Generation of Synthetic COVID-19 Chest X-ray Images and Evaluation of Their Realism and Diagnostic Consistency

Jiajia Xu (s3845567)

Supervisor:
Joost Batenburg & M.C.R. Go

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

June 18, 2026

Abstract

Generative artificial intelligence has shown increasing potential in medical image synthesis, particularly for data augmentation and pathology-aware image generation tasks. Among recent generative approaches, Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs) have demonstrated promising performance in medical imaging applications. However, the influence of external classifier-guided supervision on different generative paradigms remains insufficiently explored, especially in chest X-ray image generation for COVID-19-related tasks.

This study investigates the effects of classifier-guided supervision on chest X-ray image synthesis using two generative frameworks: Auxiliary Classifier Generative Adversarial Networks (ACGANs) and Denoising Diffusion Probabilistic Models (DDPMs). Experiments were conducted using publicly available COVID-19 and normal chest X-ray images. An external ResNet-18 classifier was incorporated to guide the generation process and improve the consistency of class discrimination in synthetic images.

The experimental results showed that the two generative models responded to classifier-guided supervision in different ways. The baseline ACGAN framework already achieved strong class-consistent generation due to its internal auxiliary classification objective, and external supervision provided only limited additional improvement. In contrast, the baseline DDPM framework generated visually smooth and anatomically coherent chest X-ray images but exhibited weak pathological discrimination capability. After introducing classifier-guided supervision, the DDPM framework achieved dramatically improved classification accuracy; however, this improvement was accompanied by degraded visual fidelity, increased Fréchet Inception Distance (FID) scores, and occasional generation instability.

The findings suggest that classifier-guided supervision introduces a significant trade-off between pathological discrimination and generative realism in diffusion-based medical image generation. While classifier guidance can enhance class-discriminative generation, excessive classifier-oriented optimization may interfere with realism-oriented denoising and reconstruction, reducing overall medical image fidelity. These observations provide useful insights for future controllable medical image synthesis and pathology-aware generative modeling research.

Contents

1 Introduction

1.1 Background

Chest X-ray imaging has played a central role in the diagnosis and monitoring of respiratory diseases, including COVID-19[1]. However, the effectiveness of data-driven approaches in this domain is often constrained by the limited availability of high-quality annotated data. This limitation becomes particularly obvious in the cases of rare diseases, where the number of available samples is originally small, while privacy concerns and regulatory restrictions further hinder the collection and sharing of medical imaging data, making it difficult to assemble sufficiently large and diverse datasets for training machine learning models.

To address these challenges, the generation of synthetic medical images has emerged as a promising direction[2]. By augmenting existing datasets with artificially generated samples, it becomes possible to improve data diversity, mitigate class imbalance, and support the training of more robust classifiers. As a result, attention has been devoted to developing generative models capable of producing realistic medical images, with approaches such as GANs[3] and diffusion models[4] demonstrating strong performance in visual image synthesis.

Nevertheless, generating visually convincing images alone is not sufficient for medical applications. While realism remains an important aspect, the ultimate goal of synthetic data generation is to produce samples that are not only plausible but also useful for downstream tasks. This introduces additional challenges in the design and evaluation of generative models, as improvements in image quality do not necessarily translate into better performance in practical settings.

Therefore, in the context of medical image synthesis, it becomes essential to investigate not only how to generate high-quality images, but also how different generative approaches influence the usefulness of the generated data. This motivates a systematic exploration of generative models and training strategies, together with appropriate evaluation methods that can capture both visual quality and task-related relevance.

1.2 Research Objectives

Building upon the challenges outlined above, this thesis focuses on improving the generation of synthetic chest X-ray images through the integration of classifier-guided supervision. Rather than solely relying on conventional generative objectives, the proposed approach investigates how task-related signals can be incorporated into the training process to influence the behavior of generative models.

To this end, the study first establishes baseline generative performance using two representative model families, namely ACGAN and diffusion-based models. These models serve as reference points for analyzing the characteristics and limitations of different generative approaches in the context of medical image synthesis.

Subsequently, classifier-guided supervision is introduced as an additional training mechanism. With the feedback from a trained classifier, the generative process is guided towards capturing class-specific patterns, allowing for a comparative analysis of its impact across model architectures.

Finally, the effect of this guidance strategy is systematically evaluated. The analysis considers both the visual quality of the generated images and their effectiveness in downstream classification tasks, providing insight into how classifier supervision influences different aspects of synthetic data generation.

1.3 Research Problem

The central problem addressed in this thesis is how different generative models can be used to produce synthetic chest X-ray images that are both visually plausible and useful for downstream tasks. While recent advances in generative modeling have enabled the creation of high-quality images, it remains unclear how different model architectures and training strategies influence the characteristics of the generated data.

In particular, this study examines two representative generative approaches, namely ACGAN and diffusion-based models, and investigates how their underlying mechanisms affect the quality and usefulness of synthetic medical images. In addition, classifier-guided supervision is explored as a complementary training strategy, allowing for an analysis of how additional task-related feedback may influence the generation process under different model frameworks.

To structure this investigation, the following research questions are considered:

- How do ACGAN and diffusion models compare in their baseline ability to generate synthetic chest X-ray images?
- What are the differences in the visual quality and task-related usefulness of images generated by these models?
- How does the inclusion of classifier-guided supervision influence the generation results within each model framework?

1.4 Thesis overview

The remainder of this thesis is organized as follows.

Section 2 introduces the necessary background on chest X-ray imaging and generative models.

Section 4 reviews related work in medical image synthesis.

Section 5 describes the methodology, including the generative models and the incorporation of classifier-guided supervision.

Section 6 presents the experimental setup and results, followed by a discussion of the findings.

Finally, Section 7 concludes the thesis and outlines directions for future work.

This work forms part of a Bachelor thesis conducted at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, under the supervision of Joost Batenburg.

2 Background

2.1 Chest X-ray Imaging and COVID-19 Diagnosis

Chest X-ray imaging is widely used as a primary diagnostic tool for detecting and monitoring respiratory diseases due to its accessibility, low cost, and rapid acquisition time [1]. During the COVID-19 pandemic, chest X-ray images have been extensively utilized to support the identification of infected patients, particularly in settings where more advanced imaging techniques such as CT scans are not readily available.

From a clinical perspective, chest X-ray images of healthy individuals typically exhibit clear lung fields with well-defined anatomical structures. In contrast, various respiratory infections introduce abnormal patterns in the lungs. For instance, bacterial pneumonia often appears as localized consolidations, while viral pneumonia tends to produce more diffuse and patchy opacities. COVID-19 pneumonia is frequently associated with ground-glass opacities and bilateral infiltrates, which manifest as hazy regions distributed across the lung area [5]. These visual differences, although sometimes subtle, are critical for both human diagnosis and automated classification systems.

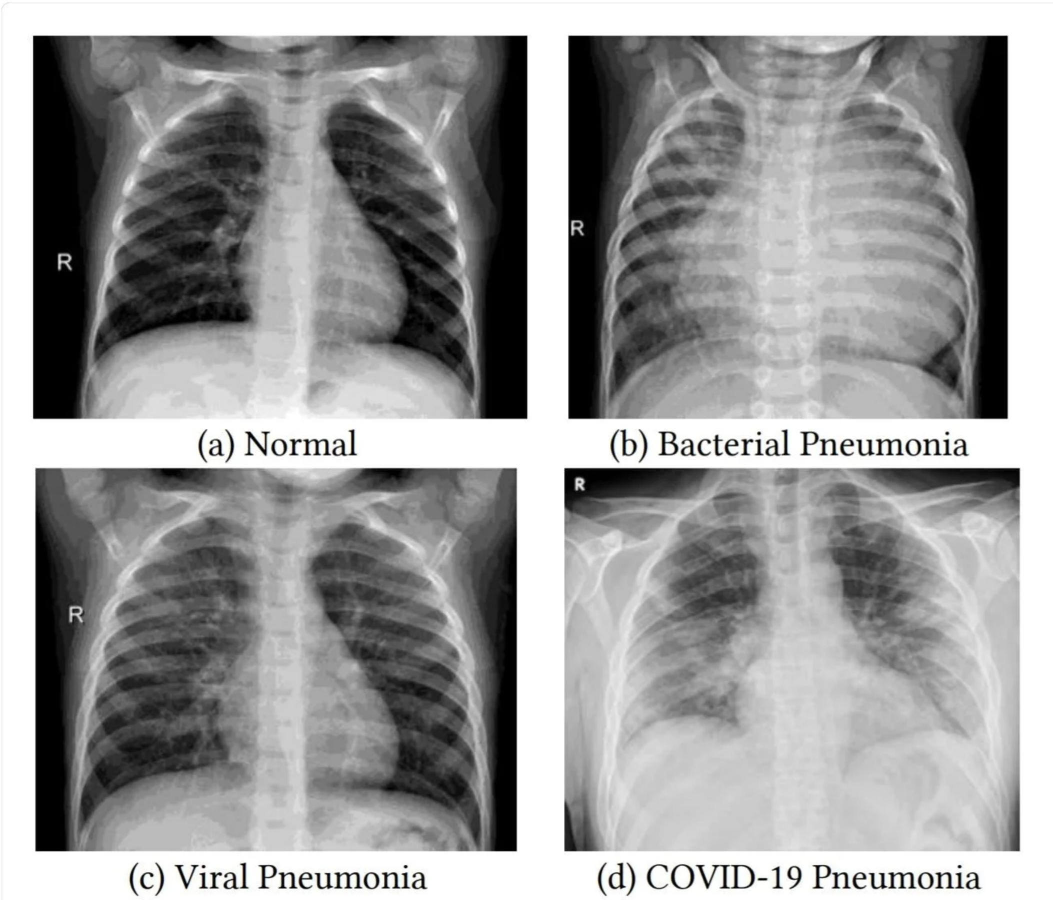

Figure 1: Examples of chest X-ray images showing different conditions: (a) normal, (b) bacterial pneumonia, (c) viral pneumonia, and (d) COVID-19 pneumonia. The images illustrate the variation in lung patterns, with increasing presence of opacities and irregular structures in infected cases.

Chest X-ray images are commonly used to distinguish between normal and pathological cases, including COVID-19. However, the variety of patterns observed across different respiratory conditions makes chest X-ray interpretation challenging. These patterns form the basis for machine learning approaches to disease detection and emphasize the importance of preserving diagnostically relevant features when generating synthetic images. If such features are not accurately captured, the generated data may appear visually plausible while lacking practical value for downstream classification tasks.

2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs), introduced by Goodfellow et al., provide a framework for learning data distributions through an adversarial training process [6]. A GAN consists of two components: a generator G, which maps a latent noise vector z sampled from a prior distribution to the data space, and a discriminator D, which aims to distinguish between real samples and those generated by G. The two networks are trained simultaneously in a minimax game, where the generator seeks to produce realistic samples that can fool the discriminator, while the discriminator attempts to correctly classify real and generated data.

Formally, the training objective can be expressed as:

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

Despite their strong performance in generating high-quality samples, standard GANs provide limited control over the generation process, as the mapping from noise to data is unconstrained. Consequently, it is difficult to guide the model to generate samples associated with specific classes or desired semantic attributes.

To address this limitation, conditional Generative Adversarial Networks (cGANs) extend the GAN framework by incorporating auxiliary information into both the generator and the discriminator [7]. As illustrated in Fig. 2, the conditioning variable is provided as input to both components. Specifically, an additional variable y, such as class labels, is provided as input to both G and D, enabling the model to learn a conditional distribution rather than an unconditional one. In this setting, the generator produces samples conditioned on y, while the discriminator evaluates whether a sample is real given the same condition.

The objective function of a conditional GAN is formulated as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|y)))]. \tag{2}$$

By incorporating conditioning information, cGANs provide an effective mechanism for controlling the generation process, which is particularly relevant for class-dependent image synthesis tasks and serves as a foundation for more advanced conditional generative models.

2.2.1 Auxiliary Classifier GAN

While cGANs introduce conditioning into the generation process, the control they provide remains implicit, as the model is not explicitly encouraged to generate samples that are class-consistent. To address this limitation, Auxiliary Classifier GANs (ACGANs) extend the conditional GAN framework by incorporating an explicit classification objective into the discriminator [8]. Specifically, in addition to predicting whether a sample is real or fake, the discriminator is also trained to predict the class label associated with each input.

This modification effectively introduces a second supervision signal. During training, the generator is encouraged not only to produce realistic samples that can fool the discriminator, but also to generate images that can be correctly classified. As a result, the model is more strongly guided to capture class-specific features, leading to improved class consistency in the generated samples.

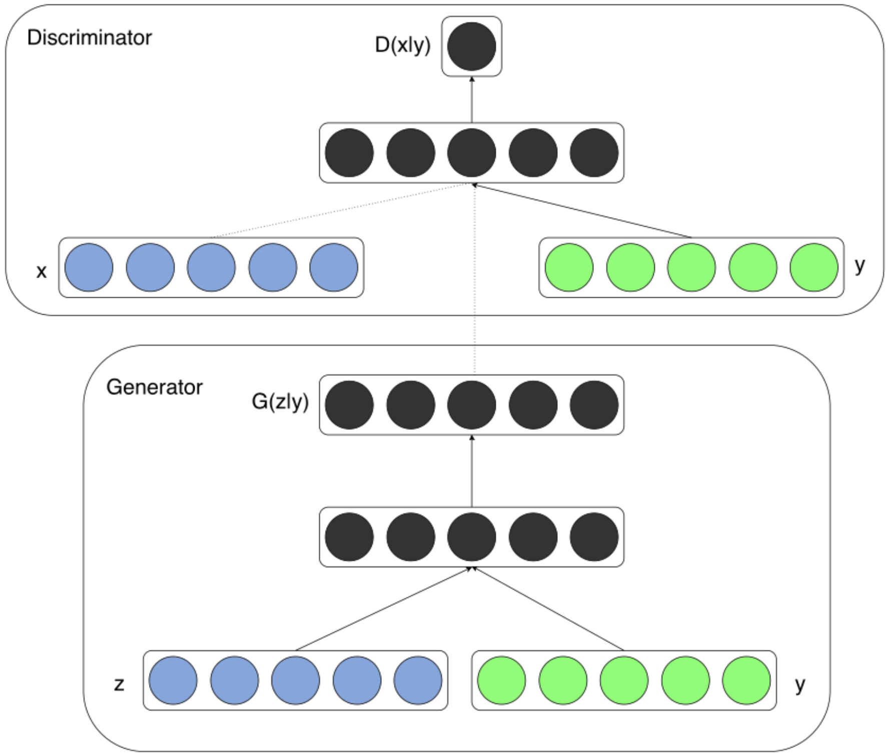

Figure 2: Conditional adversarial net

The training objective can therefore be interpreted as a combination of an adversarial loss and a classification loss:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{cls}. \tag{3}$$

Although this formulation enhances control over the generation process, it also introduces an additional constraint. In particular, the model must simultaneously optimize for visual realism and classification accuracy, which may not always align. This can lead to competing objectives during training, where improving performance on one objective may negatively affect the other.

2.3 Diffusion Models and Denoising Diffusion Probabilistic Models

Diffusion models have recently emerged as an alternative class of generative models that produce high-quality samples through a gradual denoising process [9, 10]. Instead of directly mapping a latent vector to the data space as in GANs, diffusion models learn to generate data by reversing a stochastic corruption process.

The forward process progressively adds noise to a data sample over a series of time steps until it is transformed into nearly pure Gaussian noise. This process is fixed and does not require learning. The core task of the model is then to learn the reverse process, which reconstructs the original

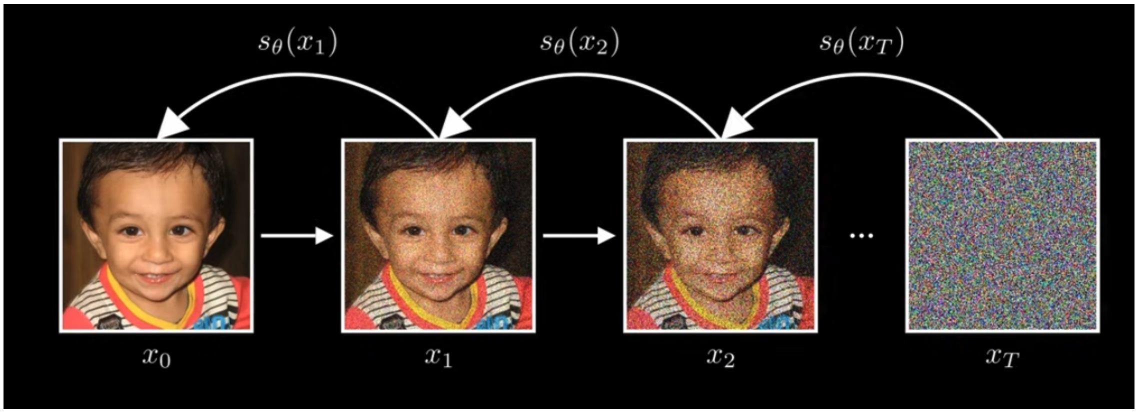

Figure 3: Illustration of the forward diffusion process. A clean image x_0 is gradually corrupted by adding noise over multiple steps until it becomes nearly pure noise x_T. The reverse process learns to recover the original image by removing noise step by step.

data by iteratively removing noise, as illustrated in 3. Denoising Diffusion Probabilistic Models (DDPMs) provide a practical formulation of this idea [10], and several subsequent works have further improved their efficiency and sample quality [11]. During training, the model is given a noisy sample at a particular time step and is trained to predict the noise component that was added. This objective can be formulated as:

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \tag{4}$$

where ϵ_θ denotes the predicted noise. By learning to estimate this noise, the model implicitly learns how to reverse the diffusion process.

At generation time, new samples are produced by starting from random noise and repeatedly applying the learned denoising steps. Each step slightly reduces the noise level, gradually transforming the sample into a structured image. This iterative procedure leads to stable training and often results in high-quality outputs.

However, unlike GAN-based models, diffusion models do not naturally incorporate conditioning mechanisms. As a result, controlling the generation process is less straightforward and often requires additional techniques, such as guidance methods applied during sampling. Since these methods intervene in the iterative denoising process, they may alter the balance of the generation dynamics and affect the quality of the final samples.

2.4 Training of Generative Models

Generative models are typically trained by optimizing an objective that encourages the model to capture the underlying data distribution. However, different models adopt different training strategies.

Generative Adversarial Networks (GANs) are trained through an adversarial process in which a generator and a discriminator are optimized simultaneously. The generator aims to produce samples that resemble real data, while the discriminator attempts to distinguish between real and generated samples. This leads to a minimax optimization problem, where the two networks are updated in opposition to each other. While this framework enables the generation of sharp and realistic images, it is also known to be sensitive to training instability and requires careful balancing between the two components.

In contrast, diffusion models are trained using a denoising objective. Instead of directly generating data, the model learns to reverse a predefined noise corruption process. At each training step, the model receives a noisy version of a data sample and is trained to predict the added noise. This formulation results in a stable and well-defined optimization objective, as it does not rely on adversarial interactions between networks.

These differences in training mechanisms lead to distinct characteristics in practice. Adversarial training encourages realism through competition, while diffusion-based training focuses on accurately modeling the data distribution through iterative refinement.

As a result, the way additional supervision signals are incorporated may have different effects depending on the underlying training paradigm.

3 Evaluation Metrics

Evaluating generative models requires considering not only the visual quality of the generated samples but also their usefulness for downstream tasks. In this work, two complementary metrics are adopted to capture these aspects: classification accuracy and Fréchet Inception Distance (FID).

To assess the consistency of the diagnosis of generated images, classification accuracy is employed. A pretrained external ResNet-18 classifier is applied to the generated images, and the predicted labels are compared with the intended generation labels. Classification accuracy is computed as

$$Accuracy = \frac{N_{\text{correct}}}{N_{\text{total}}}, \tag{5}$$

where N_{correct} denotes the number of correctly classified generated images and N_{total} denotes the total number of evaluated generated images.

In addition to task performance, the Fréchet Inception Distance (FID) [12] is used to measure the similarity between the distributions of generated and real images. FID compares feature representations extracted from a pretrained Inception-v3 network and is defined as

$$FID = \|\mu_r - \mu_g\|^2 + Tr\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{6}$$

where μ_r and Σ_r represent the mean vector and covariance matrix of features extracted from real images, and μ_g and Σ_g denote the corresponding statistics of generated images.

Lower FID values indicate that the generated samples are closer to the real data distribution in terms of both image quality and diversity.

Taken together, these metrics provide a more comprehensive evaluation of generative performance. While classification accuracy focuses on task-specific utility, FID reflects distributional similarity, highlighting the potential discrepancy between images that are useful for classification and those that represent the underlying data distribution.

4 Related Work

4.1 GAN-based Medical Image Generation

Generative Adversarial Networks (GANs) have been widely applied in medical image synthesis due to their ability to model complex data distributions and generate realistic samples. Various studies have demonstrated the effectiveness of GAN-based approaches across different imaging modalities, including CT, MRI, and X-ray, for tasks such as image reconstruction, cross-modality translation, and data augmentation [2].

A major motivation for using GANs in medical imaging is the limited availability of annotated data. To address this issue, several works have explored the use of synthetic data augmentation to improve classification performance. For example, Frid-Adar et al. [3] generated synthetic liver lesion images using GANs and showed that incorporating these images into the training set significantly improved CNN-based classification results. This demonstrates that GAN-generated data can enhance model generalization when real data is scarce.

More recently, GAN-based approaches have been applied to COVID-19 diagnosis using chest X-ray images. Waheed et al. [13] proposed CovidGAN, an Auxiliary Classifier GAN (ACGAN) framework for generating synthetic chest X-ray images. Their results show that augmenting the training dataset with generated samples can improve classification accuracy, highlighting the practical value of GANs in data-scarce scenarios.

Despite these advances, most existing works focus primarily on improving classification performance through data augmentation, while less attention has been given to systematically analyzing the quality of generated images and their impact on downstream tasks. In particular, the relationship between image realism and task-specific usefulness remains underexplored.

4.2 Diffusion Models in Medical Imaging

Diffusion models have recently become a powerful class of generative models capable of producing high-quality images. Denoising Diffusion Probabilistic Models (DDPM) introduce a framework in which data are progressively corrupted with noise and then reconstructed through a learned reverse process [10].

In medical imaging, diffusion models have been increasingly explored as an alternative to GAN-based approaches. Recent studies have demonstrated their effectiveness in generating high-fidelity synthetic images for data augmentation. For instance, Montoya-del-Angel et al. [4] applied diffusion models for mammographic image synthesis and showed that they can produce high-quality and diverse medical images, often outperforming GAN-based baselines in terms of fidelity and diversity.

Beyond unconditional generation, conditional diffusion models have also been proposed to incorporate additional information into the generation process. Hung et al. [14] introduced a conditional diffusion framework for medical image generation and demonstrated state-of-the-art performance in tasks such as super-resolution, denoising, and inpainting.

In addition to image synthesis, diffusion models have also been applied to image restoration and enhancement tasks. Chung et al. [15] proposed a regularized reverse diffusion framework for MRI denoising and super-resolution, showing that diffusion-based approaches can effectively recover high-quality images from degraded inputs. Similarly, Ma et al. [16] explored the use of pre-trained diffusion models for plug-and-play medical image enhancement, demonstrating their flexibility in improving image quality across different imaging settings. These results highlight the versatility of diffusion models across both generative and reconstruction tasks in medical imaging.

Despite these promising developments, diffusion models remain at a relatively early stage in medical image analysis compared to the more widely adopted GAN-based approaches. Therefore, further systematic investigation is needed to better understand their behavior across different downstream tasks, particularly in settings where they are directly compared with GAN-based methods.

4.3 Synthetic Data for Classification

The use of synthetic data has become an important strategy in medical imaging, where the availability of labeled data is often limited due to privacy concerns, high annotation costs, and data imbalance. Generative models provide a way to alleviate these issues by extending existing datasets with artificially generated samples.

Several studies have explored the effectiveness of synthetic data in improving classification performance. For instance, Frid-Adar et al. [3] demonstrated that incorporating GAN-generated liver lesion images into the training set can significantly improve the performance of convolutional neural networks. Similarly, Waheed et al. [13] applied an ACGAN to generate chest X-ray images for COVID-19 detection, showing that the augmented dataset leads to higher classification accuracy. More generally, Chatterjee et al. [17] combined GAN-based data augmentation with transfer learning and reported substantial improvements in classification accuracy across image recognition tasks.

Recent work has also extended this idea beyond GANs. Azizi et al. [18] showed that synthetic data generated by diffusion models can improve large-scale image classification performance, suggesting that diffusion-based approaches may offer advantages in capturing richer data distributions.

Despite these encouraging results, the impact of synthetic data on downstream tasks is not always straightforward. In particular, improvements in visual quality do not necessarily translate into better classification performance. Some studies have observed that generated images may capture superficial patterns that are sufficient for classification, while failing to accurately reflect the underlying data distribution.

These observations highlight an important distinction between image realism and task-specific usefulness. While many existing works focus on generating visually convincing samples, fewer studies systematically analyze how such data affects downstream classifiers. Understanding this relationship is particularly important when different generative paradigms, such as GANs and diffusion models, are used to produce synthetic data.

In this work, we further investigate this aspect by examining how generated images from different models influence classification performance, and whether improvements in visual fidelity lead to meaningful gains in downstream tasks.

5 Approach

5.1 Overview

In this work, we investigate the use of generative models for synthesizing COVID and non-COVID chest X-ray images to support classification tasks. To this end, two representative approaches are considered, namely an Auxiliary Classifier GAN (ACGAN) and a Denoising Diffusion Probabilistic Model (DDPM), which serve as baseline models.

Building upon these baselines, we introduce a classifier-guided supervision mechanism that is applied to both models during training. The goal is to encourage the generated images to preserve class-discriminative features that are relevant for downstream classification.

This unified setup allows for a direct comparison of how the same supervision strategy interacts with different generative paradigms. In particular, it enables an analysis of whether improvements in classification-oriented objectives lead to better generation quality, and whether the effect of such supervision is consistent across models. Furthermore, it provides a basis for examining the relationship between the visual realism of generated images and their effectiveness in downstream classification tasks.

5.2 Baseline Models

5.2.1 ACGAN

As a baseline GAN-based approach, an Auxiliary Classifier GAN (ACGAN) is adopted for conditional chest X-ray image generation. The implementation is based on the CovidGAN framework proposed by Waheed et al. [13], with modifications to accommodate the experimental setup used in this work.

The model is designed to generate COVID and non-COVID chest X-ray images conditioned on class labels, while simultaneously encouraging class consistency during training. The overall architecture of the implemented ACGAN model is illustrated in Fig. 4.

The generator receives two inputs: a latent noise vector z sampled from a Gaussian distribution and a class label y. The class label is first transformed through an embedding layer and then concatenated with the latent vector before being projected into a low-resolution feature representation. A series of transposed convolution layers is subsequently used to progressively upsample the features and generate an image of size 112×112. Batch normalization and ReLU activations are applied throughout the generator, while a Tanh activation function is used at the output layer to produce normalized image values.

The discriminator consists of multiple convolutional blocks with spectral normalization and LeakyReLU activations. Given an input image, the discriminator extracts hierarchical image features and produces two outputs. The first output estimates whether the input image is real or generated, while the second output predicts the corresponding class label. This auxiliary classification branch enables the model to learn class-dependent image characteristics in addition to adversarial discrimination.

The generator and discriminator are trained jointly using both adversarial and classification objectives. For the generator, the adversarial loss encourages the generation of realistic images, whereas the auxiliary classification loss promotes consistency between the generated image and the target class label. The overall generator objective is defined as:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda \mathcal{L}_{cls}, \tag{7}$$

where $\mathcal{L}_{adv}$ denotes the adversarial loss, $\mathcal{L}_{cls}$ represents the auxiliary classification loss, and λ controls the contribution of the classification term. In this work, λ is set to 1.0.

During training, all chest X-ray images are resized to 112×112 pixels and converted to three-channel grayscale images. The pixel values are normalized to the range $[-1, 1]$. The model is optimized using the Adam optimizer with a learning rate of 10^{-4} and trained for 2200 epochs.

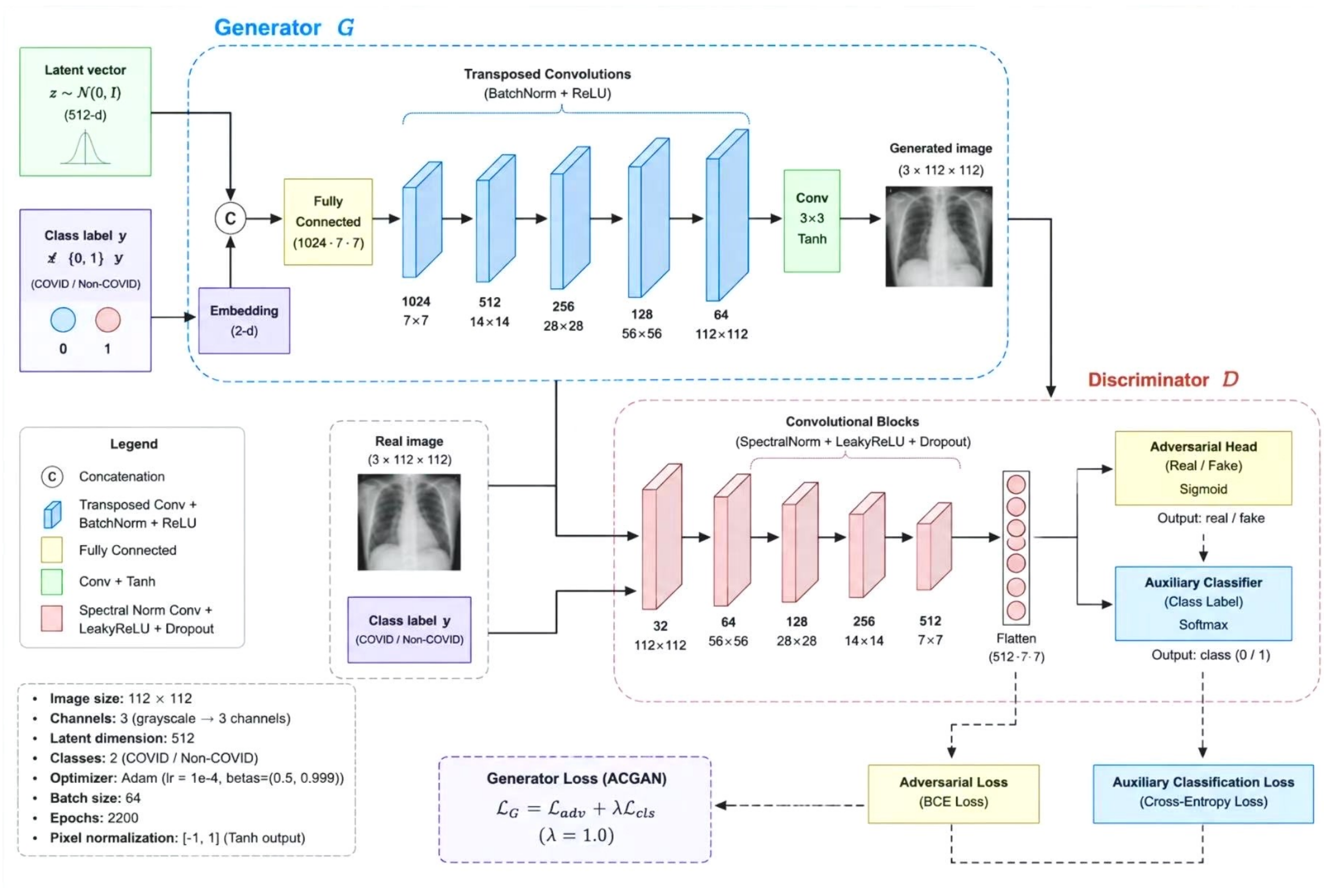

Figure 4: Architecture of the baseline ACGAN model used in this work. The generator receives a latent noise vector and a class label to generate conditional chest X-ray images, while the discriminator simultaneously predicts image authenticity and class labels through an auxiliary classification branch.

5.2.2 DDPM

As a baseline diffusion-based approach, a Denoising Diffusion Probabilistic Model (DDPM) is implemented for conditional chest X-ray image generation. The model is designed to generate COVID and non-COVID chest X-ray images by progressively denoising Gaussian noise through a conditional reverse diffusion process. The overall architecture of the implemented conditional U-Net is illustrated in Fig. 5.

During the forward diffusion process, Gaussian noise is gradually added to an input image over a predefined number of timesteps until the original image distribution is transformed into nearly pure noise. The reverse process is then learned by a neural network that predicts the noise component at each timestep and progressively reconstructs the image from noisy samples.

In this work, the DDPM baseline is implemented as a class-conditional denoising model based on a modified U-Net architecture. The network incorporates both timestep embeddings and class-label embeddings to condition the denoising process. Sinusoidal timestep embeddings are first generated and processed through a multilayer perceptron, while class labels are mapped through a learnable embedding layer. The two embeddings are subsequently fused and injected into residual blocks using Feature-wise Linear Modulation (FiLM).

The U-Net architecture consists of an encoder-decoder structure with residual convolutional blocks and skip connections. The encoder progressively extracts hierarchical image representations, while the decoder reconstructs the image through successive upsampling operations. To improve global feature modeling, a multi-head self-attention module is introduced at the bottleneck layer of the network. Compared with a standard DDPM implementation, the conditional information is more strongly emphasized by increasing the contribution of label embeddings during feature modulation.

The model is trained to predict the noise added to an image at a randomly sampled timestep. The training objective is formulated as:

$$\mathcal{L}_{DDPM} = \mathbb{E}_{x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t, y)\|^2\right], \tag{8}$$

where ϵ denotes the sampled Gaussian noise and ϵ_θ represents the predicted noise generated by the denoising network conditioned on timestep t and class label y. The model is optimized using the mean squared error (MSE) loss between the predicted and target noise.

During sampling, images are generated by iteratively denoising random Gaussian noise over $T = 1000$ timesteps. The model is trained using the Adam optimizer with a learning rate of 10^{-4} for 2000 epochs.

To maintain computational feasibility during diffusion training, the input image resolution was reduced to 64×64, compared with 112×112 used in the ACGAN baseline. This adjustment was necessary due to the substantially higher computational cost and memory requirements associated with iterative denoising and U-Net-based diffusion models.

5.3 Classifier-guided Supervision

Initial experiments showed that visually plausible synthetic chest X-ray images did not always achieve satisfactory classification performance when evaluated using an external classifier. This observation suggests that visual realism alone may not be sufficient to preserve diagnostically relevant features required for downstream classification tasks. To further encourage the generation of class-discriminative features, an additional classifier-guided supervision mechanism was introduced into both the ACGAN and DDPM frameworks.

A pretrained ResNet-18 classifier was adopted as the external supervision network. The classifier was initialized using ImageNet-pretrained weights and modified for binary classification of COVID and non-COVID chest X-ray images. After training on real chest X-ray data, the classifier parameters were frozen and used only to provide additional supervision during image generation. All generated images were first resized to 224×224, converted to grayscale three-channel images, and normalized to the range $[-1, 1]$ before being passed to the classifier.

Although the baseline ACGAN already contains an internal auxiliary classifier, its primary objective is to maintain class consistency within the adversarial learning framework. In contrast,

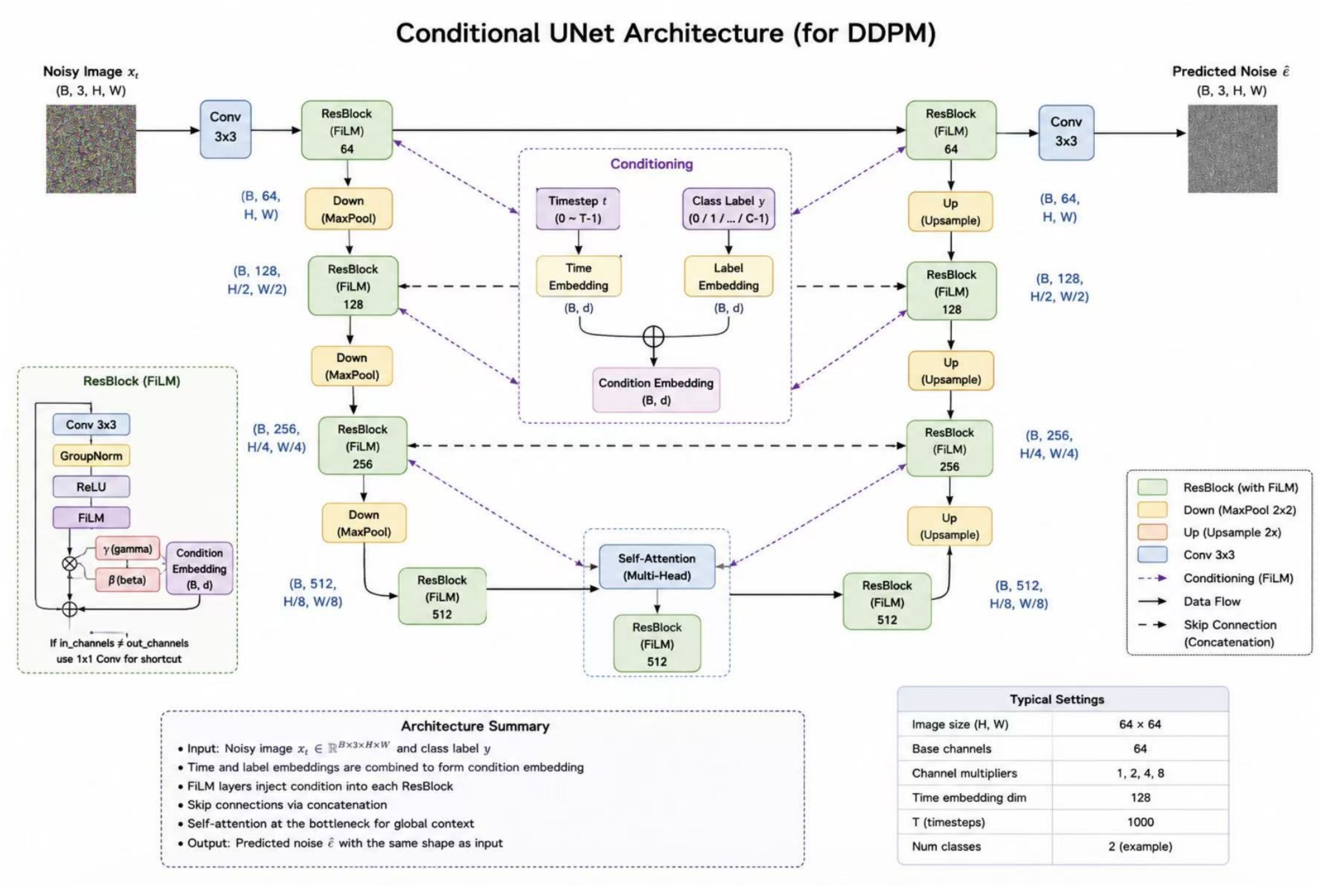

Figure 5: Conditional U-Net architecture used in the DDPM baseline. The model incorporates timestep embeddings and class-label embeddings through FiLM-based conditioning to guide the denoising process. Residual blocks, skip connections, and a self-attention bottleneck are used to improve feature representation during image reconstruction.

the external classifier introduced in this work is directly optimized for downstream classification performance. Therefore, the proposed classifier-guided supervision can be interpreted as a task-oriented supervision signal that explicitly encourages the preservation of diagnostically discriminative features rather than only improving adversarial class conditioning.

For the ACGAN-based model, the classifier supervision was incorporated directly into the generator objective. In addition to the original adversarial loss and auxiliary classification loss, an external classification loss was introduced based on the predictions of the frozen ResNet-18 classifier:

$$\mathcal{L}_G^{guided} = \mathcal{L}_{adv} + \lambda_{aux}\mathcal{L}_{aux} + \lambda_{cls}\mathcal{L}_{cls}^{ext}, \tag{9}$$

where $\mathcal{L}_{adv}$ denotes the adversarial loss, $\mathcal{L}_{aux}$ represents the internal auxiliary classification loss of ACGAN, and $\mathcal{L}_{cls}^{ext}$ corresponds to the classification loss produced by the external classifier. In this work, the weighting coefficients were set to $\lambda_{aux} = 1.0$ and $\lambda_{cls} = 0.5$. Under this formulation, the generator is encouraged not only to produce visually realistic images, but also to generate samples that can be correctly recognized by the external classifier.

For the DDPM-based model, classifier guidance was applied during both training and sampling. During training, the denoising network first predicted the clean image estimate x_0 from noisy inputs, and the external classifier loss was computed on the reconstructed image. The overall training objective was therefore defined as:

$$\mathcal{L}_{guided} = \mathcal{L}_{DDPM} + \lambda_{cls}\mathcal{L}_{cls}^{ext}, \tag{10}$$

where $\mathcal{L}_{DDPM}$ denotes the standard diffusion noise prediction loss. For the DDPM-based framework, the classifier loss weight was set to $\lambda_{cls} = 0.1$ in order to reduce excessive guidance effects observed during preliminary experiments. In addition, classifier guidance was further applied during the reverse diffusion sampling process by computing the gradient of the classifier loss with respect to the generated samples and using it to modify the predicted noise at each timestep.

Compared with the ACGAN-based approach, classifier guidance in DDPM influences the generation process more extensively due to the iterative nature of diffusion sampling. Since the denoising procedure is performed repeatedly over a large number of timesteps, classifier gradients may accumulate throughout the reverse diffusion process and progressively alter the generated image distribution. Consequently, the DDPM-based framework may become more sensitive to classification-oriented supervision than the one-step generation process used in ACGAN. The overall supervision pipelines used in the ACGAN and DDPM frameworks are illustrated in Fig. 6.

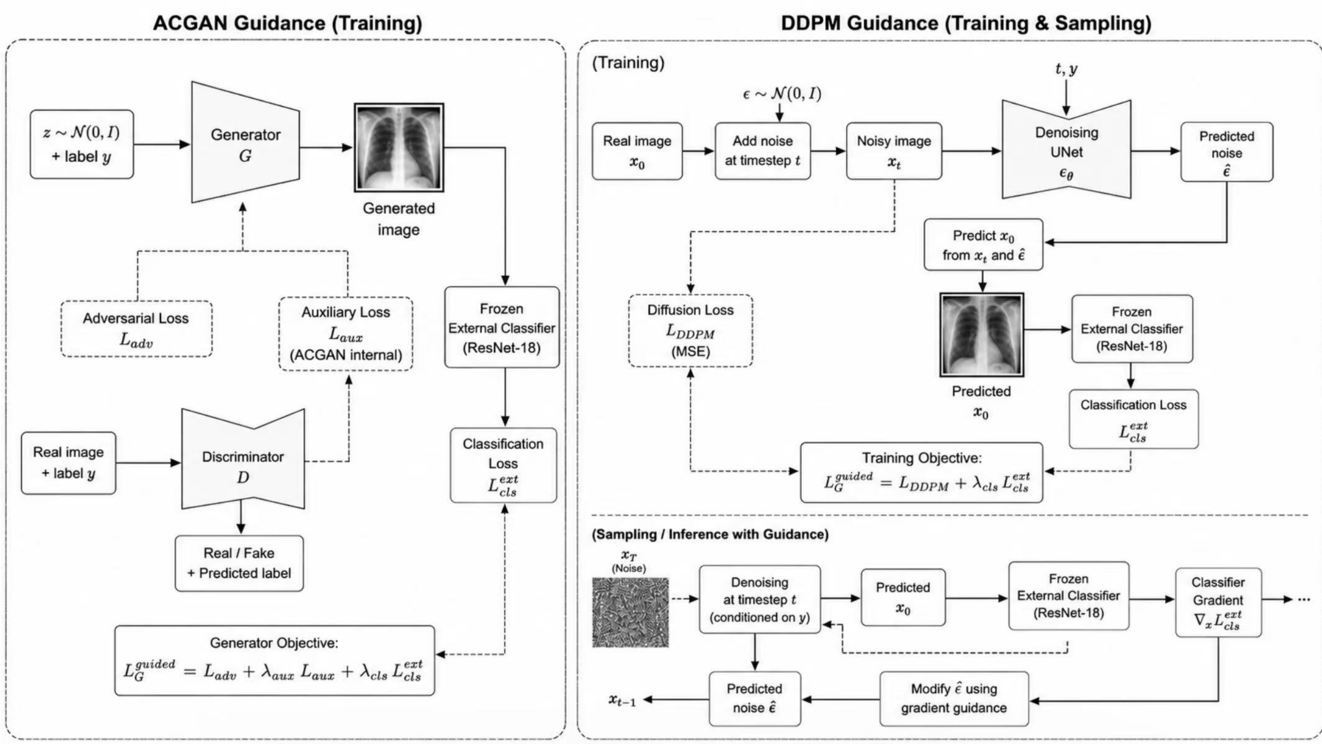

Figure 6: Overview of the classifier-guided supervision mechanisms used in the ACGAN and DDPM frameworks. In ACGAN, the external classifier loss is directly incorporated into the generator objective, while in DDPM the classifier guidance influences both denoising training and reverse diffusion sampling.

The overall objective of this supervision mechanism was not only to improve visual realism, but also to encourage the preservation of discriminative features relevant to downstream classification tasks. At the same time, introducing additional classification constraints may also alter the balance between realism and class separability, potentially affecting the stability and diversity of generated images.

5.4 Dataset and Implementation Details

5.4.1 Dataset Preprocessing

A publicly available chest X-ray dataset collected from multiple open medical imaging repositories was used in this work. The original dataset contains COVID-19, pneumonia, and normal chest X-ray images, with a total of 6432 samples, while the training, test, and validation sets are already pre-split.

For this study, only the COVID-19 and normal image subsets were used in order to formulate a binary classification setting for synthetic image generation and downstream evaluation. The pneumonia category was excluded to reduce inter-class ambiguity and focus specifically on the distinction between COVID-19 and healthy chest X-ray images.

All chest X-ray images were resized before training to match the input requirements of the corresponding generative models. For the ACGAN-based framework, images were resized to 112×112, while the DDPM-based framework used a reduced resolution of 64×64 to maintain computational feasibility during diffusion training. Pixel intensities were normalized to the range $[-1, 1]$ for all experiments.

For classifier evaluation and classifier-guided supervision, generated images were additionally resized to 224×224 and converted into three-channel grayscale representations before being processed by the external ResNet-18 classifier.

5.4.2 External Classifier Setup

An external classifier based on the ResNet-18 architecture was used for both evaluation and classifier-guided supervision. The network was initialized using ImageNet-pretrained weights and modified for binary classification of COVID-19 and normal chest X-ray images by replacing the final fully connected layer.

During classifier training, all input images were resized to 224×224, converted into three-channel grayscale representations, and normalized to the range $[-1, 1]$. The classifier was trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-4}.

After training, the classifier parameters were frozen and the network was used only for inference during generative model training and evaluation. For classifier-guided supervision, generated images were first preprocessed using the same normalization and resizing pipeline before being passed to the classifier. This consistent preprocessing procedure ensured that the classifier operated under the same input distribution throughout all experiments.

5.4.3 Training Configuration

All models were trained using the Adam optimizer with a learning rate of 1×10^{-4}. The original dataset split provided by the dataset was retained throughout all experiments, including the predefined training, validation, and test subsets. To ensure consistency across experiments, the same evaluation protocol and external classifier architecture were used for all generative models.

For the ACGAN framework, the generator and discriminator were trained for 2200 epochs with a batch size of 64. The latent vector dimension was set to 512, and the Adam optimizer parameters were configured with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For the DDPM framework, the denoising network was trained for 2000 epochs using 1000 diffusion timesteps.

The external classifier used for evaluation and classifier-guided supervision was based on a ResNet-18 architecture and trained for 10 epochs prior to guidance experiments. During classifier-guided DDPM sampling, the guidance strength parameter was set to 0.1.

Generated samples were periodically monitored during training, and intermediate image outputs were saved every 50 epochs to observe training progression and image quality changes over time.

Table 1: Training configurations used for ACGAN and DDPM experiments.

Parameter	ACGAN	DDPM
Image resolution	112×112	64×64
Batch size	64	64
Training epochs	2200	2000
Learning rate	1×10^{-4}	1×10^{-4}
Optimizer	Adam	Adam
Latent dimension	512	–
Diffusion timesteps	–	1000
Auxiliary loss weight	1.0	–
Classifier loss weight	0.5	0.1
Guidance strength	–	0.1

5.4.4 Implementation Environment

All experiments were implemented using PyTorch and conducted on a workstation equipped with an NVIDIA RTX 4090 GPU with 24 GB memory. The same hardware environment was used throughout all experiments to ensure consistency between different generative models and supervision settings.

Due to the high computational cost of diffusion-based training and sampling, the DDPM framework was trained using a reduced image resolution of 64×64. In contrast, the ACGAN-based framework was trained at a higher resolution of 112×112, since adversarial generation requires substantially lower computational resources compared with iterative diffusion sampling.

All quantitative plots, including training loss curves and classification accuracy curves, were generated directly from experimental logs using the Matplotlib library in Python.

Figures illustrating model architectures and supervision pipelines (Figures 4–6) were created by the author based on the implemented models and experimental design, with the assistance of generative AI tools for graphical layout and visualization.

6 Experiments

6.1 Experimental Setup

Four generative model configurations were evaluated in this work, including the baseline ACGAN model, the baseline DDPM model, the classifier-guided ACGAN model, and the classifier-guided DDPM model. For each model, a total of 2000 synthetic chest X-ray images were generated, consisting of 1000 COVID-19 images and 1000 normal images.

Two evaluation metrics were used throughout the experiments: Fréchet Inception Distance (FID) and classification accuracy under the external classifier. FID was computed between real COVID-19 chest X-ray images and generated COVID-19 samples in order to evaluate the fidelity of synthetic pathological features. Lower FID values indicate a smaller distributional difference between generated and real images.

In addition to FID, classification accuracy was used to evaluate whether generated images preserved recognizable class-discriminative features. For this evaluation, the frozen external ResNet-18 classifier was applied directly to generated samples, and the predicted labels were compared with the intended generation labels.

All experimental results were obtained using the final training checkpoints of each model. To ensure consistency between experiments, the same preprocessing pipeline, classifier architecture, and evaluation protocol were used across all generative frameworks.

6.2 Baseline Generation Results

6.2.1 Baseline ACGAN Results

Figure 7 shows the generator and discriminator losses during baseline ACGAN training. After an initially unstable stage during the early epochs, both losses gradually stabilized and remained within a relatively consistent range throughout the remainder of training. The discriminator loss converged near 0.5, while the generator loss fluctuated moderately around 1.0–1.5, indicating that the adversarial training process maintained a relatively stable balance without severe mode collapse.

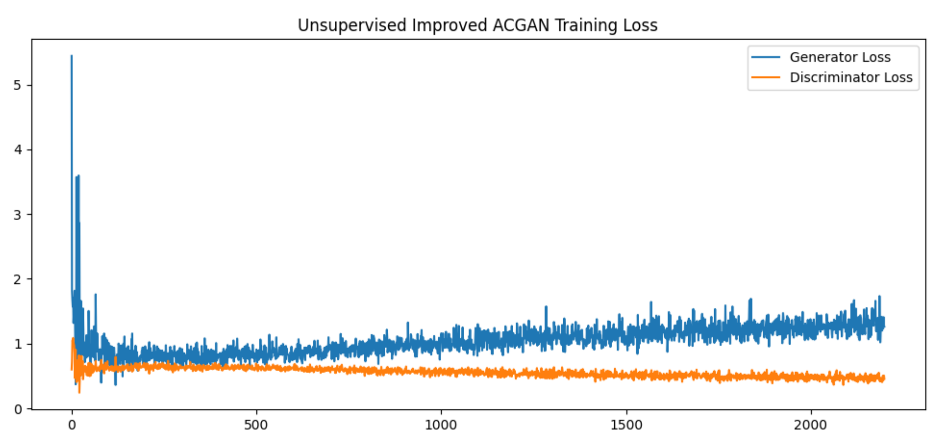

Figure 7: Generator and discriminator losses during baseline ACGAN training.

In addition to adversarial stability, the generated image classification accuracy under the external classifier rapidly improved during training and reached nearly perfect performance after the early training stage. As shown in Figure 8, the classification accuracy of generated samples approached 100% after approximately 200 epochs and remained stable throughout later training. This behavior suggests that the baseline ACGAN framework was already capable of generating highly class-consistent chest X-ray images even without external classifier-guided supervision.

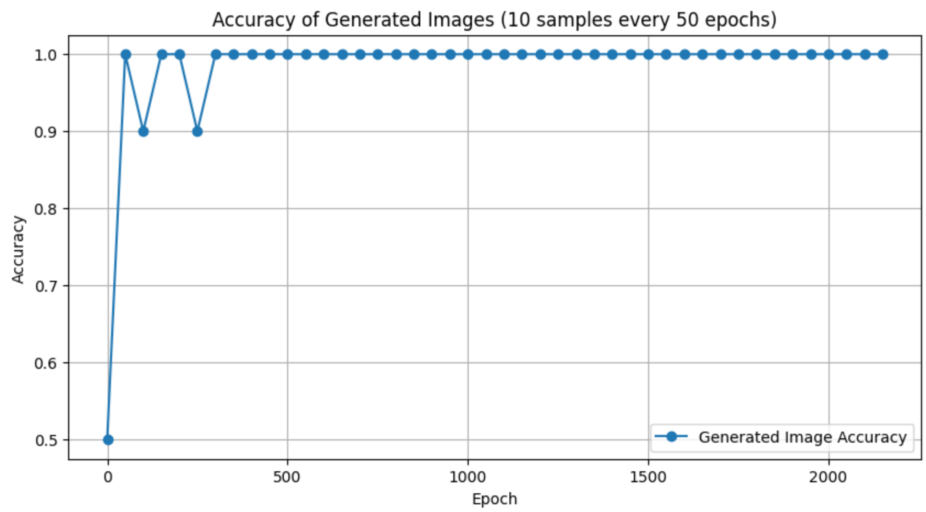

Figure 8: Classification accuracy of generated samples during baseline ACGAN training.

The progressive evolution of generated samples is illustrated in Figure 12. At epoch 0, the generated outputs consisted almost entirely of random noise without recognizable anatomical structures. By epoch 200, coarse chest cavity contours and rough lung-region layouts began to emerge, although the generated images still lacked fine structural details and remained visually blurry. In later training stages, such as epoch 2200, the generated chest X-ray images exhibited clearer anatomical structures and more realistic grayscale distributions. In particular, regions resembling the ground-glass opacity patterns commonly observed in COVID-19 chest X-ray images could also be identified in some generated COVID-19 samples. However, despite the substantial visual improvement observed during training, some synthetic samples still contained minor structural artifacts, blurry regions, and inconsistent pathological details, indicating that the generated image quality still has room for further improvement.

The strong classification consistency observed in the baseline ACGAN model is likely related to the auxiliary classification objective integrated into the ACGAN framework. Unlike standard GAN architectures, ACGAN explicitly optimizes class-discriminative generation during adversarial training. Therefore, even without additional external supervision, the model was still able to learn class-related discriminative features during image generation, which may naturally improve downstream classifier recognition performance on synthetic samples.

6.2.2 Baseline DDPM Results

Figure 13 shows the diffusion training loss during baseline DDPM training. Compared with adversarial GAN-based training, the DDPM optimization process exhibited a more stable convergence behavior without significant oscillation. The denoising loss gradually decreased throughout training, indicating that the model progressively improved its ability to reconstruct clean chest X-ray images from noisy latent states.

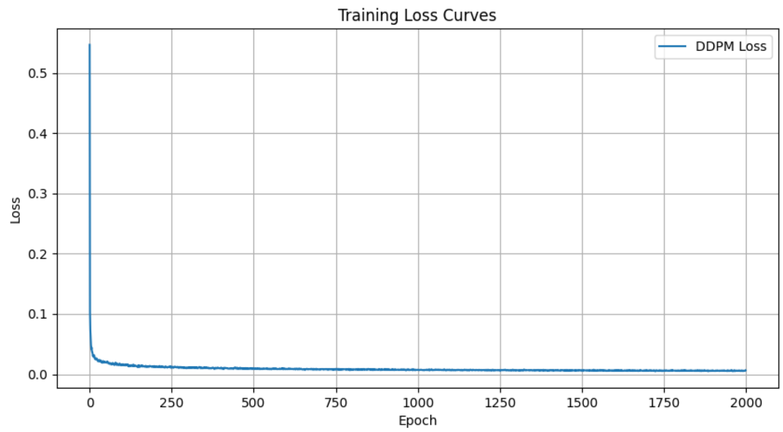

Figure 13: Training loss during baseline DDPM training.

However, unlike the baseline ACGAN model, the generated image classification accuracy under the external classifier remained relatively low throughout training. As shown in Figure 14, the classification accuracy fluctuated around 50%, which is close to random guessing performance for binary classification. Although the DDPM model was able to generate anatomically plausible chest X-ray structures, the generated samples did not consistently preserve strong class-discriminative features associated with COVID-19 and normal categories.

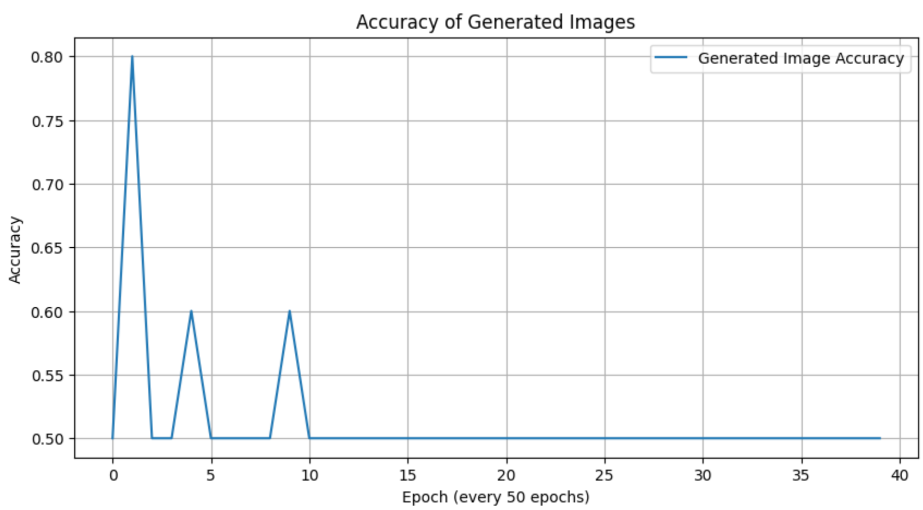

Figure 14: Classification accuracy of generated samples during baseline DDPM training.

The progressive evolution of generated samples during DDPM training is illustrated in Figure 18. At epoch 0, the generated outputs consisted entirely of random noise without visible anatomical structures. By epoch 200, coarse chest cavity structures and rough lung-region layouts began to emerge through the iterative denoising process. At later training stages such as epoch 2000, the generated chest X-ray images exhibited smoother anatomical structures and visually plausible grayscale distributions. However, compared with the baseline ACGAN model, the DDPM-generated images remained relatively blurry and contained weaker fine-grained pathological details. Qualitatively, the generated normal chest X-ray samples also appeared slightly clearer and structurally more consistent than the generated COVID-19 samples. This behavior may be related to the comparatively simpler anatomical distribution of normal chest X-ray images relative to the more complex pathological patterns associated with COVID-19 cases.

In this implementation, compared with the baseline ACGAN framework, the baseline DDPM model appeared to prioritize global anatomical realism rather than class-discriminative generation. This behavior is likely related to the diffusion objective itself, which primarily optimizes iterative denoising reconstruction instead of explicitly enforcing category separation during generation. As a result, although the generated images appeared visually smoother and structurally coherent, the generated samples remained difficult for the external classifier to distinguish reliably.

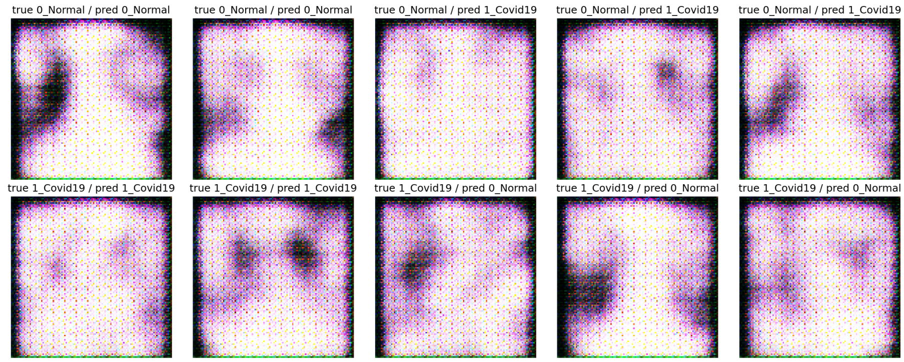

Figure 9: *
Epoch 0: Generated outputs consisted primarily of random noise without recognizable anatomical structures.

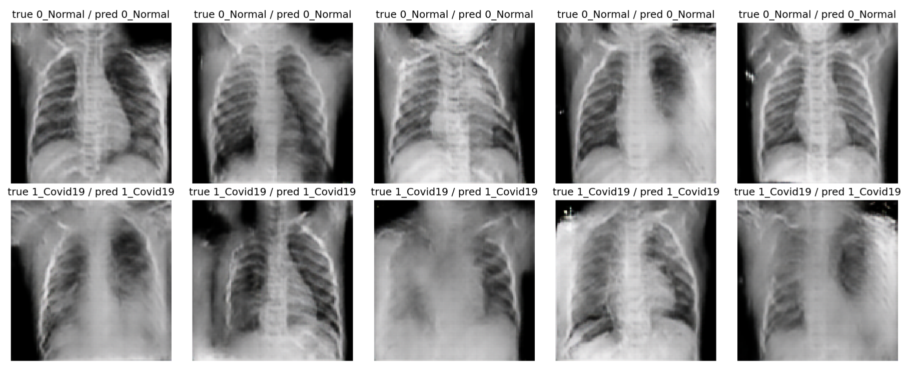

Figure 10: *
Epoch 200: Coarse chest cavity contours and rough lung-region layouts began to emerge, although structural details remained blurry.

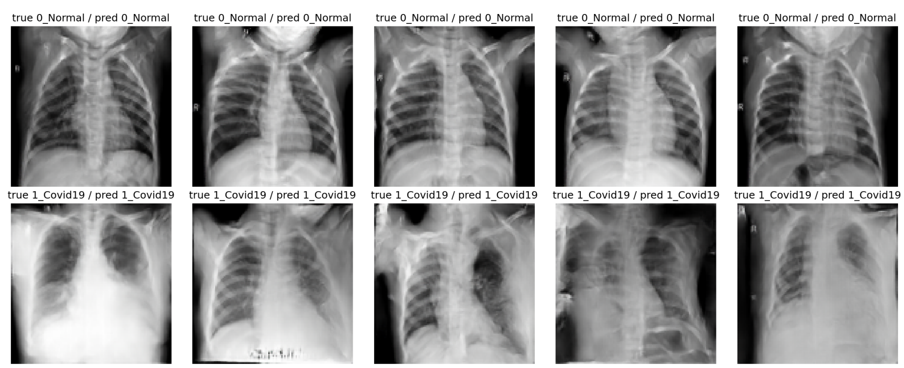

Figure 11: *
Epoch 2200: Generated chest X-ray images exhibited clearer anatomical structures and regions resembling ground-glass opacity patterns associated with COVID-19 images.

Figure 12: Progressive evolution of generated chest X-ray images during baseline ACGAN training.

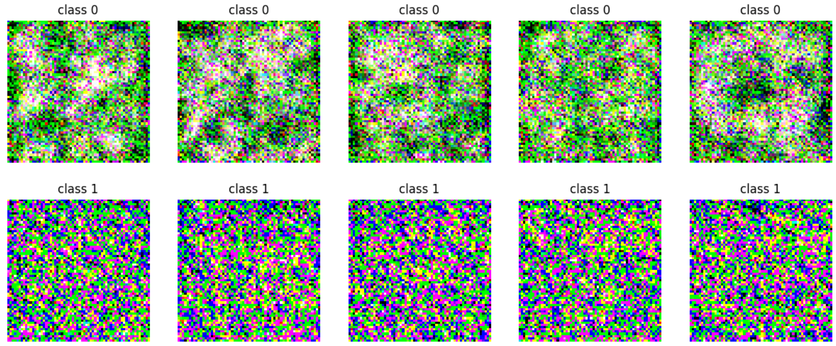

Figure 15: *
Epoch 0: Generated outputs consisted entirely of random noise without recognizable anatomical structures.

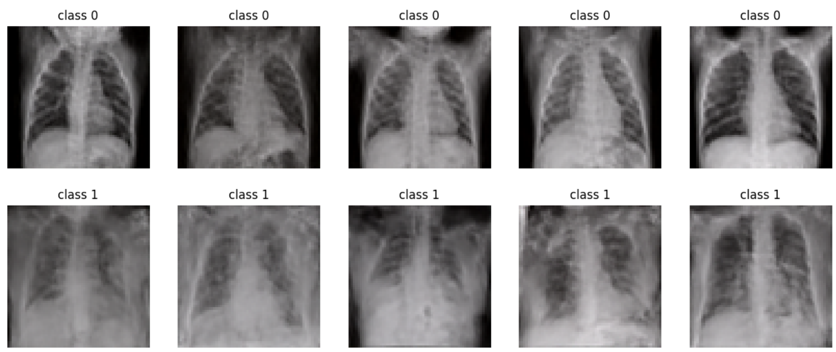

Figure 16: *
Epoch 200: Coarse chest cavity structures and lung-region layouts gradually emerged through iterative denoising.

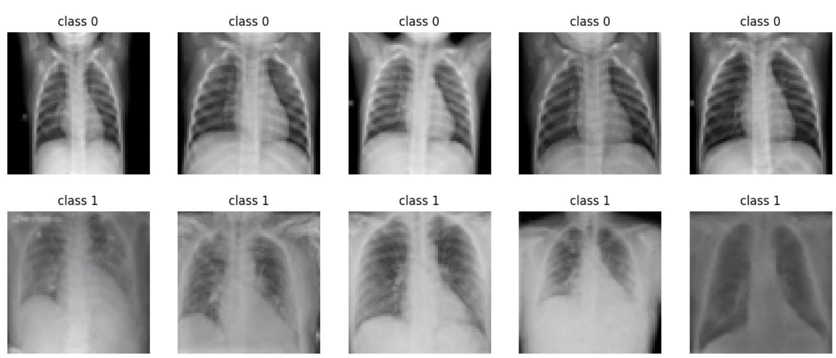

Figure 17: *
Epoch 2000: Generated chest X-ray images exhibited smoother anatomical structures and visually plausible grayscale distributions, although fine pathological details remained relatively weak.

Figure 18: Progressive evolution of generated chest X-ray images during baseline DDPM training.

6.3 Classifier-guided Generation Results

6.3.1 Classifier-guided ACGAN Results

Figure 19 shows the classification accuracy of generated samples during classifier-guided ACGAN training. Compared with the baseline ACGAN framework, the guided ACGAN model rapidly achieved perfect classification consistency under the external classifier and maintained an accuracy of 100% throughout nearly the entire training process after the initial epochs. This observation suggests that the external classifier supervision further reinforced the class-discriminative generation behavior already encouraged by the internal auxiliary classification objective of ACGAN.

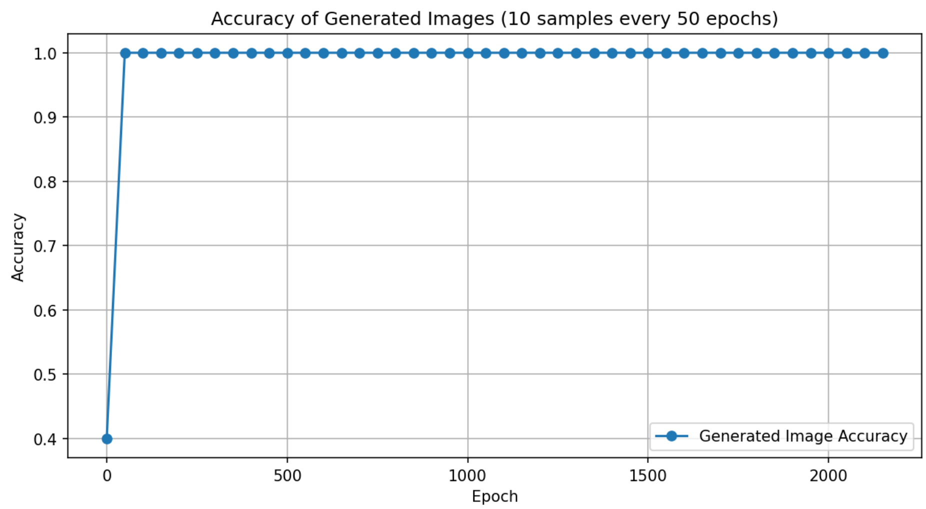

Figure 19: Classification accuracy of generated samples during classifier-guided ACGAN training.

Final generated samples produced by the classifier-guided ACGAN model are shown in Figure 20. Compared with the baseline ACGAN results, the supervised model exhibited slightly clearer anatomical structures, improved visual separation between COVID-19 and normal samples, and marginally enhanced pathological texture consistency. However, the overall visual improvement remained relatively limited, and the generated image quality was already reasonably strong in the baseline ACGAN framework prior to the introduction of external classifier guidance.

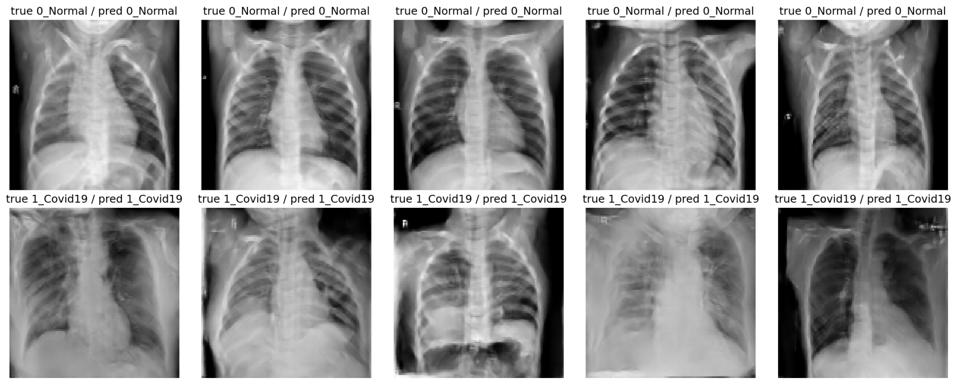

Figure 20: Generated chest X-ray samples produced by the classifier-guided ACGAN model at epoch 2200.

The relatively small improvement observed after introducing external classifier supervision is likely related to the original ACGAN architecture itself. Since ACGAN already incorporates an auxiliary classification objective during adversarial training, the generated samples were already optimized toward class-discriminative generation even before external supervision was added. As a result, the additional classifier-guided supervision appeared to provide only limited supplementary benefits for the ACGAN-based framework.

6.3.2 Classifier-guided DDPM Results

Figure 21 shows the classification accuracy of generated samples during classifier-guided DDPM training. Compared with the baseline DDPM framework, the introduction of external classifier supervision substantially improved classification consistency. After the first 50 training epochs, the generated sample classification accuracy rapidly increased to nearly 100% and remained stable throughout the remainder of training. This observation indicates that classifier-guided supervision successfully encouraged the DDPM framework to generate samples with stronger class-discriminative features.

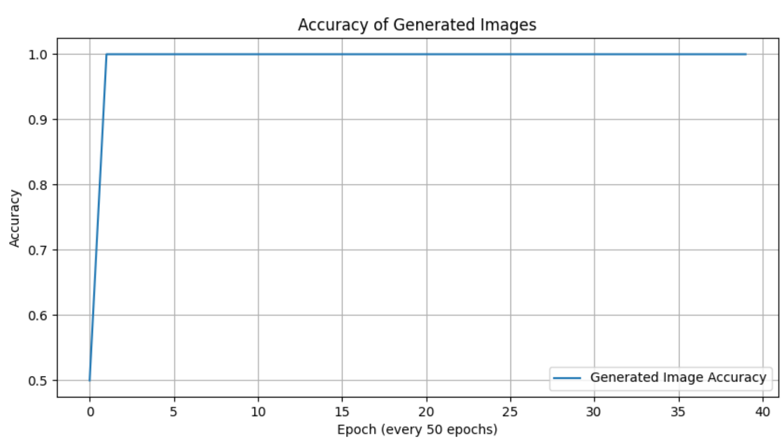

Figure 21: Classification accuracy of generated samples during classifier-guided DDPM training.

However, despite the significant improvement in classification consistency shown in the curve, classifier-guided supervision also introduced noticeable instability into the DDPM generation process. As shown in Figure 24, generated COVID-19 samples at epoch 1600 occasionally exhibited excessively dark outputs and partially collapsed pathological structures. This unexpected behavior suggests that overly strong classifier-oriented optimization may interfere with the original denoising objective of the diffusion process.

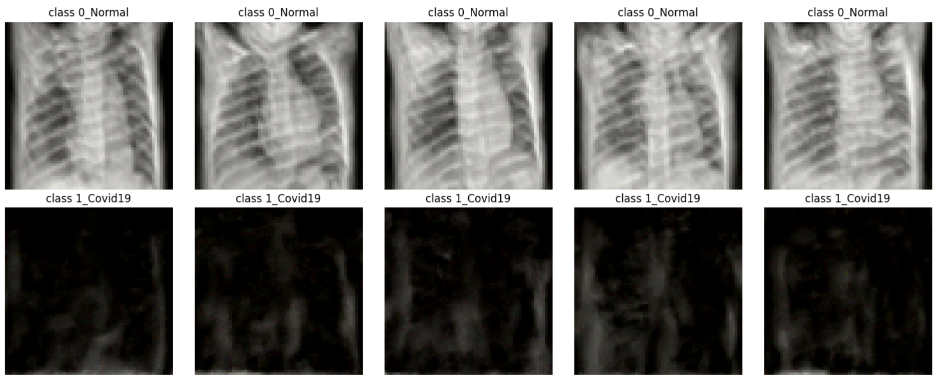

Figure 22: *
Epoch 1600: Some generated COVID-19 samples exhibited unstable pathological generation behaviors, including excessively dark outputs and partially degraded structures.

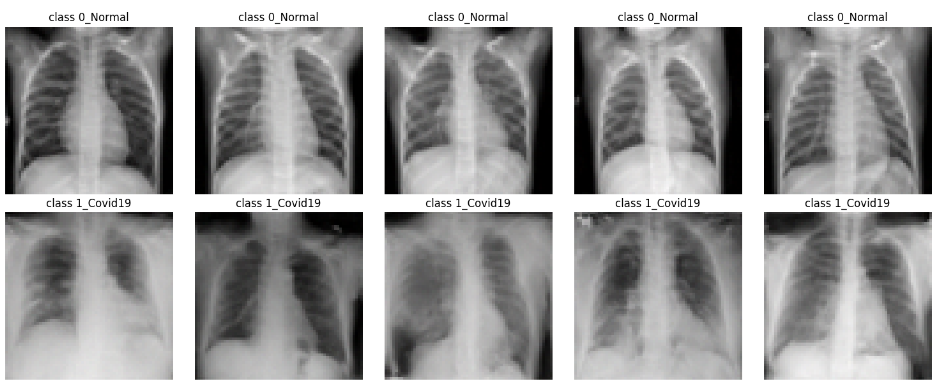

Figure 23: *
Epoch 2000: The generated samples became more stable compared with earlier training stages, although the COVID-19 samples remained less visually clear than those produced by the baseline DDPM model.

Figure 24: Generated chest X-ray samples during classifier-guided DDPM training.

At later training stages, such as epoch 2000, the generated chest X-ray images became visually more stable and structurally coherent compared with the unstable outputs observed earlier in training. Nevertheless, although classifier guidance significantly improved classification accuracy, the generated COVID-19 samples still appeared less visually clear and contained weaker pathological detail consistency compared with the baseline DDPM results. This behavior was also reflected by the increased FID score observed in the supervised DDPM framework.

Overall, the classifier-guided DDPM framework demonstrated a clear trade-off between classification consistency and generative realism. While external classifier supervision successfully enhanced class-discriminative generation, excessive guidance appeared to partially disrupt the original realism-oriented denoising behavior of the diffusion model, leading to reduced visual fidelity and occasional generation instability.

6.4 Quantitative Comparison

Table 2 summarizes the quantitative evaluation results of all generative frameworks using Fréchet Inception Distance (FID) and external classifier accuracy. Lower FID values indicate higher visual similarity between generated and real chest X-ray images, while higher classifier accuracy reflects stronger class-discriminative consistency in the generated samples.

Table 2: Quantitative comparison of all generative frameworks.

Model	FID	COVID Accuracy	Normal Accuracy
Baseline ACGAN	65	0.9982	0.9990
Guided ACGAN	62	0.9999	0.9995
Baseline DDPM	107	0.5862	0.5870
Guided DDPM	162	0.9966	0.9931

Among all evaluated models, the ACGAN-based frameworks achieved lower FID scores than the DDPM-based frameworks, indicating superior visual similarity between generated and real chest X-ray images. The classifier-guided ACGAN model achieved the best overall quantitative performance, with the lowest FID score and nearly perfect classification accuracy across both COVID-19 and normal categories. However, compared with the baseline ACGAN framework, the quantitative improvement introduced by external classifier supervision remained relatively limited, suggesting that the original ACGAN architecture had already learned strong class-discriminative generation behavior through its internal auxiliary classification objective.

In contrast, the baseline DDPM model exhibited weaker classification consistency despite producing visually smooth and anatomically coherent chest X-ray structures. The generated samples achieved classification accuracies only slightly above random guessing performance, indicating that the baseline DDPM framework struggled to preserve strong class-specific pathological features during the denoising generation process.

After introducing classifier-guided supervision, the DDPM framework achieved a dramatic improvement in classification accuracy, approaching the performance level of the ACGAN-based frameworks. However, this improvement was accompanied by a substantial increase in FID score, indicating degraded visual fidelity and reduced similarity between generated and real chest X-ray images. Combined with the qualitative observations discussed previously, these results suggest that excessive classifier-oriented optimization may partially interfere with the original realism-oriented denoising objective of diffusion models.

Overall, the quantitative results demonstrate a clear trade-off between classification consistency and generative realism within the classifier-guided DDPM framework. While external classifier supervision successfully strengthened class-discriminative generation behavior, it simultaneously reduced overall image realism and generation stability. In contrast, the ACGAN-based frameworks maintained relatively balanced performance across both realism and classification consistency due to the inherently class-aware adversarial training objective.

6.5 Discussion

The experimental results demonstrate that classifier-guided supervision affected the ACGAN and DDPM frameworks in different ways. Although both guided frameworks achieved improved classification consistency under the external classifier, the influence of classifier supervision on image realism and pathological feature preservation differed significantly between adversarial generation and diffusion-based generation.

For the ACGAN-based framework, the introduction of external classifier supervision resulted in only limited quantitative and qualitative improvement. The baseline ACGAN model had already achieved near-perfect classification accuracy prior to guidance, suggesting that the original adversarial training objective together with the auxiliary classification objective had already learned highly class-discriminative generation behavior. As a result, the external classifier guidance primarily acted as a supplementary reinforcement mechanism rather than thoroughly altering the generation process.

In contrast, the DDPM-based framework exhibited higher sensitivity to classifier-guided supervision. While the baseline DDPM model generated visually smooth and anatomically coherent chest X-ray structures, its generated samples showed weak pathological discrimination and low classification consistency. After introducing classifier-guided supervision, the DDPM framework rapidly achieved near-perfect classification accuracy. However, this improvement was accompanied by degraded visual fidelity, increased FID scores, and occasional generation instability observed in some training stages.

These observations suggest the existence of a trade-off between pathological discrimination and generative realism in classifier-guided diffusion models. The original DDPM objective primarily optimizes iterative denoising reconstruction and global image realism, whereas the external classifier objective emphasizes class-separable discriminative features. Consequently, excessive classifier-oriented optimization may partially interfere with the realism-oriented denoising process and encourage the generation model to exaggerate classifier-sensitive pathological cues rather than preserve globally realistic medical image distributions.

This behavior may also be related to the binary classification setting used in this work. Since the external classifier only needed to distinguish COVID-19 images from non-COVID-19 images, the supervision objective may have encouraged the model to focus on limited discriminative shortcuts rather than comprehensive pathological realism. As a result, generated samples could achieve high classification confidence while still deviating from realistic chest X-ray image distributions, particularly within the COVID-19 category.

In addition, the comparatively lower image resolution used in the DDPM framework may have further limited the preservation of fine-grained pathological details. Subtle COVID-19-related patterns such as diffuse opacity and ground-glass-like textures are more difficult to model under low-resolution diffusion generation settings. This may also partially explain why the generated normal chest X-ray images generally appeared more visually stable than the generated COVID-19 samples.

6.5.1 Limitations

Several limitations remain in the current study. First, the experiments were conducted only under a binary classification setting using COVID-19 and normal chest X-ray images. Second, the DDPM framework was trained using relatively low-resolution images due to computational limitations. Third, the evaluation primarily relied on FID scores and external classifier accuracy, while clinical validation by medical experts was not included. The external classifier was used both for classifier-guided supervision and for downstream evaluation, and FID might not fully reflect clinical realism in medical imaging domains. Additionally, due to computational limitations, the resolution mismatch between the ACGAN (112×112) and the DDPM (64×64) inherently biases the FID evaluation. Therefore, the reported FID scores should not be interpreted as a direct comparison of generative quality between the two model architectures. Instead, FID is used as an observational metric to evaluate the effect of classifier-guided supervision within each model framework, where changes in FID provide insight into the trade-off between realism and classification-oriented guidance.

7 Conclusion and Future Work

7.1 Conclusion

This study investigated the application of classifier-guided supervision in medical chest X-ray image generation using both ACGAN and DDPM frameworks. The experiments explored how external classifier guidance influences image realism, pathological feature preservation, and class-discriminative consistency in synthetic COVID-19 and normal chest X-ray image generation.

The experimental results demonstrated that the two generative paradigms responded to classifier-guided supervision in different ways. The baseline ACGAN framework already achieved highly class-consistent generation due to its auxiliary classification objective integrated within adversarial training. Consequently, the introduction of additional external classifier supervision produced only limited improvements in both quantitative performance and visual image quality.

On the other hand, the baseline DDPM framework generated visually smoother and anatomically coherent chest X-ray images but exhibited weak pathological discrimination capability under the external classifier. After introducing classifier-guided supervision, the DDPM framework achieved dramatically improved classification consistency. However, this improvement was accompanied by reduced visual fidelity, increased FID scores, and occasional generation instability during training.

Overall, the findings of this work suggest that classifier-guided supervision introduces a significant trade-off between pathological discrimination and generative realism in diffusion-based medical image generation. While classifier guidance can enhance class-discriminative generation behavior, excessive classifier-oriented optimization may simultaneously interfere with realism-oriented denoising reconstruction and reduce global medical image fidelity.

The results further indicate that adversarial generation and diffusion-based generation exhibit fundamentally different sensitivities to external supervision objectives in medical imaging tasks. These observations may provide useful insights for future research on controllable medical image synthesis and pathology-aware generative modeling.

7.2 Future Work

Several possible directions may further improve the effectiveness and realism of classifier-guided medical image generation.

First, future work could extend the current binary classification setting to a multi-class setting including normal, COVID-19, and other pneumonia-related chest diseases. Compared with binary supervision, multi-class guidance may provide more fine-grained pathological supervision and reduce the tendency of the classifier to rely on simplified decision-boundary shortcuts.

Second, higher-resolution diffusion generation may improve the preservation of subtle pathological details such as diffuse opacity and ground-glass-like textures commonly observed in COVID-19 chest X-ray images. Due to computational limitations, the current DDPM framework was trained only at relatively low image resolution.

Third, future studies may investigate alternative supervision strategies beyond standard classifier guidance. Since the current results suggest that classification-oriented objectives may partially conflict with realism-oriented diffusion objectives, future approaches such as classifier-free guidance, perceptual supervision, or pathology-aware feature guidance may help balance pathological discrimination and medical image realism more effectively.

Finally, future evaluation should include clinical validation from radiologists or medical experts in addition to quantitative metrics such as FID and classifier accuracy. Although the generated images demonstrated promising class-discriminative behavior, the actual diagnostic usefulness and clinical realism of synthetic medical images still require further validation in real-world medical settings.

References

[1] S. Johri, M. Goyal, S. Jain, M. Baranwal, V. Kumar, and R. Upadhyay. A novel machine learning-based analytical framework for automatic detection of covid-19 using chest x-ray images. *International Journal of Imaging Systems and Technology*, 31(3):1105–1119, 2021.

[2] Nripendra Kumar Singh and Khalid Raza. Medical image generation using generative adversarial networks: A review. In *Health Informatics: A Computational Perspective in Healthcare*, pages 77–96. Springer, 2021.

[3] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[4] R. Montoya-del Angel, K. Sam-Millan, J. C. Vilanova, and R. Martí. Mam-e: Mammographic synthetic image generation with diffusion models. *Sensors*, 24(7):2076, 2024.

[5] Jonathan P. Cohen, Paul Morrison, and Lan Dao. Chest x-ray findings in covid-19 patients presenting to the emergency department. *Journal of Healthcare Engineering*, 2021.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[7] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[8] Augustus Odena, Chris Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[9] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 2015.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[11] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171, 2021.

[12] 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. In *Advances in Neural Information Processing Systems*, 2017.

[13] Abdul Waheed, Muskan Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman, and Plácido Rogério Pinheiro. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access*, 8:91916–91923, 2020.

[14] Alex Ling Yu Hung, Kai Zhao, Haoxin Zheng, Ran Yan, Steven S. Raman, Demetri Terzopoulos, and Kyunghyun Sung. Conditional medical image generation with diffusion models. *Bioengineering*, 10(11):1258, 2023.

[15] Hyungjin Chung, Eun Sun Lee, and Jong Chul Ye. Mr image denoising and super-resolution using regularized reverse diffusion. *IEEE Transactions on Medical Imaging*, 42(4):922–934, 2023.

[16] Jiaqi Ma, Yixin Zhu, Cheng You, and Bo Wang. Pre-trained diffusion models for plug-and-play medical image enhancement. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14222 of *Lecture Notes in Computer Science*. Springer, 2023.

[17] Subhajit Chatterjee, Debapriya Hazra, Yung-Cheol Byun, and Yong-Woon Kim. Enhancement of image classification using transfer learning and gan-based synthetic data augmentation. *Mathematics*, 10(9):1541, 2022.

[18] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.