

Addressing Small and Imbalanced Medical Image Datasets Using Generative Models: A Comparative Study of DDPM and PGGANs with Random and Greedy K Sampling

Iman Khazrak*, Shakhnoza Takhirova[†], Mostafa M. Rezaee[‡], Mehrdad Yadollahi[§],
Robert C. Green II[¶], Shuteng Niu^{||}

Department of Computer Science, Bowling Green State University

Bowling Green, OH, USA

Email: *ikhazra@bgsu.edu, [†]takhirs@bgsu.edu, [‡]mostam@bgsu.edu, [§]mehrdad@bgsu.edu,
[¶]greenr@bgsu.edu, ^{||}sniu@bgsu.edu

Abstract—The development of accurate medical image classification models is often constrained by privacy concerns and data scarcity for certain conditions, leading to small and imbalanced datasets. To address these limitations, this study explores the use of generative models, such as Denoising Diffusion Probabilistic Models (DDPM) and Progressive Growing Generative Adversarial Networks (PGGANs), for dataset augmentation.

The research introduces a framework to assess the impact of synthetic images generated by DDPM and PGGANs on the performance of four models: a custom CNN, Untrained VGG16, Pretrained VGG16, and Pretrained ResNet50. Experiments were conducted using Random Sampling and Greedy K Sampling to create small, imbalanced datasets. The synthetic images were evaluated using Frechet Inception Distance (FID) and compared to original datasets through classification metrics.

The results show that DDPM consistently generated more realistic images with lower FID scores and significantly outperformed PGGANs in improving classification metrics across all models and datasets. Incorporating DDPM-generated images into the original datasets increased accuracy by up to 6%, enhancing model robustness and stability, particularly in imbalanced scenarios. Random Sampling demonstrated superior stability, while Greedy K Sampling offered diversity at the cost of higher FID scores. This study highlights the efficacy of DDPM in augmenting small, imbalanced medical image datasets, improving model performance by balancing the dataset and expanding its size. Our implementation and codes are available at https://github.com/imankhazrak/DDPM_X-Ray.

Index Terms—Medical Image Augmentation, Generative Models, Progressive Growing Generative Adversarial Networks (PGGANs), Denoising Diffusion Probabilistic Models (DDPM), Synthetic Data Integration.

I. INTRODUCTION

Medical imaging is indispensable in modern healthcare, guiding diagnostics, surgeries, treatment assessments, and disease monitoring. The growing volume of images poses challenges [1] for radiologists and physicians to maintain workflow efficiency without technological support. There are significant challenges to train accurate and reliable Machine Learning or

Deep Learning diagnostic models. Key issues include the scarcity of extensive and diverse datasets [2], stringent data privacy regulations, and inherent dataset imbalances. These imbalances can lead to biased models that struggle with rare conditions, and even minor errors can have negative implications.

Traditional data augmentation techniques, such as random rotations, flipping, cropping, and noise injection, have been widely employed to expand training datasets. While useful, these methods merely manipulate existing samples and fail to introduce the kind of fundamental variability needed for robust model training [3]. In contrast, generative models [4], GANs and DDPM, have revolutionized image synthesis by creating entirely new data points. These models offer promising solutions to the challenges of imbalanced datasets, particularly in the field of medical imaging, where the availability of labeled data is limited.

Generative models typically require large and diverse datasets [5]. This creates a paradox: if such large labeled datasets were available, effective models could be trained directly. Therefore, generative models are only viable if they can work effectively with scarce datasets. The present paper addresses this challenge by proposing a comprehensive framework for generating synthetic medical images from both small and imbalanced datasets using two generative models: PGGANs [6] and DDPM [7]. We also explore the use of two different sampling methods—Random Sampling and the Greedy K Method—to further assess their impact on model performance. This framework is rigorously evaluated to enhance diagnostic model accuracy and robustness.

This paper introduces a framework for the use of advanced generative models in medical imaging, targeting small and imbalanced datasets. Our key contributions are as follows:

- **Comprehensive Evaluation Framework:** We develop a rigorous framework that evaluates the

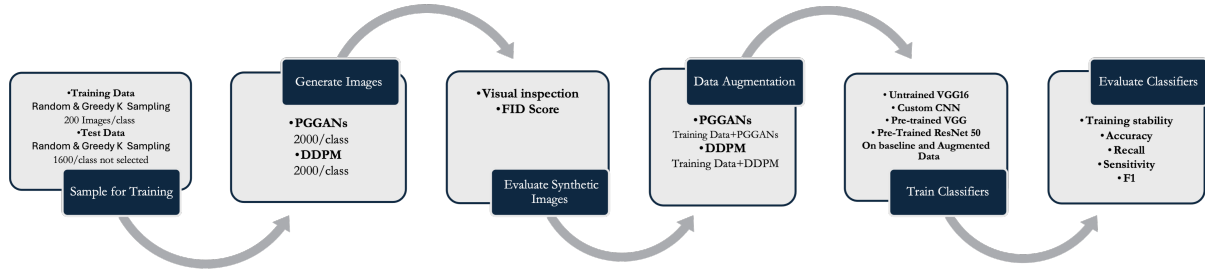


Fig. 1. Implemented comprehensive framework that rigorously evaluates the quality and performance of synthetic images generated by DDPM and PGGAN models.

quality and performance of synthetic images generated by DDPM and PGGANs. This framework consists of three stages. First, synthetic images are generated using both models. Second, the quality of these images is evaluated through visual inspection and quantitative metrics, such as Fréchet Inception Distance (FID), which measures the similarity between generated and real images, and the VGG Expert model for visual validation. Third, we examine the impact of incorporating synthetic images into small and imbalanced datasets on classification model performance.

- **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.
- **Improved Classification Performance:** By integrating synthetic images into small and imbalanced datasets, we show significant improvements in the performance of both custom CNNs and untrained VGG16 models. For example, the accuracy of untrained models trained on small datasets improved significantly.
- **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 summary, this study provides a novel approach to overcoming the limitations of small and imbalanced medical datasets by leveraging the latest generative models. It demonstrates the potential of DDPM and PGGANs to augment data in a way that not only improves the size and balance of datasets but also significantly enhances the accuracy, stability, and robustness of classification models in medical imaging. The following sections provide an overview of the

generative models, followed by a detailed discussion of DDPM and PGGANs, their applications in medical imaging, and an in-depth analysis of our methodology and results.

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

II. RELATED WORK

Generative models, especially those producing high-quality, realistic images, have gained significant attention in augmenting medical datasets, particularly for rare conditions where data scarcity and class imbalance are common. These models can be classified as latent variable generative models, which are either explicit or implicit density models.

Frameworks from other fields offer valuable insights for healthcare. For example, Sustainability Value Articulation (SVA) enhances internal and external efforts for better social and environmental outcomes by emphasizing supplier involvement and technological integration [8]. Similarly, the EV supply chain highlights the importance of systematic benchmarking and technological innovation to achieve competitive advantages in complex systems [9]. These principles align with the goals of generative models in addressing data scarcity and improving healthcare datasets, enabling long-term scalability and impact.

Explicit density models like Variational Autoencoders (VAEs), Boltzmann Machines, and DDPMs have predefined density functions, offering interpretability and training stability [7], [10], [11]. They are useful for applications like anomaly detection due to their explicit likelihood functions. However, their distributional assumptions can sometimes lead to less realistic images [12].

Implicit density models, such as GANs, do not rely on explicit likelihood functions, making them more flexible and capable of modeling complex distributions. They tend to produce more realistic images but

suffer from training instability, evaluation difficulty, and sensitivity to hyperparameters [13], [14].

A. GANs 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 [6], has significantly improved the stability and quality of GAN-generated images. PGGANs utilize a progressive training approach, starting with low-resolution images and gradually increasing the resolution as training progresses. This technique allows the model to learn coarse features before fine details, leading to more stable training and higher-quality images.

In medical imaging, GANs mainly have been used to enhance classification and segmentation deep learning models [15]. The work in [2] uses GANs on a small CT scan dataset to generate eye fundus images which confirm to the given masks. [16] also used mask to generate lung images, and only those synthetic images that fulfilled informativeness criteria calculated by Bayesian neural networks were used to improve the classifier model. In [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], brain MRI images [19] to name a few. The authors of [20] generate high resolution synthetic images of skin lesions from a dataset of 2000 dermoscopic images using multiple GANs architectures and compare their classification performances. They conclude that PGGANs could to synthesize realistic images that medical professionals upon evaluation were not able to distinguish from real ones. Results of [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.

B. 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. in 2020 [22], DDPMs have set new benchmarks in image generation quality by leveraging a sophisticated noise schedule and a robust denoising network.

The application of DDPMs in medical imaging has been explored in various tasks. Nichol and Dhariwal [23] demonstrated the effectiveness of guided diffusion models combined with upsampling techniques in improving MRI resolution, facilitating better diagnosis and treatment planning using a dataset of 10,000 MRI images. In medical imaging, the integration of explainability and trust into AI systems has proven to be crucial for clinician acceptance, particularly for life-threatening diseases like cancer. Rezaeian et al. [24] propose a dual-stage AI model for breast cancer diagnosis, introducing graduated levels of explainability such as tumor localization and probability distributions, which have been shown to significantly enhance trust in AI systems. Similarly, our work focuses on improving the robustness of AI models by addressing data scarcity and imbalance, which are pivotal challenges in building reliable diagnostic tools. Jalal et al. [25] further explored DDPMs for MRI reconstruction, showing significant improvements in image quality and robustness against noise with a dataset of 3,500 MRI images. Similarly, Wolleb et al. [26] applied DDPMs to medical image segmentation, achieving state-of-the-art results with a dataset of 7,500 images. Yoon et al. [27] compared latent DDPMs and GANs for medical image synthesis across multiple modalities using 8,000 CT and MRI images, finding DDPMs superior in image quality and diversity. Zhang et al. [28] introduced a DDPM-based X-ray Image Synthesizer using 6,000 X-ray images, demonstrating the model's ability to generate high-fidelity synthetic X-ray images, which can augment training datasets and improve diagnostic 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.

III. METHODOLOGY

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

A. Image Synthesis

1) *PGGANs*: PGGANs utilize a progressive training approach, starting with low-resolution images and gradually increasing the resolution as training progresses (Fig. 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 [29]. The adversarial loss functions for the generator (\mathcal{L}_G) and discriminator (\mathcal{L}_D) are:

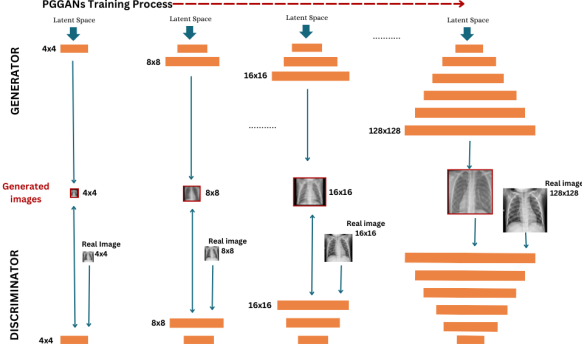


Fig. 2. PGGANs gradually increase both the resolution of the generated images and the complexity of the generator and discriminator networks during training.

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

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

where $G(z)$ represents the generated data from noise z , and $D(x)$ represents the discriminator's output for real data x [30].

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.

2) *DDPM*: 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 (Fig. 3) [7].

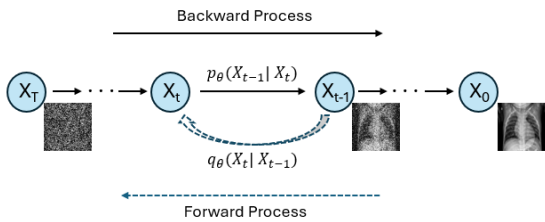


Fig. 3. The directed graphical model of DDPM. The forward process gradually adds noise to an image until it becomes pure noise. The backward process then learns to reverse this process, reconstructing the original image from noise.

The forward process adds noise to the image:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1 - \alpha_t}\epsilon_t$$

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 aims to denoise the noisy image obtained from the forward process and recover the original clean image by optimizing the variational lower bound:

$$L_{DDPM} = \mathbb{E}_{t, x_0, \epsilon} [|\epsilon - \epsilon_\theta(x_t, t)|^2]$$

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 [31], [32]. U-Net's U-shaped structure with down-sampling and upsampling paths efficiently synthesizes detailed images, incorporating time embeddings to adjust noise prediction based on the reverse process timestep [33].

B. Generated Image Assessment

a) *Visual Inspection*: Generated images are initially evaluated by visually comparing random samples to the original images.

b) *Frechet Inception Distance (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, then computing the Frechet distance between the resulting multivariate Gaussian distributions. A higher FID score indicates greater dissimilarity [34].

C. Classification Models

We evaluate the impact of synthetic images using four different classifiers: a pre-trained VGG16, a pre-trained ResNet50, an untrained VGG16, and a custom CNN. Each model is first trained on both the small and imbalanced datasets to establish baselines, followed by training on the same datasets augmented with synthetic images generated by DDPMs and PGGANs. The inclusion of an untrained VGG16 allows us to assess the direct impact of synthetic data on models learning from scratch, providing a clearer understanding of how effective the generated images are in improving generalization without the benefit of pre-learned features. This approach is particularly important in cases where pretrained models may not be available or applicable, and where the goal is to evaluate how synthetic data can help models learn directly from the augmented datasets. To assess stability and observe changes in classification metrics, each model runs five times. This approach allows for a thorough examination of how the augmented datasets influence the models' generalization performance on the test datasets, providing valuable insights into the effectiveness of synthetic data in enhancing both pre-trained and untrained model accuracy.

IV. EXPERIMENTAL RESULTS

This section presents the findings from our experiments.

TABLE I
VGG16/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 II
CUSTOM CNN MODEL SUMMARY

Layer (Type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 128, 128, 64)	1,792
batch_normalization_1 (BatchNormalization)	(None, 128, 128, 64)	256
max_pooling2d_1 (MaxPooling2D)	(None, 64, 64, 64)	0
conv2d_2 (Conv2D)	(None, 64, 64, 128)	73,856
batch_normalization_2 (BatchNormalization)	(None, 64, 64, 128)	512
max_pooling2d_2 (MaxPooling2D)	(None, 32, 32, 128)	0
conv2d_3 (Conv2D)	(None, 32, 32, 256)	295,168
batch_normalization_3 (BatchNormalization)	(None, 32, 32, 256)	1,024
max_pooling2d_3 (MaxPooling2D)	(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)

a) *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.

b) *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 256x256 pixels. However, in order to simulate real-world scenarios, two types of datasets are created: small and imbalanced datasets.

- **Small Dataset:** We sample 200 images from each class (NORMAL and PNEUMONIA) to create a balanced training dataset. The remaining 1,600 images per class form the test set. This small dataset helps in evaluating model performance under conditions where data availability is limited, mimicking clinical situations with rare medical conditions.
- **Imbalanced Dataset:** We construct an imbalanced dataset by sampling 1,500 images from the NORMAL class and 200 images from the PNEUMONIA class for training. For testing, we create three separate imbalanced test sets by randomly sampling 300 images from the NORMAL class and 100 images from the PNEUMONIA class. The models are tested on all three test sets, and the average performance metrics are computed to ensure reliable comparisons. This imbalanced dataset reflects the distribution often found in medical datasets, where certain conditions (like

pneumonia) are underrepresented.

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

- 1) **Random Sampling:** Images are selected randomly from the larger dataset, closely mimicking real-world datasets where the available data is typically uncurated and randomly distributed. This method ensures a more natural distribution of images but may not always capture the diversity of the dataset, which can limit the performance of generative models.
- 2) **Greedy K Sampling:** Images are selected based on their dissimilarity to others, ensuring a more diverse representation within the training set. This method helps to reduce computational cost by focusing on a smaller, highly diverse subset of the data, making it more efficient to generate synthetic images. The increased diversity of the selected data leads to the generation of synthetic images that exhibit greater variability, ultimately improving model generalization.

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 III)

TABLE III
DATASET OVERVIEW

Dataset Type	Sampling Method	Training Data (NORMAL, PNEUMONIA)	Test Data (NORMAL, PNEUMONIA)
Original Dataset	-	1802, 1800	-
Small Dataset	Random	200, 200	1602, 1600
	Greedy K	200, 200	1602, 1600
Imbalanced Dataset	Random	1500, 300	3 Imbalanced Test Sets (300, 100) each
	Greedy K	1500, 300	3 Imbalanced Test Sets (300, 100) each

A. Synthetic Image Generation

The PGGANs 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 [35] and [36], we generate 2,000 images per class for each model.

To train PGGANs models, random sampling from a standard normal distribution is employed for initialization. Stability in training is achieved by equalizing the learning rate, i.e. scaling the outputs right before the forward pass [6]. Convolution layers below 64-pixel resolution are set at 128 filters, while layers at 64 and 128-pixel resolution are set to 64 filters. The BATCH-SIZE is set to 4. One PGGANs 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 desired pattern of loss (Fig 4). With experimental trials relying on computed losses, we choose the checkpoint from epochs 160000 (PGGANs 160k).

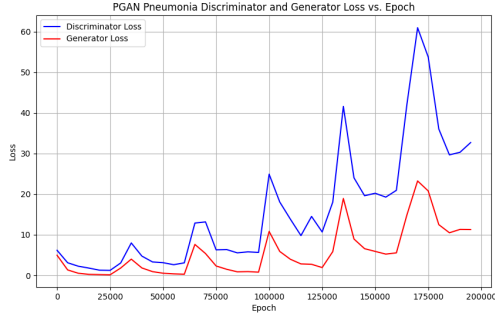


Fig. 4. PGGANs model training loss for Pneumonia class. Plot indicates the training process is stable where loss function increases as model adds new layers.

The DDPM model hyperparameters include an image size of 128 pixels, a batch size of 16, a learning rate of $1e-4$, 512 epochs, 8000 timesteps, and mixed precision (“fp16”) to reduce memory use and speed up data transfer.

B. Synthetic Image Quality Evaluation

1) *Visual Inspection*: Figs 5 and 6 showcase a visual comparison of generated images of both healthy and pneumonia-affected lungs. Although visual inspection can be subjective, the PGGANs 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.

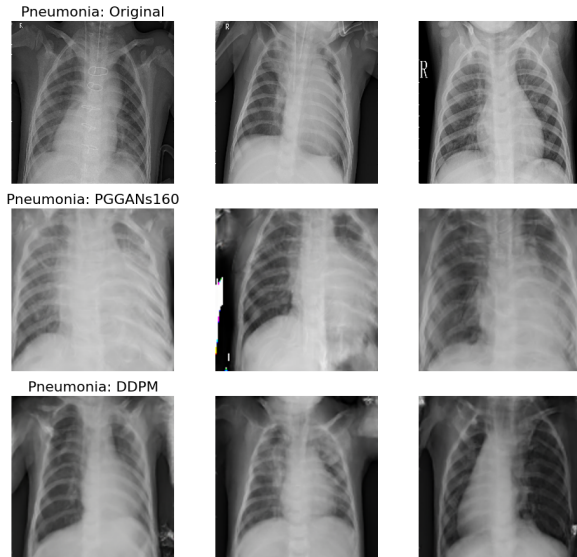


Fig. 5. Pneumonia images: Original vs. generated by DDPM vs. generated by PGGANs. Synthetic images exhibit high similarity to the originals.

2) *FID Metric*: PyTorch implementation provided by [37] is used to calculate FID scores between original data set and each model’s generated images per class (Fig.7).

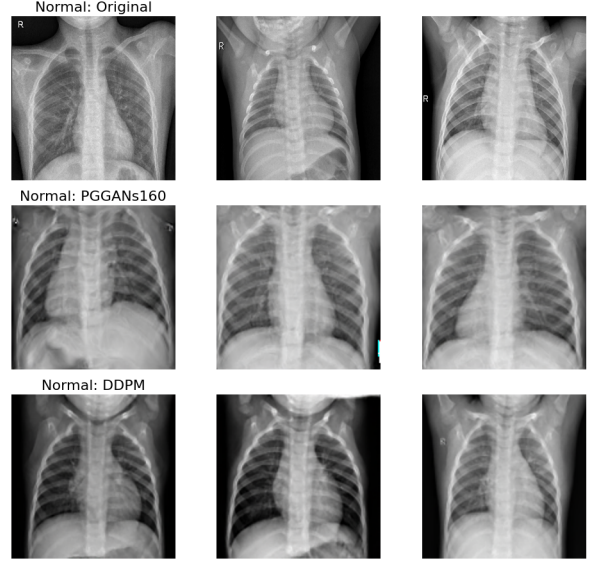


Fig. 6. NORMAL images: Original vs. generated by DDPM vs. generated by PGGANs. Synthetic images exhibit high similarity to the originals.

Fig.7 shows FID metrics for the DDPM and PGGANs models in two scenarios—Random and Greedy K sampling methods—after generating 2000 synthetic images for both the NORMAL and PNEUMONIA labels.

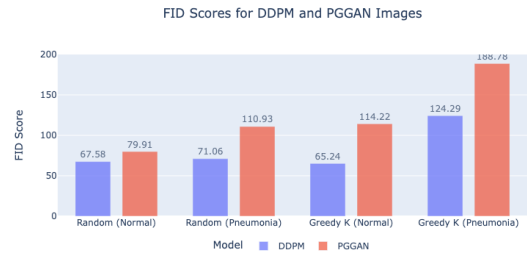


Fig. 7. Comparison of FID Scores between original and synthetically generated images using DDPM and PGGANs.

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 method compared to Random Sampling, especially for PGGANs. The Greedy K 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.

Here’s a more professional version of your section:

C. Classification Models

To assess the impact of data augmentation using generative models, we conducted two distinct experiments using two different sampling methods: Random Sampling and the Greedy K Method. 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 (ILSVRC) 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

- **Baseline:** Training dataset with a total size of 400 images.
- **DDPM Mix:** Training dataset augmented with DDPM-generated synthetic images, resulting in a total size of 4,400 images.
- **PGGANs Mix:** Training dataset augmented with PGGANs-generated synthetic images, resulting in a total size of 4,400 images.

2) Imbalanced Dataset

- **Baseline:** Training dataset with a total of 1,700 images (1,500 NORMAL, 200 PNEUMONIA).
- **DDPM Mix:** Training dataset augmented with DDPM-generated synthetic images, leading to 3,700 images (1,500 NORMAL, 2,200 PNEUMONIA).
- **PGGANs Mix:** Training dataset augmented with PGGANs-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. To ensure robustness and model stability, each model was run 5 times, with training and validation data shuffled for each run.

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 IV, V and Fig 8).

Small Dataset:

TABLE IV
ACCURACY WITH STANDARD DEVIATION (SD) FOR MODELS
USING RANDOM METHOD ON BALANCED DATASET

Model	Original	DDPM	PGGANs
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 V
ACCURACY WITH STANDARD DEVIATION (SD) FOR MODELS
USING RANDOM METHOD ON IMBALANCED DATASET

Model	Original	DDPM	PGGANs
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

- The results for the Custom CNN model show a clear improvement when using both DDPM and PGGANs-augmented datasets compared to the original dataset. On the original dataset, the model achieved an average accuracy, F1 score, recall, and precision of 0.90, with a standard deviation of 0.035, indicating moderate variability between runs. With the DDPM-augmented dataset, these metrics improved to 0.93, and the standard deviation dropped significantly to 0.011, reflecting more consistent performance. The PGGANs-augmented dataset also enhanced performance, with accuracy and other metrics increasing to 0.92. The variability was similarly reduced, with a standard deviation of 0.011.
- The results for the Untrained VGG16 model show significant improvements when using both DDPM and PGGANs-augmented datasets compared to the original dataset. With the DDPM-augmented dataset, the average accuracy, F1 score, recall, and precision improved substantially to 0.92. The standard deviation decreased to 0.021, indicating more consistent results across runs. This reflects a significant enhancement in both performance and stability due to the DDPM-generated synthetic images.
- The Pretrained VGG16 model performed strongly across all datasets, with the highest accuracy and stability observed with the DDPM-augmented dataset. The average accuracy increased to 0.95. The standard deviation remained low at 0.017, showing highly consistent performance.
- The Pretrained ResNet50 model performed best on the PGGANs-augmented dataset, achieving an average accuracy of 0.94. The standard deviation was 0.002, indicating high stable performance across multiple runs.

Imbalanced Dataset:

- While the accuracy for Custom CNN remained consistent across all datasets, the DDPM augmentation provided better stability with lower standard deviation compared to the original

and PGGANs-augmented datasets. PGGANs increased variability slightly, making DDPM the better option for stability.

- For the untrained VGG16 model with the DDPM-augmented dataset, accuracy saw a substantial improvement, reaching 0.92. Additionally, the standard deviation dropped to 0.009, indicating more consistent performance across runs and demonstrating enhanced stability compared to both the original dataset and the PGGAN-augmented dataset.
- For both the pretrained VGG16 and ResNet50 models, the augmented datasets showed a reduction in standard deviation, indicating enhanced stability and more consistent performance across runs.

2) *Greedy K Sampling*: In the next experiment, we used the Greedy K method to create small and imbalanced datasets and this section highlights the impact of synthetic data augmentation on model performance (TABLES VI, VII and Fig 8).

TABLE VI
ACCURACY WITH STANDARD DEVIATION (SD) FOR MODELS
USING GREEDY K METHOD ON BALANCED DATASET

Model	Original	DDPM	PGGANs
Custom CNN	0.89 \pm 0.035	0.93 \pm 0.011	0.93 \pm 0.020
Untrained VGG16	0.85 \pm 0.040	0.94 \pm 0.003	0.87 \pm 0.032
Pretrained VGG16	0.95 \pm 0.011	0.96 \pm 0.008	0.95 \pm 0.008
Pretrained ResNet50	0.96 \pm 0.004	0.96 \pm 0.012	0.93 \pm 0.018

TABLE VII
ACCURACY RESULTS FOR MODELS USING GREEDY K METHOD
ON IMBALANCED DATASET WITH STANDARD DEVIATION

Model	Original	DDPM	PGGANs
Custom CNN	0.95 \pm 0.014	0.95 \pm 0.012	0.97 \pm 0.012
Untrained VGG16	0.87 \pm 0.028	0.92 \pm 0.010	0.92 \pm 0.011
Pretrained VGG16	0.96 \pm 0.016	0.97 \pm 0.007	0.96 \pm 0.008
Pretrained ResNet50	0.95 \pm 0.006	0.96 \pm 0.006	0.96 \pm 0.007

- The Custom CNN model saw a significant boost in both accuracy and stability with the augmented datasets. Accuracy increased from 0.89 (original) to 0.93 (DDPM), with the standard deviation dropping from 0.035 to 0.011, indicating more stable performance.
- For the untrained VGG16, DDPM significantly boosted accuracy from 0.85 (original) to 0.94, with the standard deviation decreasing from 0.040 to 0.003, demonstrating much greater stability with DDPM augmentation.
- The pretrained VGG16 model saw an accuracy increase from 0.95 to 0.96 with the DDPM-augmented dataset, accompanied by a reduction in standard deviation from 0.011 to 0.008, indicating improved stability.
- For the pretrained ResNet50, the DDPM dataset maintained the same high accuracy of 0.96 as the original dataset, with no improvement in model stability observed after adding synthetic images.

Imbalanced Dataset:

- For Custom CNN, both DDPM and PGGANs improved accuracy, with PGGANs reaching the highest accuracy of 0.97, but DDPM offered better stability.
- For Untrained VGG16, both augmentations significantly improved accuracy to 0.92, with DDPM providing better stability due to lower variability.
- For the pretrained VGG16, PGGANs resulted in a slight accuracy increase to 0.96, while the standard deviation dropped to 0.005, indicating enhanced stability compared to the original dataset.
- Finally, for the pretrained ResNet50, PGGANs led to a slight improvement in both accuracy and stability.

The improvement in robustness by adding synthetic images can be attributed to the increased diversity and balance introduced to the training data. In small and imbalanced datasets, models are often prone to overfitting, learning patterns that are not generalizable due to the lack of variety in the data. By incorporating synthetic images, especially those generated by DDPM, the training dataset is expanded with new, diverse samples that better represent the underlying data distribution. This reduces overfitting, enabling the model to learn more generalized features. Additionally, the added synthetic data helps balance the dataset, ensuring that minority classes are adequately represented, which is crucial for improving the model's ability to correctly classify rare conditions. Together, these factors contribute to the enhanced stability and accuracy observed across multiple runs, making the models more robust to variations in the data.

V. CONCLUSION

The experiments conducted in this study evaluated the effects of data augmentation using DDPM and PGGANs on small and imbalanced medical image datasets, utilizing two sampling methods: Random Sampling and Greedy K. The use of generative models for data augmentation led to significant improvements in both the accuracy and robustness of the models. Across all experiments, the introduction of synthetic images enhanced performance, with models exhibiting greater stability and reduced variability during training. This resulted in higher accuracy and better generalization, particularly when handling limited and imbalanced datasets. Of the two generative models, DDPM consistently outperformed PGGANs, delivering superior gains in accuracy and robustness. DDPM-generated images more closely aligned with the original data distribution, contributing to more stable performance and lower variability across all datasets and sampling methods..

The FID evaluation showed that DDPM outperformed PGGANs with lower FID scores, indicating that DDPM-generated images more closely resembled the original data. Random Sampling led to lower FID



Fig. 8. Accuracy and F1 Scores for classification models on different datasets.

scores across both models, suggesting better fidelity, while the Greedy K method, though encouraging diversity, resulted in higher FID scores, particularly for PGGANs.

In terms of performance on small and imbalanced datasets, DDPM consistently improved accuracy and reduced variability across models. For instance, the Custom CNN model’s accuracy increased to 0.93 on the small dataset and 0.95 on the imbalanced dataset with DDPM, both showing reduced variability. In contrast, while PGGANs increased accuracy, it also introduced higher variability, making it less stable. This was particularly evident for the Untrained VGG16 model, where DDPM raised accuracy to 0.94 with minimal variability, while PGGANs struggled to reduce variability.

Random Sampling proved to be the most stable approach, providing consistent improvements in accuracy and reduced variability across models when paired with DDPM-generated images. This method reflects real-world scenarios, where data is often limited and randomly distributed, making it a practical choice when working with small datasets. On the other hand, Greedy K Sampling—which selects a subset of diverse images—offers a different advantage. While it led to higher variability (higher FID), especially for PGGANs, it still achieved high accuracy when combined with DDPM, as seen in the Pretrained VGG16, which reached 0.97 accuracy with minimal variability. Greedy K is particularly beneficial when we aim to train generative models on a smaller, more diverse subset of images, reducing computational costs while

generating synthetic images that not only promote diversity but also help the model achieve greater stability and accuracy. This method allows for efficient use of resources while ensuring that the generated data adds meaningful variety to the dataset, ultimately improving model performance.

The improvement in robustness from adding synthetic images comes from the increased diversity and balance they introduce. In small and imbalanced datasets, models are prone to overfitting due to limited variety. Synthetic images, especially from DDPM, expand the dataset with diverse samples, reducing overfitting and helping the model generalize better. The balanced data also ensures better representation of minority classes, improving classification of rare conditions. These factors together enhance stability, accuracy, and robustness across multiple runs.

In summary, generative models like DDPM and PGGANs offer an effective solution for handling the challenges of small and imbalanced datasets by generating synthetic images that enhance model performance. These models help improve accuracy, increase data diversity, and reduce variability, leading to more robust and stable training. While Greedy K Sampling encouraged diversity, Random Sampling proved more effective in maintaining stability, making generative models a reliable option for augmenting medical image datasets.

This study highlights the efficacy of generative models like DDPM and PGGANs in addressing challenges associated with small and imbalanced medical datasets. Future research could expand on this

by integrating interdisciplinary strategies to enhance healthcare outcomes. For instance, incorporating clinician stress indicators could improve both diagnostic accuracy and well-being [38]. Furthermore, innovation through Corporate Social Responsibility (CSR) emphasizes fostering trust and improving collaborative processes, which could inform strategies for integrating generative models in healthcare [39]. Advancements in IT demonstrate the potential of applying diversification strategies to optimize system performance, while theoretical frameworks for scalable technology adoption, as outlined in [40], provide a foundation for integrating generative methods effectively. Additionally, methods like diversity indices, such as the Shannon Diversity Index, could be used to assess the impact of generative models on healthcare datasets, helping to balance specialized data generation with broader themes in medical imaging [41]. Together, these approaches could create a more comprehensive and impactful application of generative models in medical imaging.

REFERENCES

- [1] H.-P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," *Deep learning in medical image analysis: challenges and applications*, pp. 3–21, 2020.
- [2] P. Costa, A. Galdran, M. I. Meyer, M. D. Abramoff, M. Niemeijer, A. M. Mendonça, and A. Campilho, "Towards adversarial retinal image synthesis," *arXiv preprint arXiv:1701.08974*, 2017.
- [3] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: a review," *Journal of Imaging*, vol. 9, no. 4, p. 81, 2023.
- [4] M. Frid-Adar, I. Diamant, E. Klang, M. M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, 2018.
- [5] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The power of generative ai: A review of requirements, models, input-output formats, evaluation metrics, and challenges," *Future Internet*, vol. 15, no. 8, p. 260, 2023.
- [6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
- [7] J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [8] S. Jagani, X. Deng, P. C. Hong, and N. M. Nejad, "Adopting sustainability business models for value creation and delivery: an empirical investigation of manufacturing firms," *Journal of Manufacturing Technology Management*, no. ahead-of-print, 2023.
- [9] N. M. Nejad and P. Hong, "Developing a competitive advantage in ev supply chain systems: A conceptual framework for national benchmarking studies," in *2024 Midwest Decision Sciences Institute Conference Hosted by Youngstown State University*, 2024, p. 60.
- [10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [11] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Artificial intelligence and statistics*. PMLR, 2009, pp. 448–455.
- [12] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.
- [13] S. Mohamed and B. Lakshminarayanan, "Learning in implicit generative models," *arXiv preprint arXiv:1610.03483*, 2016.
- [14] S. Tang, "Lessons learned from the training of gans on artificial datasets," *IEEE Access*, vol. 8, pp. 165 044–165 055, 2020.
- [15] S. Kazemini, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "Gans for medical image analysis," *Artificial Intelligence in Medicine*, vol. 109, p. 101938, 2020.
- [16] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 580–588.
- [17] A. Kitchen and J. Seah, "Deep generative adversarial neural networks for realistic prostate lesion mri synthesis," *arXiv preprint arXiv:1708.00129*, 2017.
- [18] M. J. Chuquicuma, S. Hussein, J. Burt, and U. Bagci, "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 (ISBI 2018)*. IEEE, 2018, pp. 240–244.
- [19] C. Bermudez, A. J. Plassard, L. T. Davis, A. T. Newton, S. M. Resnick, and B. A. Landman, "Learning implicit brain mri manifolds with deep learning," in *Medical Imaging 2018: Image Processing*, vol. 10574. SPIE, 2018, pp. 408–414.
- [20] C. Baur, S. Albarqouni, and N. Navab, "Generating highly realistic images of skin lesions with gans," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5*. Springer, 2018, pp. 260–267.
- [21] A. Beers, J. H. Brown, K. Chang, J. P. Campbell, S. Ostmo, M. F. Chiang, and J. Kalpathy-Cramer, "High-resolution medical image synthesis using progressively grown generative adversarial networks," *arXiv preprint arXiv:1805.03144*, 2018.
- [22] J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [23] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.
- [24] O. Rezaeian, A. E. Bayrak, and O. Asan, "An architecture to support graduated levels of trust for cancer diagnosis with ai," in *International Conference on Human-Computer Interaction*. Springer, 2024, pp. 344–351.
- [25] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir, "Robust compressed sensing mri with deep generative priors," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 938–14 954, 2021.
- [26] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 1336–1348.
- [27] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarbuerger, C. K. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, J. N. Kather, and D. Truhn, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, 2023.
- [28] P. Mahaulpatha, T. Abeywardane, and T. George, "Ddpm based x-ray image synthesizer," *arXiv preprint arXiv:2401.01539*, 2024.
- [29] X. Liu and C.-J. Hsieh, "Rob-gan: Generator, discriminator, and adversarial attacker," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 234–11 243.
- [30] H. Alqahtani, M. Kavakli-Thorne, and G. Kumar, "Applications of generative adversarial networks (gans): An updated review," *Archives of Computational Methods in Engineering*, vol. 28, pp. 525–552, 2021.
- [31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

- [32] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," *arXiv preprint arXiv:2112.03126*, 2021.
- [33] K. Gong, K. Johnson, G. El Fakhri, Q. Li, and T. Pan, "Pet image denoising based on denoising diffusion probabilistic model," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 51, no. 2, pp. 358–368, 2024.
- [34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
- [35] B. Liu, "Progressive GAN PyTorch," <https://github.com/odegeasslbc/Progressive-GAN-pytorch>, 2018, accessed: 2024-06-12.
- [36] HuggingFace, "Unet2d model documentation," <https://huggingface.co/docs/diffusers/en/api/models/unet2d>, 2024, accessed: 2024-06-12.
- [37] M. Seitzer, "pytorch-fid: FID Score for PyTorch," <https://github.com/mseitzer/pytorch-fid>, August 2020, version 0.3.0.
- [38] R. Wang, O. Rezaeian, O. Asan, L. Zhang, and T. Liao, "Relationship between heart rate and perceived stress in intensive care unit residents: Exploratory analysis using fitbit data," *JMIR Formative Research*, vol. 8, p. e60759, 2024.
- [39] S. K. Callaway and N. Mashhadi Nejad, "Innovating toward csr: Creating value by empowering employees, customers and stockholders," in *Decision Sciences Institute (DSI) Conference Proceedings, Atlanta, GA, USA*. Available at [Link], 2023.
- [40] S. Callaway and N. Mashhadi Nejad, "How it drives diversification and performance: An empirical study of financial services firms," *Available at SSRN 5040535*, 2024.
- [41] N. Mashhadi Nejad, M. J. Alvarado-Vargas, and M. Jalali Sepehr, "Refining literature review strategies: Analyzing big data trends across journal tiers," in *Academy of Management Proceedings*, vol. 2024, no. 1. Academy of Management Valhalla, NY 10595, 2024, p. 14852.