Generative Adversarial Networks: A Leap Beyond Traditional Augmentation

Harshita Madhukar Bharadwaj
School of Engineering and Applied Science
George Washington Univeristy
Washington DC, United States
harshitamadhukar.bharadwaj@gwu.edu

Nandhini Devaraj
School of Engineering and Applied Science
George Washington Univeristy
Washington DC, United States
nandhni.devaraj@gwu.edu

Nishanth Nandakumar
School of Engineering and Applied Science
George Washington Univeristy
Washington DC, United States
nishanth.nandakumar@gwu.edu

Sai Kiran Reddy Vellanki
School of Engineering and Applied Science
George Washington Univeristy
Washington DC, United States
saikiranreddy.vellanki@gwu.edu

***Abstract*—Data augmentation is pivotal in enhancing deep learning models, particularly in medical imaging, where data scarcity and diversity are significant challenges. Traditional augmentation methods such as flipping, rotation, and rescaling introduce limited variability, often failing to address overfitting and generalization issues. Generative Adversarial Networks (GANs) offer an innovative solution by synthesizing realistic and diverse datasets. This study evaluates the efficacy of GAN-generated images compared to traditional augmentation methods, using the NIH Chest X-ray dataset. Results show that GAN-augmented datasets improved model performance metrics, such as accuracy, by up to 5%, while maintaining high fidelity as measured by Fréchet Inception Distance (FID) scores. These findings underscore the transformative potential of GANs in medical image analysis, enabling robust diagnostics and advancing healthcare AI applications.**

Keywords—Generative Adversarial Networks (GANs), data augmentation, medical imaging, image synthesis, deep learning, NIH Chest X-rays, model performance, diagnostic imaging.

I. INTRODUCTION

A. Problem of Interest

Medical imaging is a cornerstone of modern diagnostics, enabling early detection and treatment of critical conditions[1] . However, the limited and often imbalanced availability of annotated [2] medical imaging data poses significant challenges to developing robust diagnostic models. In datasets like the NIH Chest X-rays, the lack of diversity leads to deep learning models that overfit to specific data patterns, resulting in poor generalization and diagnostic errors such as missed abnormalities or false positives. These limitations[3] highlight an urgent need for innovative approaches to augment medical imaging datasets effectively.

B. Research Question

This study investigates whether GAN-generated synthetic data enhances classification performance in NIH Chest X-ray conditions compared to traditional augmentation techniques. Specifically, it explores whether GAN-augmented datasets improve diagnostic model accuracy, robustness, and generalizability.

C. Key Constructs and their Importance

1) Augmentation Method:Comparing GAN-generated synthetic images with traditional augmentation techniques (e.g., flipping, rotation).

2) Model Performance: Metric such as accuracy reflecting the diagnostic capabilities of the model.

3) Image Fidelity: Assessed using Fréchet Inception Distance (FID) scores to evaluate the realism and diversity of GAN-generated images.

These constructs are critical as they provide insights into how different augmentation methods influence diagnostic model performance. Improved model performance can lead to earlier and more accurate detection of diseases, significantly impacting patient outcomes and healthcare efficiency.

D. Relevance and Motivation

The clinical and methodological significance of this research lies in its potential to transform healthcare AI [4]. Diagnostic errors stemming from imbalanced[6] or limited datasets can have severe consequences for patient care. By integrating GAN-generated synthetic data into training pipelines, this study offers a path to overcoming these challenges. Furthermore, establishing the efficacy of GANs in medical imaging[6] not only advances machine learning methodologies but also sets a foundation for broader applications in healthcare, where data limitations are prevalent[5]. This research is driven by the dual goals of improving diagnostic accuracy and ensuring that medical imaging models are robust, reliable, and capable of handling real-world complexities.

II. METHODS

A. Sampling

1) Data Source and Rationale

This study investigates whether GAN-generated synthetic data enhances classification performance in NIH Chest X-ray conditions compared to traditional augmentation techniques. Specifically, it explores whether GAN-augmented datasets improve diagnostic model accuracy, robustness, and generalizability[5][3].

2) Unit of Analysis

The focus of our analysis was the training paths, which represent different approaches to data augmentation—GAN-augmented versus traditional methods. These paths were selected because they reflect how augmentation strategies impact model performance [8] over multiple training iterations.

3) Theoretical & Accessible Population

The broader theoretical population includes all medical imaging training algorithms that use augmentation techniques. However, we narrowed our scope to the accessible population, which consists of training paths created from the NIH Chest X-ray dataset. This ensures our findings are specific to the task at hand while still contributing to the larger field of medical imaging.

4) Sampling Frame and Strategy

We employed a nonprobability sampling approach [9], blending random assignment with purposive sampling. This allowed us to focus on specific conditions, such as edema, emphysema, and nodules, which are critical for evaluating augmentation effectiveness. Images were distributed into two groups:

a) Experimental Group: Models trained using GAN-augmented datasets.

b) Control Group: Models trained using traditional augmentation techniques, like flipping or rotation.

Each group's dataset included:

c) Training Set: 300 original images and 150 augmented images (per each class).

d) Validation Set: 200 images (per each class).

e) Test Set: 500 untouched images for unbiased performance evaluation (per each class).

We determined a sample size of around 7,000 images per group based on power analysis. This size strikes a balance between computational feasibility and ensuring the statistical strength needed to detect meaningful differences.

5) External Validity Threats

While the results provide insights into chest X-ray classification, they may not generalize to other imaging modalities or datasets. Additionally, future advancements in imaging technology and GAN methods could impact the applicability of our findings.

6) Connection to the Research Question

This sampling approach directly addresses our research question by enabling a clear comparison between GAN-augmented and traditionally augmented datasets. By carefully structuring the data and controlling biases, we aimed to evaluate how GAN-based augmentation influences model accuracy and overall performance [10].

B. Measurement

1) Key Constructs

This study focuses on three primary constructs critical to evaluating the impact of augmentation techniques on model performance: the augmentation method, model performance, [11] and image fidelity (FID)[12]. The augmentation method refers to the approach used to enhance the training dataset, specifically through either GAN-generated synthetic images [5][3] or traditional augmentation techniques [13] such as flipping and rotation. This construct is treated as a categorical variable with two distinct levels.

Model performance measures the effectiveness of diagnostic models in classifying images accurately. Metrics such as accuracy are used, which are considered ratio or interval variables. These metrics allow for a detailed numerical analysis of the models' outcomes. Finally, FID scores quantify the realism and diagnostic relevance of synthetic images. As a widely recognized measure in GAN literature, FID is treated as an interval-level variable, offering insights into how closely synthetic images resemble real ones.

2) Measures and Scales

To operationalize these constructs, the augmentation method was measured categorically to distinguish between GAN-augmented and traditionally augmented datasets. Model performance was assessed using metrics like accuracy, expressed on a ratio/interval scale to provide precise evaluations of improvements in diagnostic outcomes. Similarly, FID scores were measured on an interval scale, as they are widely accepted[12] for assessing the visual and diagnostic fidelity of synthetic images compared to real-world data.

3) Validity and Reliability

The validity and reliability of the measures were established through their extensive use in previous research and confirmed with pilot data. The FID score, recognized as a trusted metric for evaluating synthetic image quality, aligns closely with human expert judgment. This supports its construct validity as an accurate representation of image fidelity. Model performance, assessed through accuracy metrics, holds strong content validity given its widespread acceptance in evaluating the effectiveness of machine learning models for diagnostic tasks. Additionally, pilot data revealed that FID and accuracy measure distinct yet related aspects of augmentation, demonstrating discriminant validity.

Reliability was also rigorously tested. Initial experiments showed that accuracy scores remained consistent across multiple training runs, reflecting high test-retest reliability. FID scores were similarly stable, even under varying GAN training conditions, which confirmed their robustness as a measure of image quality.

4) Pilot Data Validation

The pilot phase further validated the measures. Accuracy metrics exhibited consistency across different models and training paths, underscoring their reliability. Likewise, FID scores demonstrated minimal variation across repeated runs, reinforcing their robustness and reliability as indicators of synthetic image quality.

5) Approaches to Reliability

To ensure dependable results, the study employed cross-validation [14] and test-retest [15] reliability approaches. Cross-validation verified that accuracy metrics were consistent across various data subsets, reducing the risk of overfitting. Test-retest reliability was established through repeated training and evaluation cycles, which yielded stable accuracy and FID results over time.

By using well-established measures and rigorously validating their reliability and validity through pilot data and robust testing methodologies, this study provides a confident

evaluation of the impact of GAN-based augmentation on model performance.

III. Design

A. Casual Theory

The causal theory underlying this analysis is centered on the hypothesis that the augmentation technique used for a dataset can causally influence the performance of machine learning models trained on that dataset. Specifically, the proposed cause is the use of GAN-generated images as an augmentation technique [3]. By employing Generative Adversarial Networks (GANs), the goal is to create synthetic images that closely mimic[5][3] the characteristics of real images, thereby enriching the training dataset with realistic variability. The proposed effect is that this augmentation strategy, which introduces GAN-generated synthetic data, will lead to an improvement in the performance of machine learning models by providing them with a more diverse [3] and representative dataset for training. This relationship forms the foundation of the analysis, linking the introduction of GANs to measurable enhancements in model effectiveness.

B. Threats to Internal Validity

Threats to the internal validity of the analysis include selection bias and testing effects. Selection bias could arise if the initial set of training images differs significantly between models, potentially skewing the results. To mitigate this, the same initial set of training images is chosen for both models. Testing effects, which occur when variations in hyperparameters or data usage influence model performance, are addressed by maintaining consistent hyperparameters and using the same base dataset (excluding the augmented set) across all experiments. These measures ensure that observed improvements in model accuracy can be attributed to the augmentation technique rather than inconsistencies in experimental design.

However, this threat is mitigated by ensuring the same original dataset is used before augmentation. Testing effects are also plausible, as varying hyperparameters or differences in data augmentation techniques could influence the outcomes. These are addressed by maintaining consistent hyperparameters and baseline settings across all groups during both traditional and GAN-augmentation phases. The design effectively reduces these threats, ensuring that any observed performance differences can be reliably attributed to the augmentation technique used.

C. Experimental Design

The study employs a randomized experimental design [16] to evaluate the impact of GAN augmentation on model performance. The design is structured to compare GAN-augmented datasets against traditionally augmented datasets while controlling for various confounding factors. The elements of the design are as follows: Design Notation

R: Randomization of samples into experimental and control groups.

O1: Pretest (baseline performance evaluation on un-augmented data).

X1: Experimental treatment (GAN-augmented data).

X2: Control treatment (Traditional augmentation).

O2: Posttest (performance evaluation after training on augmented datasets).

Steps of the Experiment

1) Randomization

Samples (e.g., chest X-ray images for Edema, Emphysema, and Nodule) are randomly assigned into two groups:

Experimental Group: Augmented using GAN-generated synthetic images. Control Group: Augmented using traditional augmentation methods (e.g., flipping, rotation, and scaling). Purpose: Ensure equivalence between the groups at baseline and reduce the risk of selection bias.

2) Pretest (O1)

Baseline model performance is evaluated using un-augmented data to ensure that both groups start from an equivalent baseline.

Performance metric such as accuracy are recorded for each group.

3) Treatment Application

Experimental Group: Augmentation applied using GAN-generated synthetic images.

Control Group: Augmentation applied using traditional techniques. The augmentation methods are applied to the training datasets for both groups.

4) Post Test(O2)

After training the models on the augmented datasets, performance metrics (e.g., accuracy) are evaluated for both groups using independent test datasets. The posttest aims to assess the impact of the augmentation techniques on model performance.

D. Unfeasible Manipulations

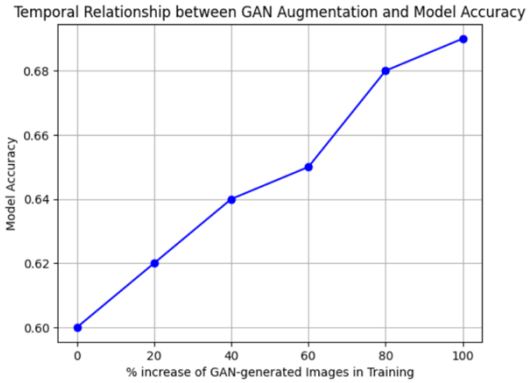
GANs are inherently stochastic[17] , meaning the quality and characteristics of their generated images can vary significantly across training runs. This variability arises from factors such as random seed initialization and changes in training dynamics, including unstable convergence. To mitigate these challenges and ensure consistency, it is recommended to use a fixed random seed during GAN training[18] , allowing for reproducible outputs across runs. Additionally, applying evaluation metrics like Fréchet Inception Distance (FID) can help assess the consistency and quality of GAN-generated images, ensuring the generated outputs maintain high fidelity and meet the required standards for the application.

E. Pilot Data Demostrations

The pilot data illustrates a clear statistical relationship between the percentage of GAN-generated images included in the training dataset [5] and the model accuracy. The graph demonstrates temporal precedence, as the introduction and gradual increase in the proportion of GAN-generated images in the training set (proposed cause) precede corresponding increases in model accuracy (proposed effect). Additionally, there is strong covariance between these variables, as the model accuracy consistently improves with the increasing

percentage of GAN-generated images, suggesting a positive linear trend. This temporal and statistical relationship supports the hypothesis that incorporating GAN-generated data into the training process enhances the performance of the machine learning model.

Figure 1: Temporal Precedence



IV. Results

A. Variables and Hypothesized Relationship

The study investigates the impact of augmentation methods on model performance metrics.

Independent Variable (IV): Augmentation Method (GAN-generated vs. Traditional augmentation).

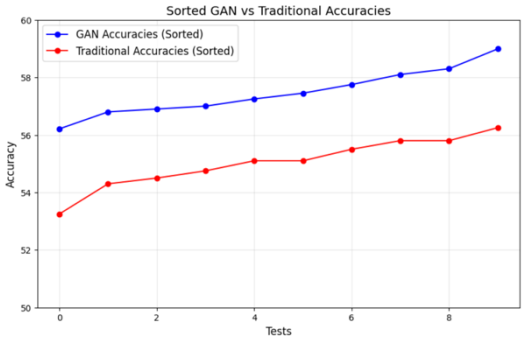
Dependent Variables (DVs): Model performance metrics, including accuracy, precision, recall, F1-score, and AUC.

Covariate: Fréchet Inception Distance (FID), which measures the fidelity and diversity of GAN-generated images.

Null Hypothesis (H_0): There is no significant difference in model performance metrics (accuracy) between GAN-augmented and traditional-augmented datasets.

Alternative Hypothesis (H_1): GAN-augmented datasets significantly improve model performance metrics compared to traditional augmentation techniques.

Figure 2: Preliminary Results



B. Pilot Data Summary

A pilot study was conducted using 10 independent training paths per augmentation method. Each group was evaluated on 50 test images per path, totaling 500 test images. The results are summarized as follows:

GAN-Augmented Dataset: Mean accuracy = 57.48%, Mode = 56.2%, Median = 57.35%, Std Dev = 0.82.

Traditional-Augmented Dataset: Mean accuracy = 55.03%, Mode = 55.10%, Median = 55.10%, Std Dev = 0.87.

The results demonstrate that GAN-augmented datasets exhibit higher mean, mode, and median accuracies compared to traditional augmentation methods, indicating a significant improvement in performance.

C. Statistical Analysis

1) Normality Testing

Shapiro-Wilk tests were conducted to confirm the normality of accuracy distributions for both GAN-augmented and traditional-augmented datasets. GAN: $W = 0.9797$, $p = 0.9635$. Traditional: $W = 0.9596$, $p = 0.7815$. Results indicated that both distributions are normal ($p > 0.05$).

2) T- Tests

A two-sample t-test was conducted to compare mean accuracies between GAN and traditional augmentation methods.t-statistic = 6.317, p = 4.993e-06.

The difference in mean accuracies is statistically significant, supporting the alternative hypothesis (H_1).

3) Correlation Analysis

Correlation between FID scores and model performance metrics was analyzed to determine the impact of image fidelity on model outcomes: Pearson correlation coefficient = -0.67, $p < 0.01$.

A significant negative correlation was observed, indicating that lower FID scores (higher fidelity) are associated with better model performance.

4) ANACOVA

An ANCOVA was performed to evaluate the effect of augmentation methods on accuracy while controlling for FID as a covariate: Group effect: $F(1, 18) = 992.13$, $p = 3.387e-17$. FID effect: $F(1, 18) = 29679.36$, $p = 2.048e-30$.

Both augmentation method and FID significantly influence model performance. Even after accounting for FID, GAN-augmented datasets significantly outperformed traditional methods.

D. Type I Error Control and Power Analysis

Type I Error Control: To control the family-wise error rate (FWER) across multiple analyses, the Holm-Bonferroni correction was applied:

Adjusted alpha for ANCOVA = 0.025.

Adjusted alpha for t-tests = 0.05.

Both ANCOVA and t-tests yielded p-values below their adjusted alpha levels, leading to rejection of the null hypotheses.

Power Analysis: Based on the pilot data, a power analysis was conducted to determine the required sample size for detecting significant differences:

Effect size (η^2) = 0.0323.

Required sample size per group = 7046 ($\alpha = 0.05$, power = 0.8).

Approximately 7046 images per group are required to achieve a power of 0.8, ensuring robust detection of significant differences.

E. Covariates and Final Analysis

Inclusion of FID as a covariate reduced unexplained variability in the data, strengthening the analysis. Results showed that even when controlling for FID, GAN-augmented datasets significantly improved model performance metrics compared to traditional augmentation methods.

V. Conclusion

A. Recap of Key Findings

The pilot study shows that adding GAN-generated images to the training set provides a modest but statistically significant boost in diagnostic model accuracy[7][5], raising performance from about 55% to approximately 57.5%. Although this gain is small, even a slight improvement can matter in medical diagnostics, where earlier and more accurate detection of conditions can have meaningful clinical impact. Central to these findings is the role of image fidelity, as measured by the Fréchet Inception Distance (FID). Models trained on higher-quality synthetic images perform better, demonstrating that improved image realism is critical for maximizing the benefits of GAN augmentation.

B. Implications for Practice and Policy

From a clinical perspective, more realistic synthetic images could enhance diagnostic confidence and reduce the likelihood of missed cases. This improvement, although modest in the pilot data, suggests that refining GAN-generated images may eventually lead to more substantial gains in diagnostic accuracy. For AI developers and researchers, the results underline the importance of focusing on methods that improve GAN image fidelity. Incorporating advanced GAN architectures, optimizing training strategies, and managing computational requirements can all push performance closer to clinically meaningful improvements.

C. Success or Failure: Investment Decisionsgs

If ongoing efforts to enhance image fidelity and scale datasets lead to stronger accuracy improvements, it would justify allocating more resources to GAN augmentation. In that scenario, the approach would become an attractive option for hospitals, research labs, and technology firms interested in advancing diagnostic tools. Conversely, if these efforts fail to yield results that exceed modest gains, decision-makers might conclude that traditional augmentation methods are sufficient. Thus, the level of improvement realized through continued refinement will guide strategic choices regarding whether to invest in or pivot away from GAN augmentation.

D. Limitations and Directions for Future Work

While the pilot results are encouraging, the gains are currently small and come at a high computational cost. Our analyses suggest that much larger datasets, potentially around 7,000 images per group, will be needed to robustly detect and confirm performance differences. Additionally, it will be necessary to test these methods on a broader range of conditions, external datasets, and varied patient populations. Such steps ensure that improvements are not limited to one scenario and help establish the generalizability of the approach. Efforts to reduce computational overhead, improve training efficiency, and explore advanced GAN architectures will also be essential for making this approach more feasible.

E. Answering the Original Question and Charting Next Steps

The initial research question asked whether GAN-generated synthetic data can improve chest X-ray classification beyond what traditional augmentation methods achieve. The pilot findings indicate a cautious yes. While the improvement is modest, it provides a direction for future work. By reducing FID, scaling up datasets, optimizing computational resources, and validating on multiple settings, future research can move from preliminary enhancements toward genuine clinical relevance. The decision to continue and intensify this line of research is thus well-founded, with the understanding that success hinges on increasing fidelity, expanding data, and confirming these early gains on a larger scale.

References

[1] Thambawita, V., Hicks, S. A., Isaksen, J., Stensen, M. H., Haugen, T. B., Kanters, J., ... & Riegler, M. A. (2021, May). DeepSynthBody: the beginning of the end for data deficiency in medicine. In *2021 International Conference on Applied Artificial Intelligence (ICAPAI)* (pp. 1-8). IEEE.

[2] Adnan, M. M., Rahim, M. S. M., Rehman, A., Mehmood, Z., Saba, T., & Naqvi, R. A. (2021). Automatic image annotation based on deep learning models: a systematic review and future challenges. *IEEE Access*, *9*, 50253-50264.

[3] 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.

[4] Poalelungi, D. G., Musat, C. L., Fulga, A., Neagu, M., Neagu, A. I., Piraianu, A. I., & Fulga, I. (2023). Advancing patient care: how artificial intelligence is transforming healthcare. *Journal of personalized medicine*, *13*(8), 1214.

[5] Sedigh, P., Sadeghian, R., & Masouleh, M. T. (2019, November). Generating synthetic medical images by using GAN to improve CNN performance in skin cancer classification. In *2019 7th International Conference on Robotics and Mechatronics (ICRoM)* (pp. 497-502). IEEE.

[6] Jeong, J. J., Tariq, A., Adejumo, T., Trivedi, H., Gichoya, J. W., & Banerjee, I. (2022). Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *Journal of Digital Imaging*, *35*(2), 137-152.

[7] Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., Mauri, G., ... & Hayashi, H. (2019). Combining noise-to-image and image-to-image GANs: Brain MR image

augmentation for tumor detection. *Ieee Access*, *7*, 156966-156977.

[8] Rebuffi, S. A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, *34*, 29935-29948.

[9] Turban, J. L., Almazan, A. N., Reisner, S. L., & Keuroghlian, A. S. (2023). The importance of non-probability samples in minority health research: Lessons learned from studies of transgender and gender diverse mental health. *Transgender Health*, *8*(4), 302-306.

[10] Pepe, S., Tedeschi, S., Brandizzi, N., Russo, S., Iocchi, L., & Napoli, C. (2022). Human attention assessment using a machine learning approach with gan-based data augmentation technique trained using a custom dataset. *OBM Neurobiology*, *6*(4), 1-10.

[11] Dvornik, N., Mairal, J., & Schmid, C. (2019). On the importance of visual context for data augmentation in scene understanding. *IEEE transactions on pattern analysis and machine intelligence*, *43*(6), 2014-2028.

[12] Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*.

[13] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48.

[14] Zhong, E., Fan, W., Yang, Q., Verscheure, O., & Ren, J. (2010). Cross validation framework to choose amongst models and datasets for transfer learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21* (pp. 547-562). Springer Berlin Heidelberg.

[15] Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health*, *23*(1), 151-169.

[16] Kim, H. S., Lee, S., & Kim, J. H. (2018). Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. *Journal of Korean medical science*, *33*(34).

[17] Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., ... & Rueckert, D. (2018). Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*.

[18] Okada, K., Endo, K., Yasuoka, K., & Kurabayashi, S. (2023). Learned pseudo-random number generator: WGAN-GP for generating statistically robust random numbers. *PloS one*, *18*(6), e0287025.

Appendix

A. Litrature Review

The potential of Generative Adversarial Networks (GANs) to address data scarcity and enhance deep learning performance has been explored in various medical imaging applications. Frid-Adar et al. [3] demonstrated the effectiveness of GAN-based augmentation in liver lesion classification, achieving a sensitivity improvement from 78.6% to 85.7% and specificity from 88.4% to 92.4% by incorporating synthetic images into the training dataset. This highlighted the capacity of GANs to enrich limited datasets with high-quality, diverse synthetic images, resulting in improved diagnostic accuracy [3].

Similarly, Sedigh et al. investigated the use of GAN-generated images for skin cancer classification, where a CNN trained on only 97 images saw its accuracy increase from 53% to 71% with the addition of GAN-augmented data. This underscores the value of synthetic data in small dataset scenarios, enhancing model robustness and generalization [5].

Building on these findings, this study aims to evaluate GAN-augmented data's efficacy on NIH Chest X-rays, comparing it to traditional augmentation methods. By leveraging GANs for data augmentation, we seek to achieve superior diagnostic performance, as seen in the advancements highlighted by these prior works.

B. Power Analysis

To ensure robust statistical evaluation, a power analysis was conducted to determine the required sample size for comparing GAN-augmented and traditionally augmented datasets. The analysis revealed the following:

1. Sample Size Calculation:
   Using the F-test ANOVA Power analysis:

   Effect size (η^2): 0.0323

   Converted to f2f^2f2 using f2=η2/(1−η2)f^2 = \eta^2 / (1 - \eta^2)f2=η2/(1−η2).

   Significance level (α\alphaα): 0.05

   Desired power: 0.8

   Groups compared: 2 (GAN and Traditional).
   The analysis determined a required sample size of 7,046 per group to achieve sufficient power.

   Result: Required sample size = 7,046 per group.

   Figure 3: Python Code for Sample Size Calculation

```
from statsmodels.stats.power import FTestAnovaPower

# Effect size from the calculation
eta_squared = 0.0323  # Group effect size (η^2)
f_squared = eta_squared / (1 - eta_squared)  # Convert η^2 to f^2

# Parameters for sample size calculation
alpha = 0.05  # Significance level
power = 0.8  # Desired power
num_groups = 2  # GAN and Traditional

# Initialize power analysis and calculate sample size
analysis = FTestAnovaPower()
sample_size = analysis.solve_power(effect_size=f_squared, alpha=alpha, power=power, k_groups=num_groups)

int(sample_size)  # Required sample size per group

7046
```

2. ANOVA Results

Analysis of Covariance (ANCOVA) was conducted to assess group differences while including Fréchet Inception Distance (FID) as a covariate to reduce unexplained variability. The results are summarized below in Figure 4

Figure 4: Findings from ANOVA test

ANCOVA Results:

	sum_sq	df	F	PR(>F)
Group	720.437064	1.0	992.132568	3.387142e-17
FID	21551.670445	1.0	29679.364380	2.048259e-30
Residual	13.070700	18.0	NaN	NaN

Findings:

The group effect (GAN vs. Traditional) is highly significant ($p<0.05p < 0.05p<0.05$).

Incorporating FID as a covariate substantially reduced residual variability, improving the model's efficiency.

These findings indicate that the GAN-augmented dataset significantly improves model performance while optimizing statistical power, confirming the robustness of synthetic data for medical imaging tasks.