures or simply artifacts stemming from the low resolution.

These insightful observations resonated well with our quantitative findings. While lower FID scores (especially from our diffusion models) did suggest improved realism, the VTT served as a powerful reminder that synthetic images aren't uniformly perfect. Their fidelity, as the expert noted, clearly varies across individual samples and different image classes.

Ultimately, the expert concluded that while these low-resolution images aren't yet ready for direct diagnostic use, they hold significant promise for us, like augmenting limited real datasets or supporting semi-supervised learning endeavors.

This qualitative assessment truly underscores the critical importance of involving human experts when working with generative models in the medical field. Unlike more conventional image domains, medical images demand an incredibly strict adherence to anatomical correctness and clinical interpretability, nuances that automated metrics alone often can't fully capture. Moving forward, our plans include integrating higher-resolution models, implementing refined architectural constraints, and embedding human-in-the-loop fine-tuning more deeply into our process to further enhance both the visual and clinical plausibility of our generated medical images.

4.3.5 Additional Experiments with ProGAN and StyleGAN3

In addition to the main set of experiments, two separate evaluations were conducted using the ProGAN and StyleGAN3 models.

The first ProGAN experiment was carried out on the Kaggle platform, where training was limited to only 30 epochs due to strict runtime constraints. Because of its progressively growing architecture, ProGAN requires significantly more time to train, even at just 30 epochs, it took approximately 9 hours to complete. Despite the limited training time, ProGAN was still able to generate visually good Chest X-Ray images, particularly in the Pneumonia class, as shown in the corresponding Fig 4.1.

A second ProGAN experiment was later performed under the same conditions as the other models (trained for 500 epochs). However, during training, the model began to produce NaN (Not a Number) loss values starting at epoch 211 and subsequently generated only black images. This behavior suggests instability in the training process, potentially might be caused by architectural complexity, learning rate issues, or high number of epochs.

As for StyleGAN3, the model is currently being trained and evaluated, but no final results are available at this stage. More details will be provided once training and testing are complete.

Figure 4.1: ProGAN Pneumonia Generated Images.

4.4 Conclusion

This chapter summarized our experimental approach and key findings using GANs and diffusion models for chest X-ray image generation, through quantitative and qualitative evaluations. In the next chapter, we present the general conclusion of this work, discuss its limitations, and propose future perspectives.

CHAPTER 5

CONCLUSION AND FUTURE PERSPECTIVES

Contents

5.1 General conclusion

In this thesis, we addressed the critical challenge of data scarcity in medical image analysis, particularly for chest X-rays, by exploring the potential of advanced generative models including GANs and diffusion models to generate realistic synthetic medical images. Through the implementation and comparison of four different models (DCGAN, WGAN-GP, DDPM, and DDIM), we evaluated their ability to produce high-quality 64×64 chest X-ray images, using both quantitative (FID) and qualitative (VTT) metrics. But both StyleGAN3 and ProGAN remain at a preliminary stage in our experiments, with only initial results available for ProGAN at this time.

Our findings demonstrate that despite limited resolution and computational constraints, certain models such as DDPM, DDIM, and even DCGAN achieved promising results in generating realistic samples. These results not only confirm the viability of generative models for medical image augmentation but also highlight the potential of simple architectures when carefully trained.

However, this study also has limitations, and the evaluation was limited to a small set of metrics and visual inspection by a single radiologist. Future research could benefit from using higher resolution images, more diverse and clinically annotated datasets, and more robust evaluation protocols involving multiple medical experts.

Moving forward, we envision expanding this work by integrating more advanced architectures, exploring hybrid approaches combining GANs and diffusion models, and applying these methods to other imaging modalities. The ultimate goal remains the same: to enhance data availability in medical imaging and support the development of more accurate and generalizable AI models for clinical use.

5.2 Limitations

Despite the contributions of this study, several limitations should be noted. First, training some of the models required a considerable amount of time, sometimes taking over three days, making it challenging to conduct multiple trials or longer training runs. This was further complicated by unstable electricity conditions at the place, where the primary workstation was located.
Additionally, all generated images were limited to a resolution of 64×64, whereas the FID metric evaluates image similarity using a pre-trained InceptionV3 model that expects inputs of 299×299, upscaling was necessary but that led to high FID scores.
Furthermore, the evaluation relied mainly on FID scores and the Visual Turing Test (VTT), other metrics such as Structural Similarity Index (SSIM) and Kernel Inception Distance (KID) were not included.
Lastly, the study focused exclusively on grayscale Chest X-Ray images, and didn't include other medical imaging types such as CT scans or MRI.

5.3 Future Perspectives

Building on the findings of this work, there are several ways we can move forward to improve and expand the results.
First, additional generative models including more advanced variants of GANs and diffusion models can be explored and adapted to synthesize not only chest X-ray images but also other types of medical images such as CT scans, ultrasound and also MRIs.
Increasing the number of training epochs and carefully tuning hyperparameters for each model may also improve performance and stability, especially for complex architectures like ProGAN and StyleGAN3. Furthermore, a detailed evaluation of data augmentation strategies can be performed by training a convolutional neural networks (CNNs) classifier on both augmented and non-augmented datasets, in order to assess whether synthetic data improves classification accuracy and model robustness.
Finally, including additional evaluation metrics such as SSIM, and KID would provide a more comprehensive assessment of image quality and realism.

BIBLIOGRAPHY

[1] What is artificial intelligence? https://www.nasa.gov/what-is-artificial-intelligence/. Accessed: 9 May 2025.

[2] Hazrat Ali, Shafaq Murad, and Zubair Shah. Spot the fake lungs: Generating synthetic medical images using neural diffusion models. https://arxiv.org/pdf/2211.00902, 2022.

[3] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. https://arxiv.org/pdf/1511.06434, 2015.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. https://arxiv.org/pdf/1710.10196, 2017.

[5] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. https://arxiv.org/pdf/2106.12423, 2021.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. https://arxiv.org/pdf/2006.11239, 2020.

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. https://arxiv.org/pdf/2010.02502, 2020.

[8] Bull. Acad. Natle Méd. What is a medical image? medical-economic considerations. https://www.academie-medecine.fr/quest-ce-quune-image-medicale-considerations-medico-economiques/. Accessed: 19 May 2025.

[9] S. K. M Shadekul Islam, MD Abdullah Al Nasim, Ismail Hossain, Md Azim Ullah, Kishor Datta Gupta, and Md Monjur Hossain Bhuiyan. Introduction of medical imaging modalities. https://arxiv.org/pdf/2306.01022, 2023.

[10] Manuel Cossio. Augmenting medical imaging: A comprehensive catalogue of 65 techniques for enhanced data analysis. https://arxiv.org/pdf/2303.01178, 2023.

[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. https://www.kaggle.com/datasets/nih-chest-xrays/data. Accessed: 19 May 2025.

[12] Chest x-ray images (pneumonia). https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia. Accessed: 19 May 2025.

[13] Iowa State University. History of radiography. https://www.nde-ed.org/NDETechniques/Radiography/Introduction/history.xhtml. Accessed: 8 April 2025.

[14] Mayo Clinic. Chest x-rays. https://www.mayoclinic.org/tests-procedures/chest-x-rays/about/pac-20393494. Accessed: 8 April 2025.

[15] Organisation mondiale de la Santé. Childhood pneumonia. https://www.who.int/fr/news-room/fact-sheets/detail/pneumonia. Accessed: 21 june 2025.

[16] Mayo Clinic. Pneumonia. https://www.mayoclinic.org/diseases-conditions/pneumonia/symptoms-causes/syc-20354204. Accessed: 21 june 2025.

[17] T. Franquet. Imaging of pneumonia: trends and algorithms. https://publications.ersnet.org/content/erj/18/1/196. Accessed: 21 june 2025.

[18] Heang-Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. In *Deep Learning in Medical Image Analysis : Challenges and Applications*, pages 3–21. Springer International Publishing, 2020.

[19] MathWorks. Medical image analysis. https://au.mathworks.com/discovery/medical-image-analysis.html. Accessed: 8 April 2025.

[20] DataCamp. Synthetic data generation: A hands-on guide in python. https://www.datacamp.com/tutorial/synthetic-data-generation. Accessed: 3 April 2025.

[21] Cole Stryker and Eda Kavlakoglu. What is artificial intelligence (ai)? https://www.ibm.com/think/topics/artificial-intelligence. Accessed: 9 May 2025.

[22] Coursera Staff. What is artificial intelligence? definition, uses, and types. https://www.coursera.org/articles/what-is-artificial-intelligence. Accessed: 9 May 2025.

[23] Michela Pellicelli. Chapter five - managing the supply chain: technologies for digitalization solutions. In Michela Pellicelli, editor, *The Digital Transformation of Supply Chain Management*, pages 101–152. Elsevier, 2023.

[24] What is machine learning? https://www.ibm.com/think/topics/machine-learning. Accessed: 9 May 2025.

[25] Coursera Staff. Supervised vs. unsupervised learning: Pros, cons, and when to choose. https://www.coursera.org/articles/supervised-vs-unsupervised-learning. Accessed: 9 May 2025.

[26] Jim Holdsworth and Mark Scapicchio. What is deep learning? https://www.ibm.com/fr-fr/think/topics/deep-learning. Accessed: 8 May 2025.

[27] Coursera Staff. What is deep learning? definition, examples, and careers. https://www.coursera.org/articles/what-is-deep-learning. Accessed: 8 May 2025.

[28] Jianqing Fan, Cong Ma, and Yiqiao Zhong. A selective overview of deep learning. https://pmc.ncbi.nlm.nih.gov/articles/PMC8300482/. Accessed: 8 May 2025.

[29] Introduction to convolution neural network. https://www.geeksforgeeks.org/introduction-convolution-neural-network/. Accessed: 18 May 2025.

[30] What are convolutional neural networks? https://www.ibm.com/think/topics/convolutional-neural-networks. Accessed: 18 May 2025.

[31] DataCamp. The complete guide to data augmentation. https://www.datacamp.com/fr/tutorial/complete-guide-data-augmentation. Accessed: 5 April 2025.

[32] Medium. Role of data augmentation in medical image analysis. https://medium.com/@mohsincsv/role-of-data-augmentation-in-medical-image-analysis-aa4cbc1e2426. Accessed: 5 April 2025.

[33] Pouya Hallaj. Data augmentation: Benefits and disadvantages. https://medium.com/@pouyahallaj/data-augmentation-benefits-and-disadvantages-38d8201aead. Accessed: 18 May 2025.

[34] Ivan Belcic. What is a generative model? https://www.ibm.com/think/topics/generative-model. Accessed: 8 May 2025.

[35] Oyedele Tioluwani. How do generative models work in deep learning? generative models for data augmentation explained. https://www.freecodecamp.org/news/generative-models-for-data-augmentation/. Accessed: 8 May 2025.

[36] Cole Stryker and Mark Scapicchio. What is generative ai? https://www.ibm.com/think/topics/generative-ai. Accessed: 19 May 2025.

[37] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. https://arxiv.org/pdf/1406.2661, 2014.

[38] Marco Del Pra. Generative adversarial networks. https://medium.com/@marcodelpra/generative-adversarial-networks-dba10e1b4424. Accessed: 19 May 2025.

[39] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. https://arxiv.org/pdf/2105.05233, 2021.

[40] AssemblyAI. Introduction to diffusion models for machine learning. https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction. Accessed: 5 April 2025.

[41] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: A review, 2023.

[42] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. https://arxiv.org/pdf/1411.1784, 2014.

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. https://arxiv.org/pdf/1703.10593, 2017.

[44] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. https://arxiv.org/pdf/1812.04948, 2018.

[45] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. https://arxiv.org/pdf/1704.00028, 2017.

[46] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. https://arxiv.org/pdf/1606.00709, 2016.

[47] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. https://arxiv.org/pdf/1701.07875, 2017.

[48] Muhammad Ubale Kirua, Bahari Belatona, Xinying Chewa, Khaled H. Almotairib, Ahmad MohdAziz Husseinc, and Maryam Aminu. Comparative analysis of some selected generative adversarial network models for image augmentation: a case study of covid-19 x-ray and ct images, 2022.

[49] Mai Feng Ng and Carol Anne Hargreaves. Generative adversarial networks for the synthesis of chest x-ray images. *Engineering Proceedings*, 2023.

[50] Bradley Segal, David M. Rubin, Grace Rubin, and Adam Pantanowitz. Evaluating the clinical realism of synthetic chest x-rays generated using progressively growing gans. https://arxiv.org/pdf/2010.03975, 2020.

[51] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. https://arxiv.org/pdf/1705.02315, 2017.

[52] Miso Jang, Hyun jin Bae, Minjee Kim, Seo Young Park, A yeon Son, Se Jin Choi, Jooae Choe, Hye Young Choi, Hye Jeon Hwang, Han Na Noh, Joon Beom Seo, Sang Min Lee, and Namkug Kim. Image turing test and its applications on synthetic chest radiographs by using the progressive growing generative adversarial network. *Scientific Reports*, 2023.

[53] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021.

[54] NVIDIA Research. Stylegan3: Official github repository. https://github.com/NVlabs/stylegan3, 2021.

[55] John R. McNulty, Lee Kho, Alexandria L. Case, Charlie Fornaca, Drew Johnston, David Slater, Joshua M. Abzug, and Sybil A. Russell. Synthetic medical imaging generation with generative adversarial networks for plain radiographs, 2024.

[56] Min Jin Chong and David A Forsyth. Analyzing and improving the image quality of stylegan. https://arxiv.org/pdf/1912.04958, 2020.

[57] Dave Bergmann and Cole Stryker. What are diffusion models? https://www.ibm.com/think/topics/diffusion-models. Accessed: 28 june 2025.

[58] Milvus. What is denoising diffusion probabilistic modeling (ddpm)? https://milvus.io/ai-quick-reference/what-is-denoising-diffusion-probabilistic-modeling-ddpm. Accessed: 31 May 2025.

[59] Iman Khazrak, Shakhnoza Takhirova, Mostafa M. Rezaee, Mehrdad Yadollahi, Robert C. Green II, and Shuteng Niu. Addressing small and imbalanced medical image datasets using generative models: A comparative study of ddpm and pggans with random and greedy k sampling. https://arxiv.org/pdf/2412.12532, 2024.

[60] Thulana Abeywardane, Tomson George, and Praveen Mahaulpatha. Ddpm-based x-ray image synthesizer. https://arxiv.org/pdf/2401.01539, 2024.

[61] Bhomik Sharma. Ddim: The faster, improved version of ddpm for efficient ai image generation. https://learnopencv.com/understanding-ddim/. Accessed: 10 June 2025.

[62] Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. https://arxiv.org/pdf/2308.02062, 2023.

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[64] Yilun Feng, Yujie Peng, Yuxuan Zhang, Xiaoxiao Luo, Yifan Yang, Yao Zhao, Matthew P Lungren, Chiyuan Zhang, Andrew Y Ng, and Pranav Rajpurkar. Medfusion: Text-to-image diffusion model for visualizing medical concepts. https://arxiv.org/pdf/2210.04133, 2022.

[65] Parisa Ashtari, Farhad Pourpanah, and Negar Ghavami. Synthetic chest x-ray generation using stable diffusion: A step toward improved disease classification. *Informatics*, 12(9):173, 2023.

[66] Mohammed Aljanabi, Salman Khan, et al. Evaluating dall-e 2 in medical image generation: Promises and limitations. *IEEE Access*, 2024. Accessed via https://ieeexplore.ieee.org/document/10440330.

[67] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. https://arxiv.org/pdf/2204.06125, 2022.

[68] Chinedu Innocent Nwoye, Rupak Bose, Kareem Elgohary, Lorenzo Arboit, Giorgio Carlino, Joël L. Lavanchy, Pietro Mascagni, and Nicolas Padoy. Surgical text-to-image generation. https://arxiv.org/pdf/2407.09230, 2024.

[69] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic

text-to-image diffusion models with deep language understanding. https://arxiv.org/pdf/2205.11487, 2022.

[70] M. Puttagunta, R. Subban, and NKB C. A novel covid-19 detection model based on dcgan and deep transfer learning. *Procedia Computer Science*, 204:65–72, 2022. Epub 2022 Sep 10, PMID: 36120410, PMCID: PMC9464299.

[71] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification. https://doi.org/10.17632/rscbjbr9sj.2, 2018.

[72] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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 (CVPR)*, pages 3462–3471, 2017. NIH Clinical Center dataset, https://nihcc.app.box.com/v/ChestXray-NIHCC.

[73] Rifatul Islam Majumder. Efficient classification of pulmonary pneumonia and tuberculosis alongside normal and non-x-ray images with minimal resources and maximum accuracy. medRxiv, 2024. Preprint. DOI: https://doi.org/10.1101/2024.12.31.24319820.

[74] Tawsifur Rahman, Amith Khandakar, Muhammad A. Kadir, Khandaker R. Islam, Khandaker F. Islam, Zaid B. Mahbub, Mohamed Arselene Ayari, and Muhammad E. H. Chowdhury. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.

[75] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. https://doi.org/10.17632/rscbjbr9sj.2, 2018.

[76] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. https://arxiv.org/pdf/1706.08500, 2017.

[77] Ayush Thakur. Comment évaluer les gan en utilisant la fréchet inception distance (fid) ? https://wandb.ai/wandb_fc/french/reports/Comment-valuer-les-GAN-en-utilisant-la-Fr-chet-Inception-Distance-FID---Vmlldzo0NjA Accessed: 13 june 2025.

[78] Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing flows are capable generative models, 2025.