Received: 1 July 2025 | Revised: 3 September 2025 | Accepted: 15 September 2025 | Published online: 17 October 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 00(00) 1-10
DOI: 10.47852/bonviewAIA52026661

BON VIEW PUBLISHING

Addressing Small and Imbalanced Medical Image Datasets Using Generative Models

Iman Khazrak[1,*], Shakhnoza Takhirova[1], Mostafa M. Rezaee[1], Mehrdad Yadollahi[1], Robert C. Green II[1] and Shuteng Niu[1,2]

[1] *Department of Computer Science, Bowling Green State University, USA*

[2] *Department of Artificial Intelligence and Informatics, Mayo Clinic, USA*

Abstract: Progress in accurate medical image classification is often hampered by concerns surrounding data privacy and scarcity of data for certain medical diseases, leading to sparsity and unbalanced datasets. To address these challenges, this study uses generative models, namely, Denoising Diffusion Probabilistic Models (DDPMs) and Progressive Growing Generative Adversarial Networks (PGGANs), for dataset improvement. In this article, we propose a framework for understanding how the resultant synthetic images generated by DDPM and PGGANs affect four different models' performance: a specially crafted Convolutional Neural Network, an untrained VGG16, a pretrained VGG16, and a pretrained ResNet50. For modeling practical constraints in real applications, experiments applied Random Sampling and Greedy K Sampling to obtain small unbalanced datasets. Synthetic image quality was also measured by applying Fréchet Inception Distance (FID), and their impact was further explored by comparing classification results with their original datasets. Experiments reveal that DDPM consistently produced images of higher realism, backed by lower FID scores, and overtakes PGGANs in augmenting classification outcomes of all investigated models and datasets. Addition of DDPM-generated images to original datasets obtained improvement of about 6% in accuracy and therefore enhanced robustness and reliability of models, specifically when datasets are unbalanced. Although Random Sampling obtained better consistency, Greedy K Sampling obtained higher variability but higher FID scores. Overall, this research identifies the potential of DDPM to effectively augment and balance sparse datasets of medical images and subsequently improve training of models and predictive outcomes.

Keywords: medical image augmentation, generative models, Progressive Growing Generative Adversarial Networks (PGGANs), Denoising Diffusion Probabilistic Models (DDPMs), synthetic data integration

1. Introduction

Medical imaging is the foundation of modern medicine, guiding diagnostics, surgery, treatment following, and disease monitoring. Growing volumes of images present challenges [1] for clinicians and radiologists to maintain productivity of workflow without computer intervention. There are significant challenges to train accurate and reliable Machine Learning or Deep Learning diagnostic classifiers. The primary concerns are the absence of complete and diverse datasets [2], rigorous data privacy legislation, and inherent dataset imbalances. These imbalances lead to biased classifiers likely to fail with rare diseases and small errors having unintended effects.

Conventional data augmentation procedures like random rotations, flipping, cropping, and noise injection have been heavily utilized to augment training sets. While beneficial, they simply reorder available samples and do not imbue the system with the type of basic variability required for comprehensive model training [3]. As opposed to conventional data augmentation procedures, generative models [4], Generative Adversarial Networks (GANs), and Denoising Diffusion Probabilistic Models (DDPMs) have changed the landscape of image synthesis by producing brand new data points. Generative models present potential solutions to the problems of imbalanced datasets, especially when applied to the realm of medical imaging where labeled data are scarce.

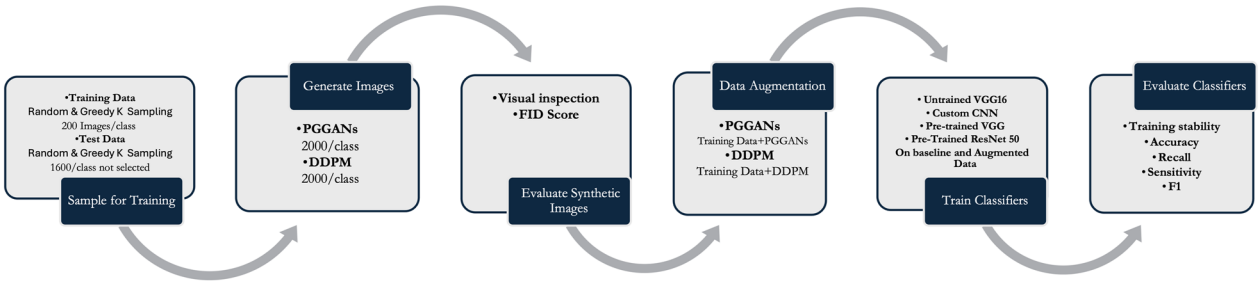
***Corresponding author:** Iman Khazrak, Department of Computer Science, Bowling Green State University, USA. Email: ikhazra@bgsu.edu

Generative models usually demand large and diverse sets [5]. This is paradoxical: if we had such large sets of labeled data available, efficient models could simply be directly trained. Generative models are thus only tenable if they are capable of functioning effectively with small sets of data. The current article meets this challenge by introducing the comprehensive framework of generating synthetic medical images from small and imbalanced sets of data with two generative models: Progressive Growing Generative Adversarial Networks (PGGANs) [6] and DDPM [7]. We further investigate applying two sampling methodologies—Random Sampling and the Greedy K Sampling methods—to also evaluate their effects on the performance of the model. This framework is extensively tested to improve the diagnostic accuracy and robustness of the model.

This work proposes a framework for applying advanced generative models to small and imbalanced medical image sets (Figure 1). Our main contributions are as follows:

1) Holistic assessment process: We construct an intensive process of assessing the quality and performance of synthetic images produced by DDPM and PGGANs. This process has three phases. First, synthetic images are produced with the two models. Second, the quality of the produced images is tested visually and with quantitative measures like Fréchet Inception Distance (FID), assessing the generated image–real image similarity, and the VGG Expert model for visual confirmation. Third, we investigate the effect of adding synthetic images to small and imbalanced sets on the performance of the classification model.

Figure 1
Framework for evaluating synthetic images



2) Generation of high-quality synthetic images: Through extensive experimentation, we demonstrate the feasibility of generating high-quality synthetic images from small medical image datasets using DDPM and PGGANs. DDPM consistently outperforms PGGANs in terms of FID scores, producing more realistic and diverse images that improve dataset size and balance.
3) Improved classification performance: By integrating synthetic images into small and imbalanced datasets, we show significant improvements in the performance of both custom Convolutional Neural Networks (CNNs) and untrained VGG16 models. For example, the accuracy of untrained models trained on small datasets improved significantly.
4) Enhanced model stability: Our findings highlight that incorporating synthetic images into the original datasets enhances the stability of both untrained and pretrained classification models. Notably, DDPM provides better stability and consistency in performance than PGGANs, especially under challenging conditions of small dataset.

In brief, our research introduces the novel method of overcoming the limitation of sparse and imbalanced medical datasets with the help of the most powerful generative models. It reveals the promise of DDPM and PGGANs to complement the data so as to not only expand and adapt the size of the sets efficiently but also enhance considerably the precision, consistency, and robustness of the classification models of the area of medical imaging. The following sections give the presentation of the overview of the generative models and detailed description of DDPM and PGGANs and the application to the area of medical imaging as well as the detailed examination of our method and our findings.

The outline of this article consists of the overview of generative models, followed by the discussion of two important methods used in the article, DDPMs and PGGANs, and their applications in medical imaging. Then, we proceed to discuss our methodology and obtained results.

2. Related Work

Generative models, particularly those generating high-quality realistic images, have been of great interest in supplementing medical datasets and especially in rare diseases where data insufficiency and class imbalance are the norm. Such models can themselves be categorized as latent variable generative models, either explicit or implicit density models.

Concepts from unrelated fields offer beneficial learnings for healthcare. As an example, Sustainability Value Articulation enhances internal and external actions toward better social and environment performance by underscoring the involvement of suppliers and technological integration [8]. Similarly, the EV supply chain emphasizes the necessity of constant benchmarking and technological development for building competitive advantages in complex systems [9]. These principles have the same goals as generative models to cope with the scarcity of data and complement the quality of healthcare data with the possibility of long-term scalability and effect.

Explicit density models such as Variational Autoencoders, Boltzmann Machines, and DDPMs possess predefined density functions and provide interpretability and stability in training [7, 10, 11]. As this class of models is beneficial for applications involving anomaly detection because of the explicit likelihood functions they possess, their distributional assumptions sometimes result in less realistic images [12].

Implicit density models like GANs lack explicit likelihood functions and thus are less restrictive and can learn complex distributions. While they generate more realistic pictures, they suffer from training instability and difficulty in evaluation as well as hyperparameter sensitivity [13, 14].

2.1. GAN family in medical imaging

GANs are prominent implicit density models that consist of two competing neural networks, a generator that creates synthetic images from a latent space and a discriminator which evaluates resemblance of generated images to real images, engaging in a zero-sum game. Generally, it is hard to train GANs due to training instability [14]. PGGANs, introduced by Karras et al. [6], have significantly improved the stability and quality of GAN-generated images. PGGANs utilize progressive training procedures, where low-resolution images are applied at initialization and escalated step by step with training progress. Such a process allows for easier training of the network to learn coarse information before fine information and generate better pictures.

In medical imaging, GANs mainly have been used to enhance classification and segmentation deep learning models [15]. The work by Costa et al. [2] uses GANs on a small CT scan dataset to generate eye fundus images which confirm to the given masks. Mahapatra et al. [16] also used mask to generate lung images, and only the synthetic images that fulfilled informativeness criteria calculated by Bayesian neural networks were used to improve the classifier model. In the study by Frid-Adar et al. [4], GANs are employed to synthesize high-quality focal liver lesions of multiple conditions to enhance a CNN classifier. Moreover, GANs have been successful at synthesizing prostate lesions [17], lung cancer nodules [18], and brain MRI images [19] to name a few. Chen et al. [20] generate high-resolution synthetic images of skin lesions from a dataset of 2,000 dermoscopic images using multiple GAN architectures and compare their classification performances. They conclude that PGGANs could synthesize realistic images that medical professionals upon evaluation were not able to distinguish from real ones. Results of the study by Park et al. [21] confirm that PGGANs can produce high-resolution images with remarkable detail and consistency, making them one of the best choices for medical image synthesis

2.2. Diffusion family in medical imaging

Diffusion models are generative models that transform noise into structured data through a sequence of steps. The DDPM [7] is a prominent model in this family, known for producing high-fidelity images by reversing a diffusion process. These models iteratively add and then remove noise from an image through two main phases: the forward process, where noise is added over several steps, and the reverse process, where the model learns to denoise the image step by step. This iterative refinement allows DDPMs to generate images with fine-grained details. Introduced by Ho et al. [7] in 2020, DDPMs have set new benchmarks in image generation quality by leveraging a sophisticated noise schedule and a robust denoising network.

Utilization of DDPMs for application to medical imaging has also been explored for varied applications. Nichol and Dhariwal [22] reported evidence of guided diffusion models and upsampling-based models to efficiently improve MRI resolution to better diagnose and plan for treatment with a dataset of 10,000 MRI images. For applications in medical imaging, combined utilization of explainability and trust in AI-based applications has also been invaluable for clinician acceptance of AI-based applications for life-threatening diseases like cancer. Rezaeian et al. [23] posit a two-stage AI architecture for diagnosing breast cancer and introduce graded explainability levels like tumor localization and probability distributions to increase trust in AI-based applications, which were found to significantly enhance trust in AI-based applications. In line with this, our current research is targeted on enhancing AI model robustness by addressing data sparsity and imbalance challenges as central challenges for building robust diagnosis-based tools. Jalal et al. [24] further explored DDPMs for MRI reconstruction and reported substantial improvement in image quality and noise robustness using a dataset of 3,500 MRI data. In line with this, Wolleb et al. [25] utilized DDPMs for application to medical image segmentation and reported state-of-the-art results using a dataset of 7,500 images. Müller-Franzes et al. [26] compared latent DDPMs and GANs for application to medical image synthesis for varied modalities using 8,000 CT and MRI data and reported DDPMs to have superior image quality and diversity when compared with alternative uses of GANs for image synthesis. Liang et al. [27] reported a DDPM-based X-ray Image Synthesizer using 6,000 X-ray image data and established the capability of generating high-fidelity synthetic X-ray images to enhance training datasets and enhance diagnosis-based model accuracy.

Most studies use large datasets for image generation or do not directly leverage generated datasets to improve model performance. In contrast, our approach uses a small dataset to generate synthetic images and shows how these images enhance model performance, addressing data scarcity and imbalance. This underscores the potential of DDPMs to transform medical imaging, making diagnostic tools more accurate, reliable, and accessible.

3. Methodology

Our research methodology includes several key phases: image synthesis, dataset augmentation, model training and fine-tuning, and performance evaluation.

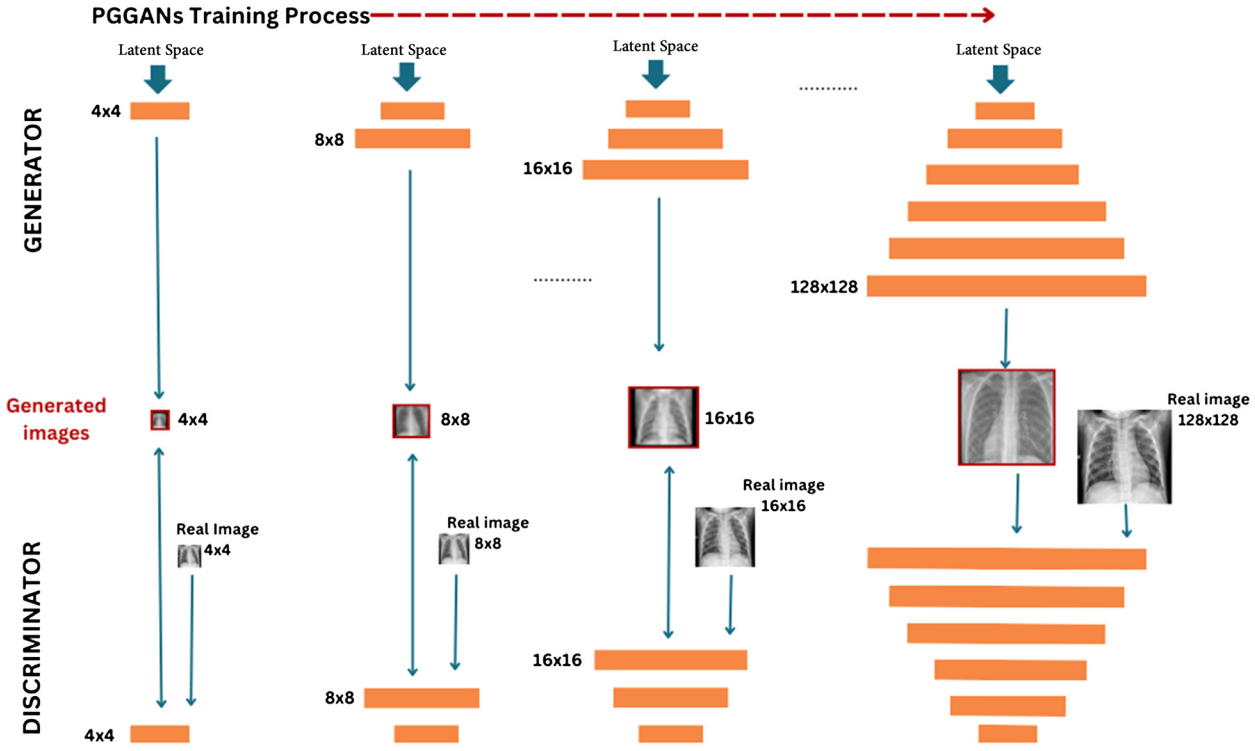
3.1. Image synthesis

3.1.1. PGGANs

PGGANs utilize a progressive training approach, starting with low-resolution images and gradually increasing the resolution as training progresses (Figure 2). This method enhances stability and image quality by incrementally increasing the complexity of the generator and discriminator networks. The generator produces data resembling real data, while the discriminator distinguishes between real and generated data [28]. The adversarial loss functions for the generator ($\mathcal{L}_G$) and discriminator ($\mathcal{L}_D$) are as follows:

$$\mathcal{L}_G = \log(1-D(G(z))) \quad (1)$$

Figure 2
PGGAN training progression



$$\mathcal{L}_D = \log(D(x)) + \log(1-D(G(z))) \quad (2)$$

where $G(z)$ represents the generated data from noise z and $D(x)$ represents the discriminator's output for real data x [29].

PGGANs adopt a step-by-step training approach, beginning with low-resolution images and advancing to higher resolutions. This progressive training allows the model to learn rough features initially and then fine-tune them for generating high-quality images. New layers are added to both networks iteratively, and the loss functions are applied at each resolution level to maintain consistency.

3.1.2. DDPMs

DDPMs synthesize images by reversing a diffusion process that gradually adds Gaussian noise to an image and then reconstructs the original image from the noise (Figure 3) [7].

The forward process adds noise to the image:

$$x_t = \sqrt{a_t}x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t \quad (3)$$

where x_t is the image at iteration t, α_t is a noise scaling factor, and ε_t is the Gaussian noise added at iteration t [7]. The backward process is aimed at denoising the noisy image obtained from the forward process and recovering the original clean image by optimizing the variational lower bound:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,x_0,\varepsilon}\left[|\varepsilon - \varepsilon_\theta(x_t,t)|^2\right] \quad (4)$$

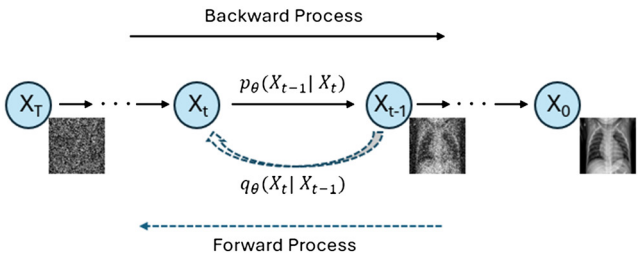
Here, ε represents Gaussian noise and ε_θ is the noise predicted by the model.

The U-Net architecture, adapted for use in DDPMs, excels in the reverse diffusion process by predicting and removing noise added during the forward phase [13, 30]. U-Net's U-shaped structure with downsampling and upsampling paths efficiently synthesizes detailed images, incorporating time embeddings to adjust noise prediction based on the reverse process timestep [31].

3.2. Generated image assessment

1) Visual inspection: Generated images are initially evaluated by visually comparing random samples to the original images.
2) FID: The FID score quantifies the distributional similarity between real and generated images. It is calculated by extracting features from an InceptionV3 model for both real and generated images and then computing the Fréchet distance between the resulting multivariate Gaussian distributions. A higher FID score indicates greater dissimilarity [32].

Figure 3
Directed graphical model of DDPM



3.3. Classification models

We compare the impact of synthetic images using four separate classifiers: pretrained VGG16 and ResNet50 (Table 1), an untrained VGG16, and a self-built CNN (Table 2). Each of these models is first trained on both imbalanced and small datasets to establish baselines before training on augmented versions of both datasets using DDPM- and PGGAN-created synthetic images. Inclusion of an untrained VGG16 allows for assessing the direct impact of synthetic data on

Table 1
VGG16 and ResNet50 model summary

Layer (Type)	Output Shape	Param #
vgg16/resnet50 (Functional)	(None, 7, 7, 512)	14,714,688
flatten (Flatten)	(None, 25,088)	0
dense (Dense)	(None, 512)	12,845,568
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 2)	1,026
Total params:		27,561,282 (105.14 MB)
Trainable params:		12,846,594 (49.01 MB)
Non-trainable params:		14,714,688 (56.13 MB)

Table 2
Custom CNN model summary

Layer (Type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 128, 128, 64)	1,792
batch_normalization_1 (BatchNorm)	(None, 128, 128, 64)	256
max_pooling2d_1 (MaxPool2D)	(None, 64, 64, 64)	0
conv2d_2 (Conv2D)	(None, 64, 64, 128)	73,856
batch_normalization_2 (BatchNorm)	(None, 64, 64, 128)	512
max_pooling2d_2 (MaxPool2D)	(None, 32, 32, 128)	0
conv2d_3 (Conv2D)	(None, 32, 32, 256)	295,168
batch_normalization_3 (BatchNorm)	(None, 32, 32, 256)	1,024
max_pooling2d_3 (MaxPool2D)	(None, 16, 16, 256)	0
flatten (Flatten)	(None, 65536)	0
dense_1 (Dense)	(None, 256)	16,777,472
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 128)	32,896
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 2)	258
Total params:		17,183,234 (65.55 MB)
Trainable params:		17,182,338 (65.55 MB)
Non-trainable params:		896 (3.50 KB)

scratch-training models and gaining a better understanding of how effective generated image data are at improving generalization without having learned information to fall back on. That is particularly of relevance in scenarios in which pretrained models are not usable or irrelevant and in which we are concerned with discovering how synthetic data are able to help models learn directly off of augmented datasets. Each of our models is run five times to assess stability and monitor changing classification metrics. Each of these actions allows for an in-depth observation of how models' generalization performance on test datasets is impacted by augmented datasets and provides rational insights on how effective using synthetic data is able to improve both untrained and pretrained model accuracy.

4. Experimental Results

This section presents the findings from our experiments.

1) Computation resources: Our experiments are conducted on a Pitzer GPU cluster node from OSC (Ohio Supercomputer Center) with Dual NVIDIA Volta V100 GPUs with 32 GB GPU memory and 48 cores per node at 2.9 GHz. We used Python for the implementation, PyTorch for the generative models, and TensorFlow for the classification models.
2) Dataset: The original dataset for this study, sourced from Kaggle, consists of chest X-ray (CXR) images categorized into two classes: 1,802 NORMAL and 1,800 PNEUMONIA. Each image is originally 256 × 256 pixels in size. However, in order to simulate real-world scenarios, two types of datasets are created: small and imbalanced datasets.
   a. Small dataset: We choose 200 images per class (PNEUMONIA and NORMAL) so that we get a balanced training set. The remaining 1,600 per class are taken for the test set. Such a mini dataset is ideal to evaluate model performance when data are limited and data availability is low, like in clinical applications where unusual medical conditions are involved.
   b. Imbalanced dataset: We generate an imbalanced dataset by randomly selecting 1,500 images of the NORMAL class and 200 images of the PNEUMONIA class for training. For validation and test purposes, we generate three different imbalanced test sets by randomly selecting 300 images of the NORMAL class and 100 images of the PNEUMONIA class. Each of these test sets is used for validation and test for three different models, and the average of the performance measures is taken for correct comparison. This is an imbalanced dataset of the type often found in medical datasets where some of the conditions are not sufficiently represented (e.g., pneumonia).

To ensure diversity and robustness, two different sampling methods are employed:

1) Random sampling: Images are randomly selected from the full dataset, similarly to datasets in practice where available data are frequently uncurated and randomly sampled. This allows for a more natural sampling of images but is not necessarily capable of capturing the diversity of the dataset and thus often constrains generative model performance.
2) Greedy K sampling: Images are sampled according to their dissimilarity to others for obtaining higher diversity representation of the training set. Computational cost is minimized by looking at only a smaller set of very diverse data on which to create synthetic images at an efficient rate. Higher diversity of sampled data leads to creating synthetic images with higher variability, then facilitating better generalization of models.

The combination of these approaches results in four distinct datasets: a small and an imbalanced dataset for each sampling method. These datasets are then used for training the classification models and assessing the impact of synthetic images generated by DDPM and PGGANs (Table 3).

4.1. Synthetic image generation

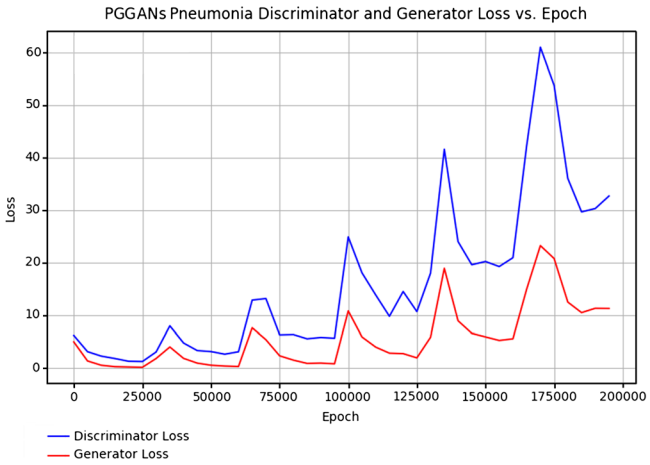
The PGGAN and DDPM models are trained separately for each class in the training dataset, producing a total of four models using 200 images from the small dataset for each sampling method. Leveraging the code from Hugging Face, we generate 2,000 images per class for each model.

To train PGGAN models, Random Sampling from a standard normal distribution is employed for initialization. Stability in training is achieved by equalizing the learning rate, that is, scaling the outputs right before the forward pass [6]. Convolution layers below a 64-pixel resolution are set at 128 filters, while layers at 64- and 128-pixel resolutions are set to 64 filters. The BATCH-SIZE is set to 4. One PGGAN model is trained per class using the Adam optimizer and the Wasserstein loss, each for 200,000 epochs. Due to computational constraints, the models did not converge, though the training process was stable and followed a desired pattern of loss (Figure 4). Each spike reflects the network's temporary destabilization when a new resolution level is introduced, followed by a return to more stable behavior as the model adapts to the increased complexity. This pattern shows that PGGANs maintained balanced training despite fluctuations, adapting effectively during progressive layer growth. With experimental trials relying on computed losses, we choose the checkpoint from epochs 160,000 (PGGANs 160k).

Table 3
Dataset overview

Dataset Type	Sampling Method	Training Data (NOR, PNE)	Test Data (NOR, PNE)
Original dataset	-	1802, 1800	-
Small dataset	Random	200, 200	1602, 1600
	Greedy K	200, 200	1602, 1600
Imbalanced dataset	Random	1500, 300	3 sets of (300, 100)
	Greedy K	1500, 300	3 sets of (300, 100)

Figure 4
PGGAN training loss—PNEUMONIA class



The DDPM model hyperparameters include an image size of 128 pixels, a batch size of 16, a learning rate of 1e−4, 512 epochs, 8,000 timesteps, and mixed precision ("fp16") to reduce memory use and speed up data transfer.

4.2. Synthetic image quality evaluation

4.2.1. Visual inspection

Figures 5 and 6 showcase a visual comparison of generated images of both healthy and pneumonia-affected lungs. Although visual inspection can be subjective, the PGGAN images from the 160k checkpoint are visually appealing but occasionally display defect patches. In contrast, the DDPM-generated images demonstrate a closer resemblance to the original data, exhibiting superior visual fidelity.

Figure 5
Pneumonia images: original vs. generated by DDPM vs. generated by PGGANs. Synthetic images exhibit high similarity to the originals

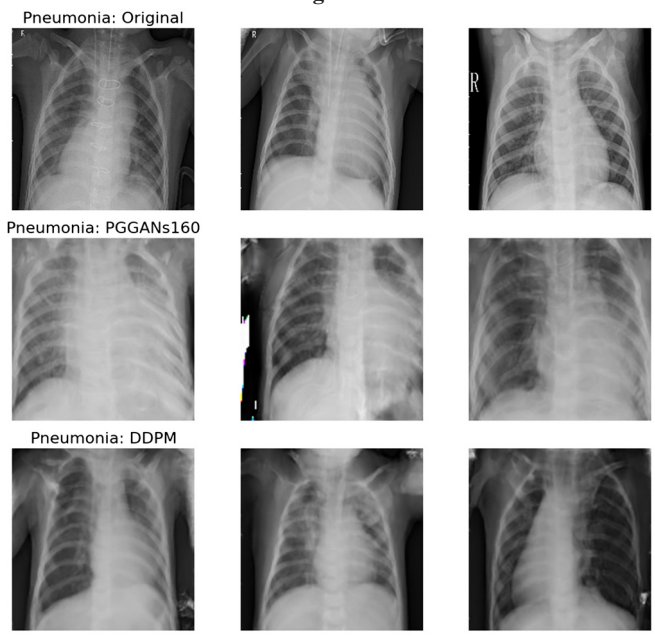
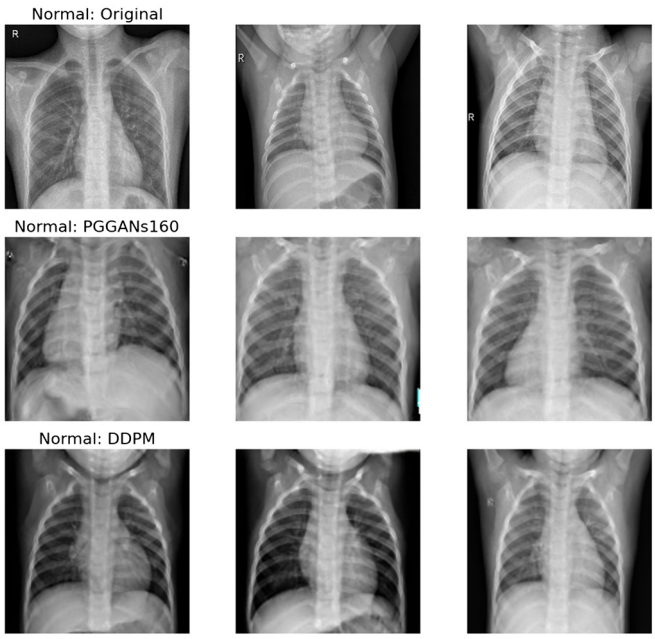


Figure 6
Normal images: original vs. synthetic



4.2.2. FID metric

PyTorch implementation provided by Khazrak et al. [33] is used to calculate FID scores between the original dataset and each model's generated images per class (Figure 7).

Figure 7 shows FID metrics for the DDPM and PGGAN models in two scenarios—Random and Greedy K Sampling methods—after generating 2,000 synthetic images for both the NORMAL and PNEUMONIA labels.

Across both sampling methods, the FID scores for DDPM are consistently lower than those for PGGANs, indicating that DDPM generates more realistic synthetic images that are closer to the real data distribution. In both models, the FID values for the PNEUMONIA label are higher than those for the NORMAL label, suggesting that generating realistic PNEUMONIA images is more challenging.

However, the impact of the sampling method is evident in the significantly higher FID scores observed with the Greedy K Sampling method compared with Random Sampling, especially for PGGANs. The Greedy K Sampling method selects more diverse and distinct samples, which encourages the generative models to produce a wider variety of images, including rare or uncommon patterns in the dataset. While this increases image diversity, it also makes it harder for the models to maintain fidelity to the real data, leading to higher FID scores—particularly for the PNEUMONIA label, which is more difficult to generate accurately.

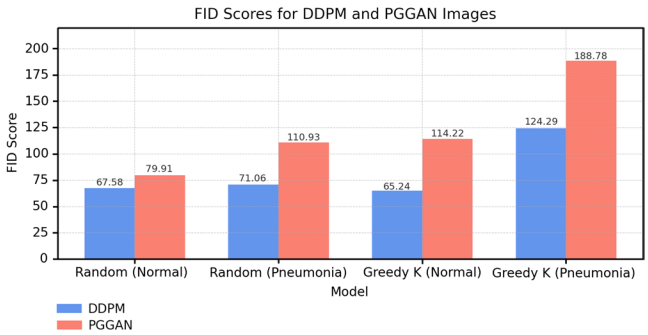
4.3. Experimental evaluation of classification models

To assess the impact of data augmentation using generative models, we conducted two distinct experiments using two different sampling methods: the Random Sampling and Greedy K Sampling methods. For each sampling method, we performed two experiments—one on a small dataset and another on an imbalanced dataset.

Four models were evaluated in each experiment: a custom CNN (trained for 20 epochs), an untrained VGG16 (10 epochs), a pretrained VGG16 (5 epochs), and a pretrained ResNet50 (5 epochs). The pretrained weights for both the VGG16 and ResNet50 models were sourced from the ImageNet Large Scale Visual Recognition Challenge dataset, which consists of 1.2 million images categorized into 1,000 classes. The following datasets were used for model training for each sampling method:

1) Small dataset:
   a. Baseline: Training dataset with a total size of 400 images.
   b. DDPM Mix: Training dataset augmented with DDPM-generated synthetic images, resulting in a total size of 4,400 images.

Figure 7
FID score comparison



c. PGGAN Mix: Training dataset augmented with PGGAN-generated synthetic images, resulting in a total size of 4,400 images.

2) Imbalanced dataset:
   a. Baseline: Training dataset with a total of 1,700 images (1,500 NORMAL, 200 PNEUMONIA).
   b. DDPM Mix: Training dataset augmented with DDPM-generated synthetic images, leading to 3,700 images (1,500 NORMAL, 2,200 PNEUMONIA).
   c. PGGAN Mix: Training dataset augmented with PGGAN-generated synthetic images, leading to 3,700 images (1,500 NORMAL, 2,200 PNEUMONIA).

All hyperparameters were kept consistent across experiments. Models were evaluated using accuracy, recall, precision, and F1 Score.

Table 4
Accuracy ± SD: Random Sampling, Balanced Dataset

Custom CNN	0.90 ± 0.035	**0.93 ± 0.011**	0.92 ± 0.011
Untrained VGG16	0.86 ± 0.039	**0.92 ± 0.018**	0.88 ± 0.016
Pretrained VGG16	0.93 ± 0.007	**0.95 ± 0.005**	0.92 ± 0.019
Pretrained ResNet50	0.93 ± 0.015	0.93 ± 0.017	**0.94 ± 0.002**

Table 5
Accuracy ± SD: Random Sampling, Imbalanced Dataset

Custom CNN	0.95 ± 0.014	**0.95 ± 0.011**	0.95 ± 0.018
Untrained VGG16	0.86 ± 0.028	**0.92 ± 0.008**	0.89 ± 0.030
Pretrained VGG16	0.94 ± 0.019	0.94 ± 0.019	**0.94 ± 0.018**
Pretrained ResNet50	0.95 ± 0.017	**0.95 ± 0.010**	0.95 ± 0.016

To ensure robustness and model stability, each model was run five times, with training and validation data shuffled for each run.

4.3.1. Random sampling

We employed random sampling to generate both small and imbalanced datasets, simulating real-world scenarios where data distribution is often uneven and limited in size. In this section, we present the results of these experiments, highlighting the impact of synthetic data augmentation on model performance under these challenging conditions (Tables 4 and 5 and Figure 8).

1) Small dataset:

The custom CNN performance on DDPM- and PGGAN-augmented datasets improves distinctly on both datasets in comparison to using the base dataset itself. On the original dataset, the algorithm performed with mean accuracy, F1 Score, recall, and precision of 0.90 and standard deviation (SD) of 0.035, demonstrating medium variability between runnings. For the DDPM-augmented dataset, these were elevated to 0.93 and SD was lowered to 0.011 to reflect more consistent performance. For both datasets generated by PGGANs to expand them, performance was also improved and accuracy and other measures were elevated to 0.92. Variability was also lowered to 0.011 SD.

For untrained VGG16 model results, significant improvements were obtained when using both DDPM- and PGGAN-augmented datasets on the original dataset. For the DDPM-augmented dataset, mean accuracy, F1 metric, recall, and precision were improved drastically to 0.92. SD dropped to 0.021, indicating a much stable result per run. There is a significant increase in both performance and stability obtained due to the DDPM-generated synthetic data.

Pretrained VGG16 exhibited stable performance on all datasets, and maximum stability and accuracy were obtained on the DDPM-augmented dataset. Average accuracy was increased to 0.95. SD was kept low at 0.017 and accounted for highly consistent performance.

Figure 8
Accuracy and F1 scores across datasets

Pretrained ResNet50 performed best on an augmented PGGAN dataset with an average accuracy of 0.94. SD was 0.002, indicating high stable performance on a number of runs.

2) Imbalanced dataset:

Even though custom CNN's performance was consistent for all datasets, DDPM augmentation had greater stability with lower SD than original and PGGAN-augmented datasets. PGGANs marginally raised variability, and so, DDPM is chosen for stability.

For the untrained VGG16 network on DDPM-augmented data, accuracy also increased significantly to 0.92. SD was also reduced to 0.009 to capture more stable performance in each run and to reflect better stability than both original data and PGGAN-augmented data.

For both pretrained VGG16 and ResNet50 networks, the augmented datasets had smaller SD, indicating greater stability and reduced run-to-run performance variability.

4.3.2. Greedy K sampling

In the next experiment, we used the Greedy K Sampling method to create small and imbalanced datasets and this section highlights the impact of synthetic data augmentation on model performance (Tables 6 and 7 and Figure 8).

The custom CNN model made a spectacular increase in accuracy and stability with the augmented datasets. Accuracy was increased to 0.93 (DDPM) from 0.89 (original), and SD decreased to 0.011 from 0.035 to display a stable improvement in performance.

For untrained VGG16, DDPM raised accuracy considerably from 0.85 (original) to 0.94 and decreased SD from 0.040 to 0.003, considerably higher stability with DDPM augmentation.

For pretrained VGG16, accuracy was increased from 0.95 to 0.96 by using the DDPM-augmented dataset with a corresponding SD drop from 0.011 to 0.008, exhibiting increased stability.

For pretrained ResNet50, the DDPM dataset also maintained the original's same level of high accuracy of 0.96 and no increased model stability was observed after adding artificially generated data.

1) Imbalanced dataset:

For custom CNN models, both PGGANs and DDPM improved accuracy and PGGANs obtained the highest accuracy of 0.97 yet DDPM obtained higher stability. For untrained VGG16, both augmentations significantly increased accuracy to 0.92 and DDPM obtained better stability due to less variability.

For pretrained VGG16, PGGANs obtained a slightly higher accuracy of 0.96 and a SD of 0.005 indicating superior stability than that for the original dataset.

Table 6
Accuracy ± SD: Greedy K Sampling, Balanced Dataset

Custom CNN	0.89 ± 0.035	**0.93 ± 0.011**	0.93 ± 0.020
Untrained VGG16	0.85 ± 0.040	**0.94 ± 0.003**	0.87 ± 0.032
Pretrained VGG16	0.95 ± 0.011	**0.96 ± 0.008**	0.95 ± 0.008
Pretrained ResNet50	**0.96 ± 0.004**	0.96 ± 0.012	0.93 ± 0.018

Table 7
Accuracy ± SD: Greedy K Sampling, Imbalanced Dataset

Custom CNN	0.95 ± 0.014	0.95 ± 0.012	**0.97 ± 0.012**
Untrained VGG16	0.87 ± 0.028	**0.92 ± 0.010**	0.92 ± 0.011
Pretrained VGG16	0.96 ± 0.016	**0.97 ± 0.007**	0.96 ± 0.008
Pretrained ResNet50	0.95 ± 0.006	**0.96 ± 0.006**	0.96 ± 0.007

Finally, for pretrained ResNet50, PGGANs obtained a moderate improvement on both stability and accuracy.

Augmentation for robustness improvement using synthetic images is attributed to augmented diversity and balance injected in training data. For smaller and unbalanced datasets, models overfit and learn non-generalizable patterns due to unchanging data. Augmentation using synthetic images, specifically those generated by DDPM, increases diversity and robustness to training data by adding fresh and diverse samples to the training dataset better representing underlying data distribution. Overfitting is reduced, and the model is able to learn generalized characteristics that are not specific to data variations used to train it but to underlying data distribution. Additionally, augmentation by synthetic data helps in balancing datasets such that minority classes are represented well to facilitate better identification of rare states by the model by enhancing better robustness of it. Through these two processes, increased stability and accuracy are obtained in successive runnings and models are less affected by data variations and are therefore robust. Our code and implementation are also available in publicly available code repository [34].

5. Limitations and Future Work

1) Limitations: This study is limited to a single public CXR dataset and a binary task (NORMAL vs. PNEUMONIA), without external cross-dataset validation. To reduce computational cost and enable a controlled DDPM–PGGAN comparison under small/imbalanced data, we generated images at 128 × 128 resolution ; this choice may limit fine-grained anatomical detail. DDPM is more computationally expensive than PGGANs in both training and inference. As with any synthetic augmentation, there is a risk of overfitting to the synthetic distribution or artifacts not present in the real data. We did not employ clinician review to clean or prune synthetic images prior to training in this study; by contrast, in related work on laryngeal lesion classification, two domain experts screened DDPM-generated images to remove unrealistic samples before model training [35]. Finally, each configuration was run five times; we report mean SD and refrain from null hypothesis testing, which may limit statistical power.
2) Future work: We will (i) conduct cross-dataset (out-of-distribution) validation to assess generalization across institutions; (ii) investigate domain adaptation techniques to mitigate source–target shift; (iii) evaluate rare-disease and extreme-imbalance cohorts where augmentation is most impactful; (iv) study higher-resolution synthesis (e.g., 256–512 px) and quantify trade-offs using FID, SSIM, PSNR, and downstream accuracy; (v) extend beyond binary CXR to multi-class tasks and additional modalities (e.g., CT and MRI); and (vi) increase the number of runs (e.g., 10) to enable adequately powered hypothesis testing where appropriate.

Note: Models trained with DDPM-augmented data consistently achieved higher accuracy and F1 scores, especially in smaller and imbalanced datasets. The reduction in SD across runs indicates improved training stability and better generalization, with DDPM showing more consistent benefits compared with PGGANs.

6. Conclusion

This research used data augmentation by DDPM and PGGANs on small-sized and imbalanced medical image datasets formed by Random Sampling and Greedy K Sampling. In all the experiments, the inclusion of synthetic images increased the classification accuracy, improved generalization, and minimized variability across runs. Comparing the two networks, DDPMs universally beat PGGANs by generating data closer to the original distribution, resulting in low FID

scores and producing a more stable behavior under varying classifiers and sampling regimes.

Random Sampling was the stronger alternative, providing constant increments toward accuracy and stability, whereas Greedy K Sampling contributed added diversity with higher variability. However, DDPM demonstrated great proficiency under both schemes, achieving a balance between fidelity and diversity that benefited models ranging from customized CNNs to pretrained models.

The general enhancements come from the added balance and diversity inherent to synthetic data that lessen overfitting and enhance minority class representation. These benefits were most apparent within untrained models, where synthetic augmentation enabled networks to learn more generalized characteristics and greater stability than when using real data.

Overall, generative models—particularly DDPM—are an efficient and dependable solution to the long-lasting problems of small-sized and imbalanced medical datasets. Generative models improve accuracy, stability, and data variability, aiding the construction of stabler diagnostic models.

Acknowledgement

An earlier version of this work was made available as a preprint [36].

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/imankhazrak/DDPM_X-Ray.

Author Contribution Statement

Iman Khazrak: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Shakhnoza Takhirova:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation,Writing – original draft, Writing – review & editing, Visualization. **Mostafa M. Rezaee:** Validation, Investigation, Resources, Writing – review & editing. **Mehrdad Yadollahi:** Investigation, Resources, Writing – review & editing. **Robert C. Green II:** Methodology, Writing – review & editing, Supervision. **Shuteng Niu:** Methodology, Writing – review & editing, Supervision.

References

[1] Chan, H.-P., Samala, R. K., Hadjiiski, L. M., & Zhou, C. (2020). Deep learning in medical image analysis. In G. Lee & H. Fujita (Eds.), *Deep learning in medical image analysis: Challenges and applications* (pp. 3–21). Springer. https://doi.org/10.1007/978-3-030-33128-3_1

[2] Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abramoff, M., Mendonca, A. M., & Campilho, A. (2018). End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, *37*(3), 781–791. https://doi.org/10.1109/TMI.2017.2759102

[3] Kebaili, A., Lapuyade-Lahorgue, J., & Ruan, S. (2023). Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, *9*(4), 81. https://doi.org/10.3390/jimaging9040081

[4] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, *321*, 321–331. https://doi.org/10.1016/j.neucom.2018.09.013

[5] Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative AI: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260. https://doi.org/10.3390/fi15080260

[6] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of International Conference on Learning Representations*, 1–26.

[7] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 574.

[8] Jagani, S., Deng, X., Hong, P. C., & Mashhadi Nejad, N. (2024). Adopting sustainability business models for value creation and delivery: An empirical investigation of manufacturing firms. *Journal of Manufacturing Technology Management*, *35*(2), 360–382. https://doi.org/10.1108/JMTM-03-2023-0099

[9] Nejad, N. M., & Hong, P. (2024). Developing a competitive advantage in EV supply chain systems: A conceptual framework for national benchmarking studies. In *Midwest Decisions Sciences Conference*, 60–91.

[10] Chen, Y., Liu, J., Peng, L., Wu, Y., Xu, Y., & Zhang, Z. (2024). Auto-encoding variational Bayes. *Cambridge Explorations in Arts and Sciences*, *2*(1), 1–8. https://doi.org/10.61603/ceas.v2i1.33

[11] Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann machines. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, *5*, 448–455.

[12] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, *37*, 2256–2265.

[13] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., ..., & Yang, M.-H. (2024). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, *56*(4), 1–39. https://doi.org/10.1145/3626235

[14] Tang, S. (2020). Lessons learned from the training of GANs on artificial datasets. *IEEE Access*, *8*, 165044–165055. https://doi.org/10.1109/ACCESS.2020.3022820

[15] Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., & Mukhopadhyay, A. (2020). GANs for medical image analysis. *Artificial Intelligence in Medicine*, *109*, 101938. https://doi.org/10.1016/j.artmed.2020.101938

[16] Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., & Reyes, M. (2018). Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *Medical Image Computing and Computer Assisted Intervention: 21st International Conference*, 580–588. https://doi.org/10.1007/978-3-030-00934-2_65

[17] Xu, I. R. L., van Booven, D. J., Goberdhan, S., Breto, A., Porto, J., Alhusseini, M., ..., & Arora, H. (2023). Generative adversarial networks can create high quality artificial prostate cancer

magnetic resonance images. *Journal of Personalized Medicine*, *13*(3), 547. https://doi.org/10.3390/jpm13030547

[18] Chuquicusma, M. J. M., Hussein, S., Burt, J., & Bagci, U. (2018). How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, 240–244. https://doi.org/10.1109/ISBI.2018.8363564

[19] Bermudez, C., Plassard, A. J., Davis, L. T., Newton, A. T., Resnick, S. M., & Landman, B. A. (2018). Learning implicit brain MRI manifolds with deep learning. In *Medical Imaging 2018: Image Processing: Proceedings of SPIE*, *10574*, 105741L. https://doi.org/10.1117/12.2293515

[20] Chen, Y., Yang, X.-H., Wei, Z., Heidari, A. A., Zheng, N., Li, Z., ..., & Guan, Q. (2022). Generative adversarial networks in medical image augmentation: A review. *Computers in Biology and Medicine*, *144*, 105382. https://doi.org/10.1016/j.compbiomed.2022.105382

[21] Park, H. Y., Bae, H.-J., Hong, G.-S., Kim, M., Yun, J., Park, S., ..., & Kim, N. (2021). Realistic high-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: Visual turing test. *JMIR Medical Informatics*, *9*(3), e23328. https://doi.org/10.2196/23328

[22] Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, *139*, 8162–8171.

[23] Rezaeian, O., Bayrak, A. E., & Asan, O. (2024). An architecture to support graduated levels of trust for cancer diagnosis with AI. In *HCI International 2024 Posters: 26th International Conference on Human-Computer Interaction*, 344–351. https://doi.org/10.1007/978-3-031-61966-3_37

[24] Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A. G., & Tamir, J. (2021). Robust compressed sensing mri with deep generative priors. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 1145.

[25] Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., & Cattin, P. C. (2022). Diffusion models for implicit image segmentation ensembles. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, *172*, 1336–1348.

[26] Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., ..., & Truhn, D. (2023). A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, *13*(1), 12098. https://doi.org/10.1038/s41598-023-39278-0

[27] Liang, Z., Xue, Z., Rajaraman, S., & Antani, S. (2024). Covid-19 pneumonia chest X-ray pattern synthesis by stable diffusion. In *2024 IEEE Southwest Symposium on Image Analysis and Interpretation*, 21–24. https://doi.org/10.1109/SSIAI59505.2024.10508671

[28] Liu, X., & Hsieh, C.-J. (2019). Rob-GAN: Generator, discriminator, and adversarial attacker. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11226–11235. https://doi.org/10.1109/CVPR.2019.01149

[29] Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of generative adversarial networks (GANs): An updated review. *Archives of Computational Methods in Engineering*, *28*(2), 525–552. https://doi.org/10.1007/s11831-019-09388-y

[30] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention: 18th International Conference*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[31] Gong, K., Johnson, K., El Fakhri, G., Li, Q., & Pan, T. (2024). PET image denoising based on denoising diffusion probabilistic model. *European Journal of Nuclear Medicine and Molecular Imaging*, *51*(2), 358–368. https://doi.org/10.1007/s00259-023-06417-8

[32] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*, 2017.

[33] Khazrak, I., Zainaee, S., M. Rezaee, M., Ghasemi, M., & C. Green, R. (2025). Feasibility of improving vocal fold pathology image classification with synthetic images generated by DDPM-based GenAI: A pilot study. European Archives of Oto-Rhino-Laryngology, *282*(8), 4139–4153. https://doi.org/10.1007/s00405-025-09443-4

[34] Wang, R., Rezaeian, O., Asan, O., Zhang, L., & Liao, T. (2024). Relationship between heart rate and perceived stress in intensive care unit residents: Exploratory analysis using fitbit data. *JMIR Formative Research*, *8*, e60759. https://doi.org/10.2196/60759

[35] Mashhadi Nejad, N., Alvarado-Vargas, M. J., & Jalali Sepehr, M. (2024). Refining literature review strategies: Analyzing big data trends across journal tiers. *Academy of Management Proceedings*, *2024*(1), 14852. https://doi.org/10.5465/AMPROC.2024.14852abstract

[36] Khazrak, I., Takhirova, S., Rezaee, M. M., Yadollahi, M., Green II, R. C., & Niu, S. (2024). Addressing small and imbalanced medical image datasets using generative models: A comparative study of ddpm and pggans with random and greedy k sampling. *arXiv Preprint:2412.12532*. https://doi.org/10.48550/arXiv.2412.12532

How to Cite: Khazrak, I., Takhirova, S., Rezaee, M. M., Yadollahi, M., Green II, R. C., & Niu, S. (2025). Addressing Small and Imbalanced Medical Image Datasets Using Generative Models. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA52026661