CSE 291 Project Final Report: [MIGFLaGAN] Medical Image Generation through Federated Learning and General Adversarial Networks

Sehaj Dhillon
Department of Computer Science
UC San Diego La Jolla, CA 92093
sedhillon@ucsd.edu

Jeffrey Yang
Department of Computer Science
UC San Diego La Jolla, CA 92093
jey013@ucsd.edu

Daniel Zhao
Department of Computer Science
UC San Diego La Jolla, CA 92093
djzhao@ucsd.edu

Abstract

Medical image generation is known to be a very challenging task, especially while maintaining patient privacy. For instance, some methods of generative modeling require extensive amounts of data to train on. Without proper protocols and security measures in place, patient privacy could be at risk. Medical imagery remains a critical piece of infrastructure in the patient-care and medical research space. Maintaining patient privacy remains difficult due to the way state-of-the-art models are developed, trained, and evaluated. Generative Adversarial Networks (GANs) and Federated Learning are techniques that have been utilized to combat such an issue individually, however, a combination of the two remains rather unexplored. We propose a combination of GANs and Federated Learning for the medical image generation task in an effort to maintain patient privacy while also training a high-quality class-conditional medical image generator. We experiment using Chest X-Ray imagery and compare against baselines of a classic, centralized cGAN, evaluating using existing generative model evaluation techniques. https://github.com/jey013ucsd/cse291-group11-final-project

1 Introduction

In healthcare, privacy is one of the most critical aspects of the patient-care experience. Patients expect their health information to remain secure and private and local and federal laws like HIPAA mandate such expectations. Advancements in Generative AI have led to the ability to generate images for multiple domains, such as art and even medical imagery. However, in order to train and develop these models, large amounts of data are needed, posing a grave privacy concern for patients as they may be unaware how their information is processed or stored.

1.1 Problem Definition.

Generative Adversarial Networks (GANs) Goodfellow et al. [2014] have been explored for medical imagery as a means to develop photo-realistic medical cases for multiple use cases, whether training, education, or prediction. This can help reduce the reliance on patient data and alleviate some of the privacy concerns related to medical imagery.

Preprint. Under review.

However, challenges still remain with GAN medical imagery. These models need to be trained somehow so there is still concern for patient privacy. In addition, some find that while GANs can generate quire realistic medical imagery, they are still unable to reproduce the "full richness of medical datasets" Skandarani et al. [2023]. Further research is required for advancing medical image generation that retains patient privacy while also improving the fidelity of medical imagery. We aim to explore this research problem.

1.2 Problem Significance.

Generating medical imagery is rooted in generative AI in it of itself. Image generation is a relatively growing field where much of the frontier is either unexplored or currently being explored. Designing methods to generate images while protecting privacy has remained challenging. By pursuing this direction, we can learn and explore the different methods of deep learning which allow for image generation while also promoting privacy practices for model development. This directly supports society as a whole as it can drive medical practice positively while putting patients at ease regarding the use of their very personal medical data. Our project aims to utilize a mixture of a conditional GAN and Federated Learning to generate medical images while keeping patient data private. We consider a method where clients (ex. hospitals) do not need to directly share patient data for model development but rather contribute to a shared global model without exposing sensitive data used for training.

1.3 Technical Challenge.

Related work has found vulnerabilities in federated GANs where back-door attacks may be possible. Jin and Li [2023]. As such, maintaining patient privacy remains a difficult aspect of this research while remaining a crucial area to explore. Gathering data for this project is also difficult as patient data is either scarce or difficult to access within a short time-frame as that of this course. However, we have identified datasets that can be utilized for our use case. Implementing this research is also difficult as it will require advanced compute to generate such imagery, however the availability of resources from the course should be sufficient for our task. Related works have identified evaluation criteria for GAN-based medical image generation Skandarani et al. [2023] which we intend to build upon to evaluate our method as well.

1.4 State-of-the-Art.

As mentioned earlier, some limitations of state-of-the-art (SOTA) papers are the inability to reproduce the same quality of medical imagery as those contained in medical datasets Skandarani et al. [2023]. In addition, regular GAN or Federated Learning papers, while relevant to medical imaging and the concerns regarding privacy and security, remain sparse. There is not much research in the direction of combining Federated Learning and GANs. Some research even highlights areas like One-shot Federated learning for generative models like GANs remains unexplored and open Gargary and Cristofaro [2024]. They highlight how privacy and integrity for non-GAN-based Federated learning remains an area that requires exploring. As a result, SOTA papers are scarce in this direction and those that exist continue to attempt to bridge the gap of image generation and privacy.

1.5 Contributions

- Introduce an implementations of a federated conditional GAN for medical imaging.
- Demonstrate high-quality class-conditional image generation feasibility w/o sharing patient data.
- Provide an empirical comparison between centralized & federated training for generative modeling.
- Analyze training stability challenges unique to federated GANs and propose mitigation strategies.
- Create a simulation environment for multi-institution medical GAN training using real medical image partitions.
- Establish quantitative and qualitative evaluation pipelines for federated medical image generation.

2 Related Work

2.1 Conditional GANs in Medical Imaging

Generative Adversarial Networks (GANs) have emerged as a versatile tool in medical image analysis, especially in settings where labelled data are limited. Recent reviews have highlighted multiple applications of GANs in medical and molecular imaging, including augmentation, modality conversion, denoising, super-resolution, and image generation conditioned on disease severity or radiogenomics attributes. [Yi et al., 2019, Koshino et al., 2021] In particular, ”conditional” variants of GANs (cGANs) (where the generator is conditioned on a label or attribute) enable more controlled image synthesis (e.g., generate an X-ray with vs. without disease). For example, a survey of medical-image synthesis found that cGAN and DCGAN were among the most frequently used architectures across modalities such as MRI, CT, and X-ray. [Yi et al., 2019] Finally, transfer learning or fine-tuning of pretrained models (e.g., GAN generators or discriminators) is increasingly explored in medical imaging contexts to compensate for small datasets or domain shifts. [Han et al., 2019] Most conditional GAN work in medical imaging is done in centralized settings where labels are aggregated. This is seen in Yi et al. [2019] where they highlight using public datasets. This leaves open how well conditional generation works when labels are distributed across institutions where they cannot be shared due to privacy concerns.

2.2 Federated Learning for Medical Imaging

Federated Learning (FL) is a distributed training paradigm in which multiple clients (e.g., hospitals) train local models using their private data and share only model updates (not raw data) with a central aggregator. This allows training across institutions without data sharing, addressing privacy and governance concerns. [McMahan et al., 2023] In the healthcare domain, there has been a growing body of work applying FL to imaging tasks: a survey reported that neural networks and medical imaging are among the most common machine-learning and data-types in healthcare FL studies. Specific to medical image analysis, one survey categorised FL methods by client end, server end, and communication aspects, and discussed real-world implementation challenges such as heterogeneous (non-IID) data distributions and deployment barriers. [Guan et al., 2024] Thus, FL is particularly relevant in medical imaging, where privacy, data-sharing restrictions, and cross-institution variability are major concerns. Further, most Federated Learning applications in medical imaging focuses on distriminative tasks like classification and segmentation rather than generation, leaving image generation to be under-explored Guan et al. [2024].

2.3 Bridging GANs and Federated Learning

Recently, research has begun to explore the intersection of GANs and FL: synthesising medical images in a federated manner can in principle afford privacy-preserving augmentation across sites. For example, FedMed-GAN proposed federated domain translation for unsupervised brain image synthesis. [Wang et al., 2023] However, the literature remains comparatively sparse; one paper investigated back-door attacks in federated GANs for medical image synthesis, highlighting additional vulnerabilities when combining the two paradigms. [Jin and Li, 2023]

While GAN-based image generation is well explored, and FL for medical imaging is increasingly studied, the synthesis of conditional GANs in a federated, multi-site setting remains an emerging area. Most existing work focuses on unconditional generation or domain translation, leaving label-conditioned synthesis for targeted medical data augmentation largely unexplored in federated contexts. Thus, while both GAN-based image synthesis and Federated Learning are well explored independently, together their intersection for conditional medical image generation is still under-explored.

2.4 Alternative Methods

Other generative models have also been explored in medical imaging, such as variational autoencoders (VAEs), diffusion models, and auto-regressive transformers. However, these models can require additional computational cost and overhead even if generation may result in slightly better results. We therefore focus on conditional GANs as a tractable and controlled approach to explore federated conditional synthesis. Methods like diffusion models can require many iterative steps that can be computationally expensive, whereas GANs, during inference, require a forward pass and can be less computationally expensive. As a result, while training may be simpler than GANs, inference can be more complex with other models.

2.5 Our Method

In this work, we implement and evaluate a federated conditional GAN for medical image synthesis, comparing its performance against a centralized baseline. Specifically, we: (1) simulate multiple client hospitals with local data partitions, (2) apply conditional GAN training under the federated learning paradigm using FedAvg aggregation, and (3) assess whether federated training can approach centralized performance using standard image quality metrics. **We acknowledge that GAN training instability may be amplified in federated settings, and prepare fallback strategies including fine-tuning pretrained models or maintaining local discriminators if full federated convergence fails.** Our goal is to provide empirical evidence on the feasibility of privacy-preserving conditional image synthesis for medical data augmentation. We aim to introduce an improved training method that preserves privacy by reducing transmission of medical data while maintaining performance on par with centralized data performance through this combination GAN-Federated Learning method.

3 Methodology

3.1 Problem Setting.

The central problem addressed in this work is how to train a high-quality, class-conditioned medical image generator without requiring any institution to share its private data.

Specifically, the objective of this project is to evaluate how a conditional Generative Adversarial Network (cGAN) can be trained effectively in a federated learning setting, where multiple clients collaboratively contribute to the training of a shared global model without exposing sensitive training data 1. We aim to assess whether such a federated setup can achieve comparable generative quality and realism to a conventionally trained, centralized model.

3.2 Idea Summary.

To address this problem, we propose the following federated learning setup for a conditional generative adversarial network to allow for decentralized training on private medical image data. At a high level, our approach extends traditional, state-of-the-art cGAN frameworks by introducing federated averaging across distributed generator–discriminator pairs while maintaining class-conditioning and privacy constraints. Each client maintains local copies of the generator and discriminator, trains on its own data partition, and then participates in a global aggregation step that combines model parameters across institutions without sharing raw data.

Past federated learning work focuses more on segmentation and classification tasks; our application of federated learning for image generation tasks is relatively less explored as a research area. The biggest risks to our approach are the training difficulties inherent to GAN frameworks. GANs are known to suffer from instability, convergence failure, and mode collapse, and these challenges are expected to be amplified in a decentralized training context where parameter updates across clients may diverge. If the full federated set up fails to converge, alternative solutions include fine-tuning a pretrained model instead of training from scratch, or only aggregating the generator while keeping discriminators local.

Additionally, if we simply split the dataset into equal parts to simulate multiple clients, it would make the assumption that data is independently and identically distributed across clients. This is almost never guaranteed in real settings. As such, if the model demonstrates stable convergence under the IID assumption, subsequent experiments will deliberately relax this condition by introducing controlled data augmentations (e.g. applying varying levels of noise or color shifts) to simulate heterogeneous client distributions and evaluate the robustness of the federated training process.

3.3 Description.

Formally, a cGAN consists of two networks, a generator G and a discriminator D. Their training objective is formulated as a two-player minimax game between the two networks:

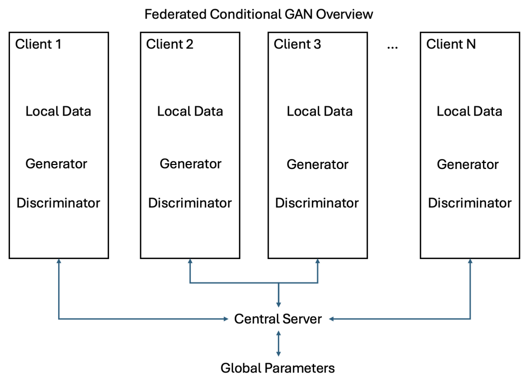


Figure 1: Methodology Description

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

The discriminator aims to maximize the objective by correctly distinguishing real samples x from samples generated by the generator $G(z \mid y)$. Concurrently, the generator aims to minimize the objective by producing samples $G(z \mid y)$ that the discriminator misclassifies as real.

In a federated learning setup, we introduce N clients $\{C_1, C_2, \ldots, C_N\}$. Each client C_i holds a private dataset $D_i = \{(x_i, y_i)\}_{j=1}^{m_i}$ of size m_i. Each client locally trains a copy of the generator-discriminator pair, parametrized by θ_G^i and θ_D^i. After a set number of local updates K, each client transmits their parameters to a central server, which performs federated aggregation by computing an average of all local parameters weighted by their individual client's dataset size:

$$\theta_G = \sum_{i=1}^{N} \frac{m_i}{M} \theta_G^i, \quad \theta_D = \sum_{i=1}^{N} \frac{m_i}{M} \theta_D^i,$$

Our generative model is designed to learn the conditional distribution

$$p_\theta(x \mid y),$$

a high-dimensional distribution over 128×128 chest X-ray images. The generator defines a transformation

$$G(z, y; \theta_G) : \mathbb{R}^{d_z} \times \{0, 1\} \to \mathbb{R}^{128 \times 128 \times 3},$$

which maps a noise vector $z \sim p_z(z) = \mathcal{N}(0, I)$ and class embedding y to an image.

The discriminator estimates two quantities: an adversarial score $D_{\text{adv}}(x)$ used to approximate the divergence between p_{data} and p_G, and a class-prediction head $D_{\text{cls}}(x)$ that encourages the generator to produce class-consistent samples. The generator objective is therefore a combination of an adversarial term and a conditional consistency term. Both networks are optimized using hinge losses, which provide stable gradients in high-resolution image synthesis.

3.4 Implementation.

We implemented our model in **PyTorch** and ran all experiments on an NVIDIA A40 GPU through RunPod. The implementation was *not* built on a specific existing public codebase; we wrote the federated GAN pipeline from scratch, using only the standard FedAvg algorithm and PyTorch GAN training patterns as conceptual references. For architectures, we used a **conditional GAN** for its ability to steer generation towards specific attributes, with class embeddings in the generator and a projection discriminator. Hyper-parameters were tuned empirically via small grid/hand search, with final choices of batch size 32, learning rate 2×10^{-4} for generator and 8×10^{-5} for discriminator, z-dim 128, 10 rounds, 3 local epochs, and $\alpha = 0.5$.

4 Experiments

4.1 Datasets and Tools.

We use the Chest X-Ray Images (Pneumonia) dataset, containing 5,863 labeled chest X-ray images across two classes: NORMAL and PNEUMONIA. All images are resized to 128×128 grayscale and then partitioned into multiple federated clients according to the chosen Dirichlet α configuration.

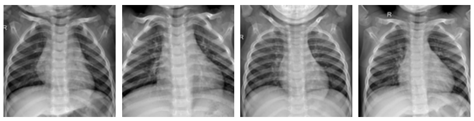


Figure 2: Example NORMAL images from the training dataset.

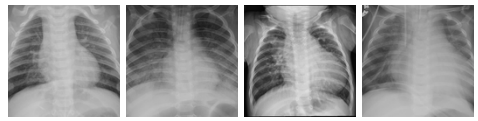


Figure 3: Example PNEUMONIA images from the training dataset.

All experiments are implemented in PyTorch, and FedML is used to simulate the federated learning environment.

4.2 Baselines.

We compare our federated cGAN against several centralized generative models trained on the full dataset:

- **ACGAN Baseline** [Odena et al., 2016]: A class-conditional GAN whose discriminator jointly performs real/fake discrimination and class prediction. This serves as our strongest conditional baseline.
- **Unconditional GAN**: A non-conditional generator and discriminator trained to model the marginal distribution $p(x)$ without conditioning on class labels. This provides an upper bound on achievable KID scores, since the model solves an easier task than conditional GANs.
- **Variational Autoencoder (VAE)**: A reconstruction-based generative model with a Gaussian latent prior. While VAEs typically produce blurrier samples than GANs, they provide a likelihood-based baseline and help contextualize adversarial model performance.

All baselines are trained for 125 epochs at 128×128 resolution using batch size 64, hinge losses for GAN models, and Adam optimizers with learning rates 2×10^{-4} (generator) and 8×10^{-5} (discriminator). For conditional GANs, we employ class-balanced sampling and maintain an exponential moving average of generator weights for improved stability. Including both unconditional and conditional models enables us to disentangle two effects: (1) the inherent difficulty of learning class-conditional image distributions (where conditional GANs typically underperform unconditional ones due to reduced per-class data), and (2) the additional degradation caused by federated training and non-IID data.

4.3 Evaluation Metrics.

We use the following evaluation metrics:

- **Kernel Inception Distance (KID):** A polynomial-kernel MMD measure between real and generated image features. Unlike FID, it is unbiased even when computed on small real datasets, making it suitable for medical imaging settings with limited test samples. All KID reports are evaluated on 500 samples between generated images and training images.
- **Training Stability Indicators:** Discriminator and generator loss trends are tracked to compare convergence behavior across centralized and federated setups.

We also generate samples from each model to allow for a qualitative visual analysis of model performance.

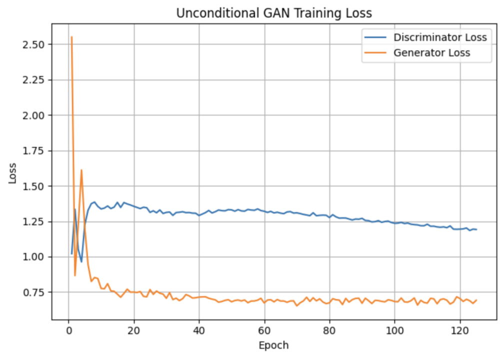


Figure 4: Training loss curves for the baseline **unconditional GAN**.

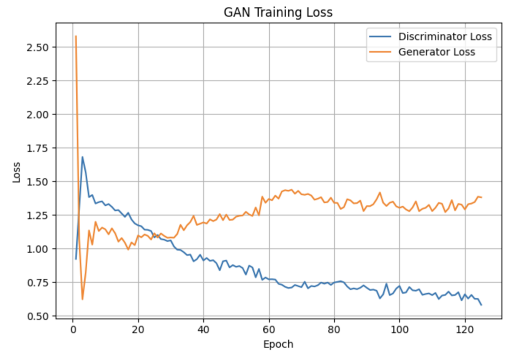


Figure 5: Training loss curves for the baseline **centralized cGAN**.

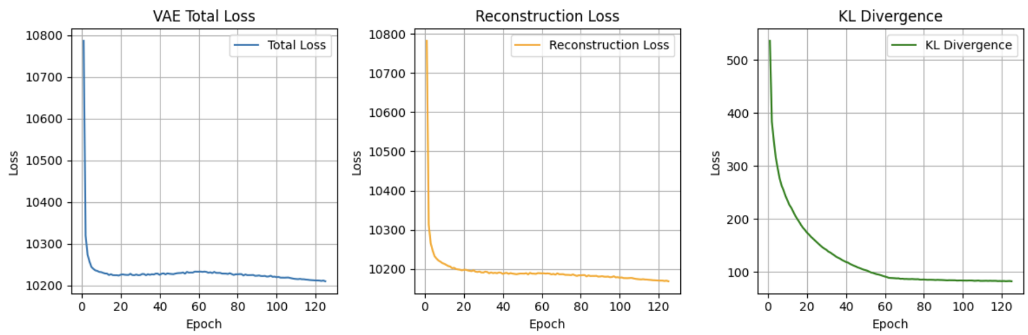


Figure 6: Training loss curves for the **VAE baseline**.

4.4 Quantitative Results.

We evaluate all models using KID (lower is better), computed separately for NORMAL and PNEUMONIA classes and averaged into a single KID-AVG score. Table 1 summarizes the baseline scores for the unconditional GAN, VAE, and centralized cGAN (no federated learning). As expected, the unconditional GAN provides the strongest baseline (KID-AVG $= 0.0687$), while the centralized cGAN achieves moderate image quality (KID-AVG $= 0.2887$).

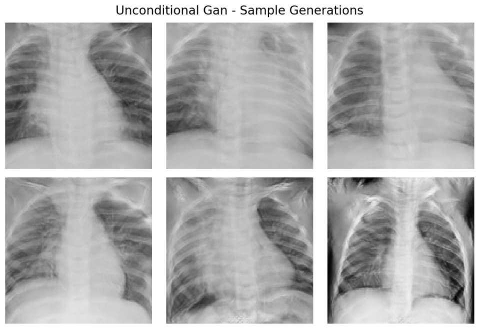
Table 1: KID scores for baseline models (lower is better).

Model	KID-NORMAL	KID-PNEUM	KID-AVG
Unconditional GAN	**0.0827**	**0.0546**	**0.0687**
VAE	0.5622	0.5672	0.5647
Centralized cGAN	0.2322	0.3452	0.2887

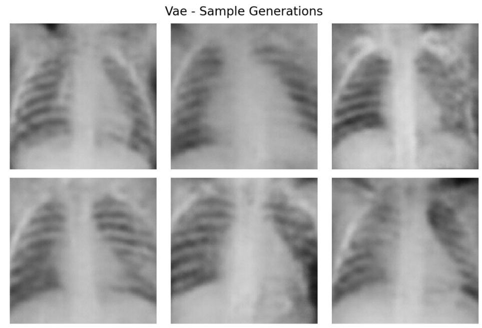
To analyze how federated learning conditions affect image quality, we organize results into three categories: (1) non-IID severity controlled by Dirichlet α, (2) number of clients, (3) communication frequency (rounds vs. local epochs). A smaller α corresponds to *more skewed* client label distributions and thus more challenging federated optimization.

4.5 Ablative Studies.

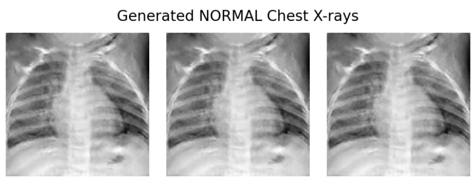
Discussion of α Ablation. Surprisingly, the best-performing federated model occurs at **highly non-IID conditions** ($\alpha = 0.1$), where KID-AVG drops to **0.1656**, outperforming the centralized cGAN by a large margin. Moderate non-IID also improves over some baseline settings, with a general trend downwards as α decreases (besides $\alpha = 0.4$, which we believe is an outlier with bad random initialization). We hypothesize that this improvement arises because non-IID partitions implicitly



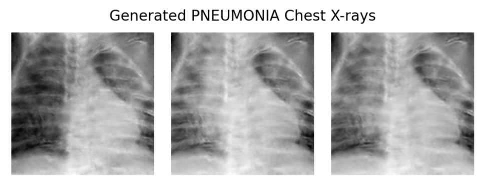
(a) Unconditional GAN



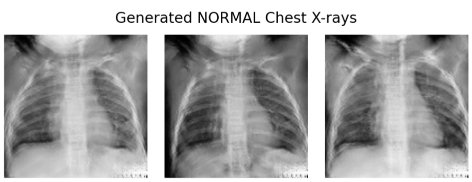
(b) VAE



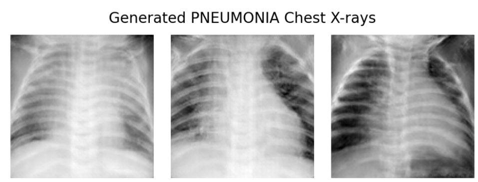
(c) cGAN (NORMAL)



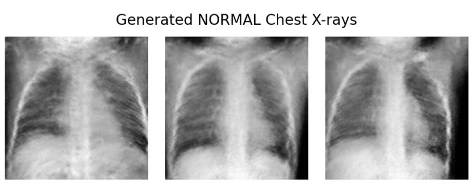
(d) cGAN (PNEUMONIA)



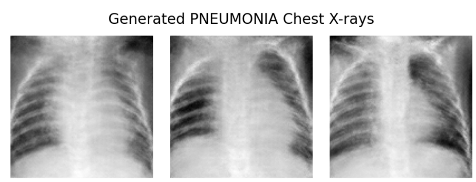
(e) Fed α=0.1 (NORMAL)



(f) Fed α=0.1 (PNEUMONIA)



(g) Fed IID (NORMAL)



(h) Fed IID (PNEUMONIA)

Figure 7: Qualitative comparison of centralized baselines (Rows 1–2) and federated models (Rows 3–4). Each row displays NORMAL and PNEUMONIA samples side-by-side for visual comparison.

regularize GAN training: each client's discriminator specializes on a different subset of the data distribution, forcing the global generator to satisfy multiple "expert" discriminators and thereby reducing mode collapse. In contrast, IID training produces nearly identical gradients across clients, which homogenizes updates and provides no additional diversity beyond centralized training.

For comparison, we also contrast the best non-IID setting against the IID case.

Client Count Ablation. We next vary the number of clients under a fixed non-IID configuration ($\alpha = 0.5$), seen in table 3. Client-level heterogeneity increases with more participants, and performance declines for both very few clients (underfitting, poor global diversity) and very many clients (overly fragmented updates). The best KID is achieved with the standard 5-client setup.

Communication Frequency Ablation. Finally, we study the trade-off between more communication rounds and more local training, keeping total local epochs approximately fixed, seen in table 4. Increasing the number of communication rounds improves global model synchronization and stabilizes generator updates. As shown, **more frequent communication (60 rounds)** outperforms

Table 2: Federated cGAN performance under varying non-IID severity (Dirichlet α). Lower KID indicates better image quality.

Experiment (5 clients, 25 rounds, 5 local epochs)	KID-NORMAL	KID-PNEUM	KID-AVG
$\alpha = 1.0$ (mild non-IID)	0.5922	0.1495	0.3709
$\alpha = 0.9$	0.6101	0.2401	0.4251
$\alpha = 0.8$	0.6077	0.2205	0.4141
$\alpha = 0.7$	0.5808	0.1974	0.3891
$\alpha = 0.6$	0.5071	0.2434	0.3753
$\alpha = 0.5$	0.5748	0.2258	0.4003
$\alpha = 0.4$	0.8280	0.2108	0.5194
$\alpha = 0.3$	0.5820	0.1559	0.3690
$\alpha = 0.2$	0.2992	0.1407	0.2199
$\alpha = 0.1$ **(highly non-IID)**	**0.2227**	**0.1085**	**0.1656**

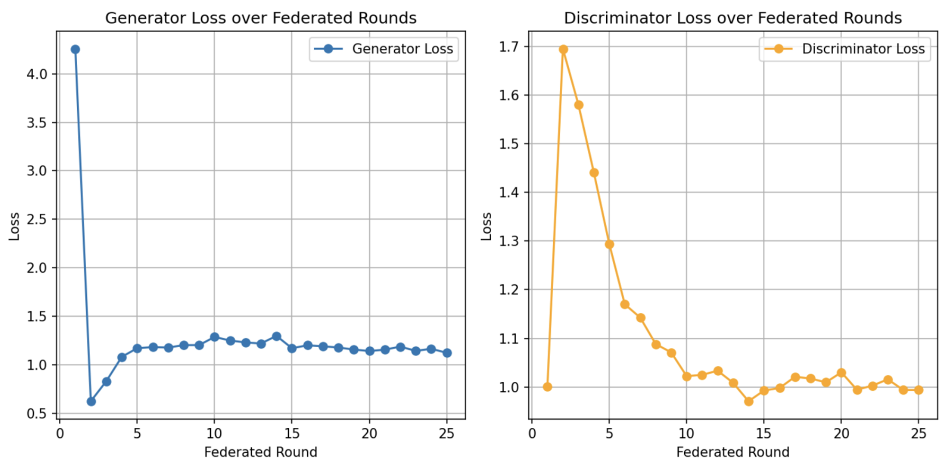


Figure 8: Federated cGAN training loss for $\alpha = 0.1$ (best-performing model).

heavier local training with fewer rounds (20 rounds), despite using the same total number of local epochs.

Qualitative Results. To complement the quantitative KID evaluation, we perform a qualitative assessment of generated samples for both the NORMAL and PNEUMONIA classes. Across both centralized and federated models, NORMAL samples generally show clearer and more uniform lung regions: the shading inside the lungs is relatively even, and there are no large bright patches. Overall, these images look smooth and consistent, with minimal distortion. PNEUMONIA samples, in contrast, tend to show more uneven or textured regions inside the lung fields; these appear more "cloudy" or patchy compared to the NORMAL class. Federated models with strong performance (especially $\alpha = 0.1$) achieve qualitatively similar structure, lung boundaries, and interior shading to the centralized cGAN, suggesting that high-quality generative models can be learned even under realistic, privacy-preserving data fragmentation.

5 Conclusion and Discussion

Achievements. Throughout this project, we successfully implemented, trained, and evaluated both centralized and federated conditional GANs for chest X-ray image generation. We established strong baselines using an unconditional GAN, a VAE, and a centralized cGAN, achieving KID-AVG scores of 0.0687, 0.5647, and 0.2887 respectively. Our key contribution is a comprehensive experimental study of federated adversarial learning under varying non-IID conditions. Surprisingly, we find that **high heterogeneity dramatically improves performance**: the best federated model ($\alpha = 0.1$) achieves a KID-AVG of **0.1656**, outperforming the centralized cGAN by a large margin. We further

Table 3: Effect of number of clients on KID performance (non-IID $\alpha = 0.5$).

Experiment (25 rounds except where noted)	KID-NORMAL	KID-PNEUM	KID-AVG
e6: 3 clients (21 rounds $\times$ 6 epochs)	0.6852	0.4702	0.5777
e4: 5 clients (25 rounds $\times$ 5 epochs)	0.5748	0.2258	0.4003
e7: 10 clients (30 rounds $\times$ 4 epochs)	0.6227	0.3600	0.4914

Table 4: Effect of communication patterns (rounds vs. local epochs).

Experiment (5 clients, $\alpha = 0.5$)	KID-NORMAL	KID-PNEUM	KID-AVG
e8: 60 rounds $\times$ 2 epochs (120 epochs total)	0.5126	0.1929	0.3527
e9: 20 rounds $\times$ 6 epochs (120 epochs total)	0.6311	0.2053	0.4182

provided detailed qualitative comparisons showing that highly non-IID training yields sharper and more realistic samples than IID or mildly non-IID setups. Overall, our work demonstrates that federated GANs can remain stable and even excel under extreme data fragmentation, which is generally an encouraging finding for privacy-preserving medical image synthesis.

Lessons Learned. This project highlighted the unique challenges of adversarial training in a federated environment. While supervised FL often degrades under non-IID partitions, GANs behave very differently: discriminator specialization across clients can act as implicit regularization, reducing mode collapse and improving sample diversity. Through experimentation, we learned the importance of careful hyperparameter tuning (synchronization frequency, local epochs, and client count) and the value of strong centralized baselines to contextualize federated results. We also learned that qualitative evaluation is essential for medical imaging tasks, since numerical metrics alone cannot capture subtle structural differences in lung fields. Given more time, we would explore architectural improvements like multi-discriminator aggregation or diffusion-based federated models to further enhance stability and fidelity.

Bottlenecks and Mitigation. The primary bottleneck was training instability, which compounded under federated updates where client drifts could push the generator into collapse. We mitigated this by (1) reducing local update steps, (2) increasing communication rounds to keep models synchronized, and (3) using EMA on generator weights even in the federated setting. Another bottleneck was evaluating medical image quality; due to the lack of a lot of data, it was hard to use traditional metrics such as FID to evaluate (due to limited sample size). We resolved this by combining KID (which works better with smaller sample size) with structured qualitative analysis of NORMAL and PNEUMONIA visual characteristics.

Team Responsibilities. The team began as three Computer Science Master's students with backgrounds in machine learning through coursework and research experience. Unfortunately Jeffrey had to drop out of the course midway through. Responsibilities were divided broadly across three areas: model development, experimental setup, and analysis. Team members collaborated on designing the model architecture, preprocessing the dataset, implementing the training pipelines, and evaluating performance. Jeffrey contributed to the project idea formation and implemented the baseline cGAN that was later used for the experiments. Sehaj contributed to the project idea formation, model development, and contributed to documentation, report writing, and final presentation preparation, coordinated the deliverables. Also facilitated implementation of proposal and milestone feedback. Daniel conducted Federated Learning experiments and reports on the experiments. Also wrote evaluation code, ablation studies, and implemented the VAE and unconditional GAN baselines.

6 LLM Usage

In this paper, LLMs were used to word sentences, reformat grammar, spelling, and punctuation. In addition, LLMs were used to perform look ups for general concepts and were reaffirmed with identifying sources which confirmed the information presented by the LLM. Any sources utilized in such a manner are included in the references of this paper.

References

Ashkan Vedadi Gargary and Emiliano De Cristofaro. A systematic review of federated generative models, 2024. URL https://arxiv.org/abs/2405.16682.

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

Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey, 2024. URL https://arxiv.org/abs/2306.05980.

Changhee Han, Kohei Murao, Tomoyuki Noguchi, Yusuke Kawata, Fumiya Uchiyama, Leonardo Rundo, Hideki Nakayama, and Shin'ichi Satoh. Learning more with less: Conditional pggan-based data augmentation for brain metastases detection using highly-rough annotation on mr images. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 119–127. ACM, November 2019. doi: 10.1145/3357384.3357890. URL http://dx.doi.org/10.1145/3357384.3357890.

Ruinan Jin and Xiaoxiao Li. Backdoor attack and defense in federated generative adversarial network-based medical image synthesis, 2023. URL https://arxiv.org/abs/2210.10886.

Kazuhiro Koshino, Rudolf A. Werner, Martin G. Pomper, Ralph A. Bundschuh, Fujio Toriumi, Takahiro Higuchi, and Steven P. Rowe. Narrative review of generative adversarial networks in medical and molecular imaging. *Annals of Translational Medicine*, 9(9):821, 2021. doi: 10.21037/atm-20-6325. URL https://atm.amegroups.org/article/view/62414.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023. URL https://arxiv.org/abs/1602.05629.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. URL https://arxiv.org/abs/1610.09585. Revised July 2017.

Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 9(3), 2023. ISSN 2313-433X. doi: 10.3390/jimaging9030069. URL https://www.mdpi.com/2313-433X/9/3/69.

Jinbao Wang, Guoyang Xie, Yawen Huang, Jiayi Lyu, Yefeng Zheng, Feng Zheng, and Yaochu Jin. Fedmed-gan: Federated domain translation on unsupervised cross-modality brain image synthesis, 2023. URL https://arxiv.org/abs/2201.08953.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, December 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.101552. URL http://dx.doi.org/10.1016/j.media.2019.101552.