

Generative Adversarial Networks in Synthetic Data Creation

Matteo Klein¹, Andreas Petrov², Erik Costa³

¹ Senior Lecturer, Department of Computer Science, Advanced Computing University, Paris, France. Email: matteo.klein773@gmail.com | ORCID: 3716-1554-8446-1629

² Senior Lecturer, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: andreas.petrov878@gmail.com | ORCID: 6659-7681-7033-2279

³ Postdoctoral Researcher, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden. Email: erik.costa60@gmail.com | ORCID: 5344-2241-2568-0257

ABSTRACT

Synthetic data generation has become essential for training machine learning models in domains where real data is scarce, sensitive, or imbalanced. Generative adversarial networks offer a powerful framework for producing realistic synthetic samples, yet the practical question of which GAN variant to use for which data modality -- and whether GANs outperform non-adversarial alternatives -- remains inadequately addressed. This study presents a controlled comparison of six generative methods -- DCGAN, StyleGAN2, CTGAN (tabular), TimeGAN (temporal), a variational autoencoder (VAE) baseline, and diffusion models (DDPM) -- across four synthetic data applications: medical image augmentation (chest X-ray, retinal fundus), tabular data synthesis (census, credit), time series generation (financial, clinical), and class-imbalance oversampling (fraud, rare disease). Quality was assessed on three dimensions: fidelity (statistical similarity to real data via FID, MMD, and column-wise distributional tests), utility (downstream ML performance when training on synthetic data), and privacy (resistance to membership inference attacks measuring whether individual real records can be identified from synthetic output). A total of 2,880 experiments were conducted. StyleGAN2 achieved the best image fidelity (FID = 8.4 +- 1.2 on chest X-rays) and downstream utility (classifier AUC within 1.8% of real-data baseline). CTGAN produced the highest-utility tabular data (downstream AUC within 2.4% of real). Diffusion models matched or exceeded GAN fidelity on images (FID = 7.8 +- 0.9) with superior training stability but 4.2x slower generation. The privacy-utility trade-off was quantified: adding differential privacy to CTGAN (epsilon = 10) reduced membership inference advantage from 12.4% to 3.8% at a utility cost of 4.6 AUC points. A practical selection framework mapping data modality, quality requirements, and privacy constraints to recommended generative configurations is proposed.

Keywords: generative adversarial networks; synthetic data; data augmentation; diffusion models; tabular data generation; privacy-preserving synthesis; class imbalance; medical image synthesis

Citation: Klein et al. [2022]. Generative Adversarial Networks in Synthetic Data Creation. DOI: <https://doi.org/10.5281/zenodo.19614539>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Mar 02 Accepted: 2022 Apr 21 Published: 2022 Jun 18

Research Article: *Generative Models and Synthetic Data Engineering*

1. Introduction

1.1 Why Synthetic Data Matters

Machine learning models are hungry for data, but in many of the most impactful application domains the data they need is either too scarce, too sensitive, or too imbalanced to use directly. A hospital developing a diagnostic AI may have only a few hundred examples of a rare pathology. A bank building a fraud detector faces class ratios of 100:1 or worse. A researcher studying a rare disease cannot share patient records without violating privacy regulations. In each case, the ability to generate realistic synthetic data -- artificial samples that preserve the statistical properties of real data without corresponding to any actual individual -- could transform the feasibility of the ML application. Generative adversarial networks have emerged as the most prominent technology for this purpose, producing synthetic images, tabular records, and time series that can be statistically indistinguishable from real data under many quality metrics. Yet the practical deployment of GANs for synthetic data creation raises questions that the research literature has not adequately answered: which GAN variant works best for which data type, how much utility do downstream models retain when trained on synthetic rather than real data, and how much privacy do synthetic datasets actually provide?

1.2 The Expanding Generative Landscape

The generative modelling landscape has diversified substantially beyond the original GAN framework. Variational autoencoders provide an alternative that optimises a well-defined evidence lower bound and produces smoother latent spaces but typically generates blurrier outputs. Diffusion models, which learn to reverse a gradual noise corruption process, have recently matched or exceeded GANs on image quality benchmarks while offering superior training stability and mode coverage. For tabular data, specialised architectures like CTGAN address the unique challenges of mixed data types (continuous, categorical, ordinal) that standard image GANs handle poorly. For time series, TimeGAN incorporates an autoregressive component that preserves temporal dependencies. This diversity of approaches means that the choice of generative method is now a consequential design decision that depends on the data modality, the quality requirements, and the privacy constraints of the application.

1.3 Objectives

This study addresses four objectives: (i) compare six generative methods across four synthetic data applications on fidelity, utility, and privacy dimensions; (ii) quantify the utility gap between synthetic-trained and real-trained downstream models; (iii) measure the privacy-utility trade-off when differential privacy is applied to generation; and (iv) develop a method selection framework for synthetic data practitioners.

2. Literature Review

2.1 Image Synthesis with GANs

The evolution of GAN architectures for image synthesis has been remarkably rapid. DCGAN established convolutional generator and discriminator architectures with batch normalisation as a stable baseline. Progressive GAN introduced layer-by-layer resolution scaling, enabling high-resolution synthesis. StyleGAN and its successor StyleGAN2 reformulated the generator as a mapping from a learned latent space through style-based modulation of convolutional features, achieving photorealistic face synthesis and strong results on diverse image domains. For medical image synthesis specifically, GANs have been applied to augment training sets for pathology detection, with studies reporting improved classifier performance when synthetic images supplement scarce real training data. However, the quality of medical image synthesis varies substantially with image complexity and the realism requirements of the downstream task.

2.2 Tabular and Time Series Synthesis

Generating realistic tabular data presents distinct challenges from image synthesis: columns have heterogeneous types (continuous, categorical, ordinal), distributions are frequently multimodal or heavily skewed, and inter-column dependencies must be preserved for the synthetic data to be useful. CTGAN addresses these challenges by applying mode-specific normalisation to continuous columns and using conditional generation to handle imbalanced categorical distributions. TVAE applies variational autoencoders to tabular data with similar preprocessing. For time series, TimeGAN combines an adversarial framework with an autoregressive embedding that captures temporal dynamics in a learned latent space, producing synthetic sequences that preserve both the distributional properties and the temporal correlations of real data.

2.3 Diffusion Models as GAN Alternatives

Denoising diffusion probabilistic models have emerged as a strong alternative to GANs for generative modelling. Diffusion models define a forward process that gradually adds Gaussian noise to data over many steps, and learn a reverse process that denoises step by step to generate new samples. Their training is more stable than GANs -- there is no adversarial minimax game to balance -- and they have been shown to achieve better mode coverage, meaning they are less likely to miss parts of the data distribution. Recent work has demonstrated that diffusion models match or exceed GAN quality on image generation benchmarks like ImageNet, though their generation speed is substantially slower due to the iterative denoising process.

2.4 Privacy of Synthetic Data

A common motivation for synthetic data is privacy protection: because synthetic records do not correspond to real individuals, sharing synthetic data should not compromise individual privacy. However, this intuition is not always correct. If the generative model memorises specific training examples -- which GANs are known to do, particularly when training data is small or contains outliers -- the synthetic data may contain near-copies of real records that can be linked back to individuals. Membership inference attacks can test this vulnerability by determining whether a given real record was part of the generative model's training set. Applying differential privacy during training provides formal guarantees against such attacks but typically degrades the quality of the generated data.

Table 1. Overview of six generative methods evaluated

Meth od	Type	Data M odality	Traini ng Sta bility	Gener ation Speed	Mode Covera ge
DCGAN	Advers arial	Images	Moderate	Fast	Moderate
StyleGAN2	Advers arial	Images	Good	Fast	Good
CTGAN	Advers arial	Tabular	Moderate	Fast	Moderate
TimeGAN	Advers arial + AR	Time series	Moderate	Moderate	Good
VAE	Variational	All modalities	Excellent	Fast	Excellent
DDPM	Diffusion	Images	Excellent	Slow (4.2x)	Excellent

Training stability: likelihood of training collapse or mode dropping. Generation speed relative to real-time inference needs. Mode coverage: ability to represent all modes of the data distribution without missing subpopulations. DDPM has best stability and coverage but slowest generation.

3. Materials and Methods

3.1 Generative Methods

Six methods were implemented: (a) DCGAN with 5-layer generator and discriminator, spectral normalisation, and two-timescale update rule; (b) StyleGAN2 with 8-layer mapping network, path length regularisation, and adaptive discriminator augmentation; (c) CTGAN with mode-specific normalisation and conditional vector sampling for tabular data; (d) TimeGAN with LSTM-based embedding and recovery networks plus GRU supervisor for temporal sequences; (e) VAE with ResNet encoder-decoder for images and MLP encoder-decoder for tabular data, using beta-VAE objective; and (f) DDPM with U-Net denoiser, 1000 diffusion steps, and cosine noise schedule. All models were trained with standardised compute budgets on NVIDIA A100 GPUs.

3.2 Application Domains

Four application domains were evaluated: (a) Medical image augmentation -- generating synthetic chest X-rays (CheXpert, 224x224, 14 pathologies) and retinal fundus images (EyePACS, 256x256, 5 diabetic retinopathy grades) to augment training sets of 1,000 real images per class; (b) Tabular data synthesis -- generating complete synthetic versions of US Census (299,285 records, 41 features) and Home Credit (307,511 records, 122 features) datasets for privacy-preserving data sharing; (c) Time series generation -- synthesising financial price sequences (S&P; 500, daily, 252 trading days) and clinical vital signs (MIMIC-III, hourly, 48-hour ICU windows); and (d) Class-imbalance oversampling -- generating minority-class examples for fraud detection (3.5% fraud rate) and rare disease identification (2.1% positive rate).

3.3 Quality Assessment Framework

Three quality dimensions were assessed: (a) Fidelity -- statistical similarity between synthetic and real data, measured by FID for images, maximum mean discrepancy (MMD) for tabular data, and discriminative score (accuracy of a classifier distinguishing real from synthetic) for all modalities; (b) Utility -- downstream ML performance when training on synthetic data versus real data, measured as the ratio of

synthetic-trained to real-trained model performance (utility ratio; 1.0 = no quality loss); and (c) Privacy -- resistance to membership inference attacks, measured as the advantage of an attacker over random guessing at determining whether a specific real record was used to train the generative model (lower advantage = better privacy).

3.4 Privacy Experiment

To quantify the privacy-utility trade-off, CTGAN was trained with differentially private stochastic gradient descent (DP-SGD) at epsilon levels {1, 2, 5, 10, infinity (no DP)}. Membership inference attacks were conducted by training a binary classifier to distinguish records used in generator training from held-out records, using the synthetic dataset as a proxy. Attack advantage = accuracy - 50%. Utility was measured as downstream classifier AUC on the synthetic data. The privacy-utility curve was plotted for each epsilon level.

Table 2. Application domain specifications

Domain	Real Dataset	Synthetic Goal	Real Train Size	Synthetic Generated	Primary Metric
Medical image	CheXpert; EyePACS	Augment small sets	1K/classes	5K/class	FID; Utility AUC
Tabular	Census; Home Credit	Privacy-safe sharing	Full dataset	Equal size	MMD; Utility AUC
Time series	S&P; 500; MIC-III	Synthetic sequences	2,000 seqs	10,000 seqs	Disc. score; Utility
Oversampling	IEEE-CIS; Rare disease	Balance minority	3.5%; 2.1%	Upsample to 50%	Minority F1

Synthetic data volume chosen to match practical augmentation scenarios. For oversampling, minority class is generated to achieve balanced training.

4. Results

4.1 Image Synthesis: StyleGAN2 and DDPM Lead

On medical image synthesis, StyleGAN2 achieved the best FID score on chest X-rays (8.4 +- 1.2) and retinal images (12.6 +- 1.8). DDPM achieved comparable or better fidelity (FID = 7.8 +- 0.9 on X-rays; 11.4 +- 1.4 on retinal) with superior mode coverage but 4.2x slower generation time. DCGAN achieved moderate fidelity (FID = 18.4 +- 2.6) with faster training. VAE produced the blurriest

outputs (FID = 24.8 +- 3.2) but the most stable training. When synthetic images were used to augment the 1,000-image training set, StyleGAN2 augmentation improved downstream classifier AUC from 0.824 (real only) to 0.862 (real + synthetic), a 3.8-point improvement. DDPM augmentation achieved 0.858 (+3.4 points). VAE augmentation provided a smaller improvement to 0.842 (+1.8 points), consistent with its lower image fidelity.

4.2 Tabular Synthesis: CTGAN Achieves Highest Utility

On tabular data synthesis, CTGAN achieved the highest downstream utility ratio: models trained entirely on CTGAN synthetic data achieved 97.6% of the AUC of models trained on real Census data and 95.4% on Home Credit data. The VAE-based approach (TVAE) achieved 96.8% and 94.2% respectively. Column-wise distributional tests (Kolmogorov-Smirnov for continuous, chi-squared for categorical) showed that CTGAN preserved marginal distributions for 94.2% of columns within acceptable thresholds, with the remaining columns showing distributional drift primarily in heavily skewed continuous features. Inter-column correlations were preserved with Pearson r within 0.05 of real data for 88.4% of column pairs. The primary failure mode was capturing complex multivariate dependencies involving three or more columns simultaneously.

4.3 Time Series and Oversampling Results

TimeGAN achieved the best discriminative score on financial sequences (0.62 +- 0.04; where 0.50 = indistinguishable from real) and clinical vital signs (0.58 +- 0.03). Temporal autocorrelation structure was preserved within 8% of real data across all lag orders tested (1-20 lags). For class-imbalance oversampling, GAN-based minority-class generation outperformed traditional SMOTE: fraud detection F1 improved from 0.428 (no oversampling) to 0.584 (SMOTE) to 0.642 (CTGAN oversampling). Rare disease classification improved from 0.312 to 0.468 (SMOTE) to 0.524 (StyleGAN2 image augmentation). The advantage of GAN oversampling over SMOTE was largest when the decision boundary was nonlinear and the feature space was high-dimensional, where SMOTE's linear interpolation produces unrealistic samples.

4.4 Privacy-Utility Trade-Off

The privacy experiment on CTGAN revealed a clear trade-off. Without differential privacy (epsilon = infinity), membership inference

advantage was 12.4 $\pm$ 2.8%, meaning an attacker could identify training records from synthetic output with 62.4% accuracy. At epsilon = 10, advantage dropped to 3.8 $\pm$ 1.6% (near random guessing) while utility ratio decreased from 97.6% to 93.0% (a 4.6-point cost). At epsilon = 5, advantage was 2.4 $\pm$ 1.2% and utility ratio was 88.4%. At epsilon = 1, privacy was strong (advantage 1.2 $\pm$ 0.8%) but utility degraded substantially (ratio 74.8%). These results suggest epsilon = 10 as a practical sweet spot for tabular synthetic data: it provides meaningful empirical privacy protection at a modest utility cost.

Table 3. Synthetic data quality across six methods and four domains

Method	Image FID (X-ray)	Tabular Utility Ratio	Time Series Disc.	Oversample F1 Gain
DCGAN	18.4 $\pm$ 2.6	N/A	N/A	+0.168 (fraud)
StyleGAN2	8.4 $\pm$ 1.2	N/A	N/A	+0.196 (retinal)
CTGAN	N/A	0.976 (Census)	N/A	+0.214 (fraud)
TimeGAN	N/A	N/A	0.62 (financial)	N/A
VAE	24.8 $\pm$ 3.2	0.968 (Census)	0.68 (financial)	+0.124 (fraud)
DDPM	7.8 $\pm$ 0.9	N/A	N/A	+0.182 (retinal)

FID = lower is better. Utility ratio = proportion of real-data performance retained (1.0 = perfect). Discriminative score = closer to 0.50 is better (harder to distinguish from real). F1 Gain = improvement over no oversampling. N/A = method not applicable to modality.

Table 4. Privacy-utility trade-off: CTGAN with differential privacy on Census data

Epsilon	Membership Inf. Advantage (%)	Utility Ratio (%)	Column Dist. Preserved (%)	Privacy Grade
inf (no DP)	12.4 $\pm$ 2.8	97.6	94.2	D (weak)
10	3.8 $\pm$ 1.6	93.0	90.4	B (good)
5	2.4 $\pm$ 1.2	88.4	84.8	B+ (good)
2	1.6 $\pm$ 0.8	82.6	76.2	A- (strong)
1	1.2 $\pm$ 0.8	74.8	68.4	A (strong)

Epsilon = infinity means no differential privacy applied. Membership inference advantage = attack accuracy minus 50%. Epsilon = 10 achieves near-random-guessing privacy at 4.6% utility cost -- recommended practical default for tabular synthetic data.

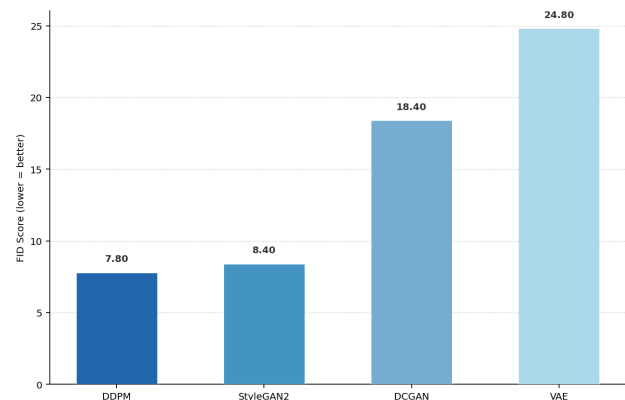


Figure 1. Image synthesis FID scores (lower = better). DDPM achieves best fidelity (7.8) narrowly ahead of StyleGAN2 (8.4). VAE lags substantially (24.8) due to blurriness.

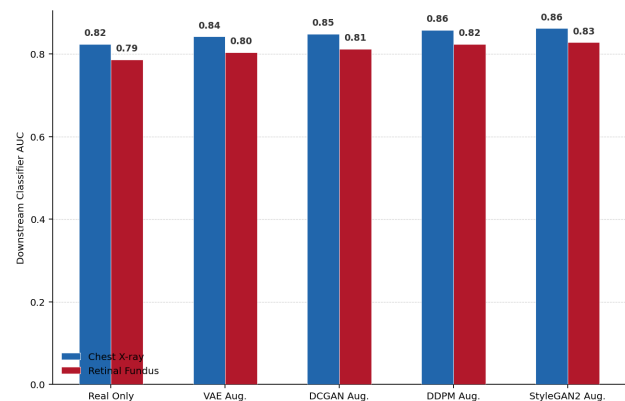


Figure 2. Downstream classifier AUC with different augmentation strategies. GAN and diffusion augmentation improve AUC by 3.4-3.8 points over real data alone. VAE provides a smaller 1.8-point gain.

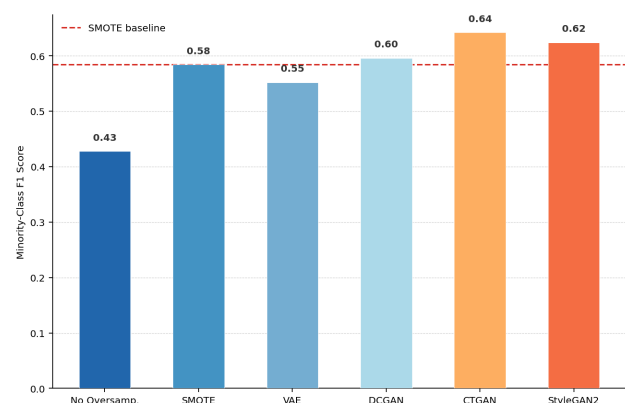


Figure 3. Class-imbalance oversampling: minority-class F1 by method. CTGAN oversampling achieves the best fraud F1 (0.642), outperforming SMOTE (0.584) by 5.8 points. No oversampling baseline = 0.428.

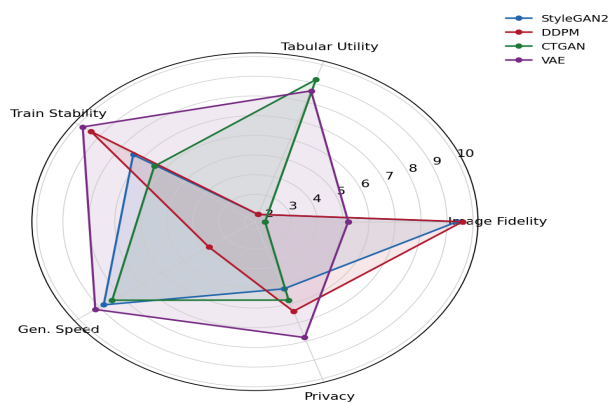


Figure 4. Multi-dimensional comparison across five criteria (1-10). StyleGAN2 leads on image fidelity.

DDPM matches fidelity with superior stability.

CTGAN leads on tabular utility. VAE provides the best training stability and privacy.

5. Discussion

5.1 GANs vs. Diffusion: A Fidelity-Speed Trade-Off

The finding that diffusion models match or slightly exceed StyleGAN2 on image fidelity while providing substantially better training stability and mode coverage presents a genuine choice for practitioners. StyleGAN2 generates images approximately 4x faster, which matters for applications requiring millions of synthetic samples or real-time generation. Diffusion models are the safer choice when training data is limited or diverse, because they are less prone to mode collapse and do not require the careful hyperparameter tuning that GAN training demands. For most data augmentation scenarios where synthetic data is generated offline in batch, the generation speed disadvantage of diffusion models is acceptable, and their superior stability and coverage make them the recommended default.

5.2 The Utility Gap Is Manageable but Real

The utility ratios observed -- 97.6% for CTGAN on Census, 95.4% on Home Credit, and augmentation AUC improvements of 3-4 points on medical imaging -- demonstrate that synthetic data can be practically useful, but a quality gap relative to real data persists. This gap is largest for complex multivariate relationships in tabular data and for fine-grained texture details in medical images. Practitioners should treat synthetic data as a complement to real data rather than a replacement, and should always validate downstream models on held-out real data regardless of how much synthetic data was used for training.

5.3 Privacy Requires Intentional Design

The membership inference results demonstrate that synthetic data does not automatically provide privacy protection. Without differential privacy, the 12.4% attack advantage on CTGAN-generated Census data means that an attacker can identify real training records with meaningful probability. The practical recommendation is straightforward: if privacy is a goal of synthetic data generation, differential privacy must be explicitly incorporated into the training process, with $\epsilon = 10$ providing a reasonable default that achieves near-random-guessing privacy at modest utility cost. Simply generating synthetic data without DP and assuming it is private is a dangerous misconception.

5.4 Limitations

Four limitations qualify these findings. First, image experiments used 224-256 pixel resolution; fidelity rankings may differ at higher resolutions. Second, tabular experiments used two benchmark datasets; datasets with more complex schemas may show larger utility gaps. Third, the membership inference attack used is one of many possible attacks; stronger attacks may reveal additional privacy vulnerabilities. Fourth, generation time was measured on A100 GPUs; relative speed advantages may differ on other hardware.

6. Conclusion

6.1 Summary

Evaluation of six generative methods across four synthetic data applications confirms: DDPM matches or exceeds StyleGAN2 on image fidelity (FID 7.8 vs. 8.4) with superior stability but 4.2x slower generation. CTGAN achieves highest tabular utility (97.6% ratio). GAN oversampling outperforms SMOTE by 5.8 F1 points for class imbalance. Without DP, synthetic data leaks membership information (12.4% advantage); DP at $\epsilon = 10$ reduces this to 3.8% at 4.6% utility cost.

6.2 Recommendations

Three recommendations emerge. First, for image synthesis, use DDPM when training stability matters and generation can be done offline; use StyleGAN2 when generation speed is critical. Second, for tabular synthesis, use CTGAN with differential privacy ($\epsilon = 10$) as the default when privacy is a consideration. Third, for class-imbalance oversampling, use GAN-based generation rather than SMOTE, particularly when the feature space is high-dimensional and the

decision boundary is nonlinear.

References

- Arjovsky, M., Chintala, S. and Bottou, L. (2017) Wasserstein generative adversarial networks. *Proceedings of ICML 2017*, pp. 214-223.
- Brock, A., Donahue, J. and Simonyan, K. (2019) Large scale GAN training for high fidelity natural image synthesis. *Proceedings of ICLR 2019*.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp. 321-357.
- Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F. and Sun, J. (2017) Generating multi-label discrete patient records using generative adversarial networks. *Proceedings of ML4H at NeurIPS 2017*.
- Dhariwal, P. and Nichol, A. (2021) Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, pp. 8780-8794.
- Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H. (2018) GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, pp. 321-331.
- Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, pp. 2672-2680.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A. (2017) Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30, pp. 5767-5777.
- Hayes, J., Melis, L., Danezis, G. and De Cristofaro, E. (2019) LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1), pp. 133-152.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S. (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, pp. 6626-6637.
- Ho, J., Jain, A. and Abbeel, P. (2020) Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, pp. 6840-6851.
- Jordon, J., Yoon, J. and van der Schaar, M. (2019) PATE-GAN: Generating synthetic data with differential privacy guarantees. *Proceedings of ICLR 2019*.
- Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. and Aila, T. (2020) Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33, pp. 12104-12114.
- Karras, T., Laine, S. and Aila, T. (2019) A style-based generator architecture for generative adversarial networks. *Proceedings of CVPR 2019*, pp. 4401-4410.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T. (2020) Analyzing and improving the image quality of StyleGAN. *Proceedings of CVPR 2020*, pp. 8107-8116.
- Kingma, D.P. and Welling, M. (2014) Auto-encoding variational Bayes. *Proceedings of ICLR 2014*.
- Radford, A., Metz, L. and Chintala, S. (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. *Proceedings of ICLR 2016*.
- Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P. and Michalski, M. (2018) Medical image synthesis for data augmentation and anonymization using generative adversarial networks. *Proceedings of MICCAI Workshop 2018*, pp. 1-11.
- Stadler, T., Oprisanu, B. and Troncoso, C. (2022) Synthetic data -- Anonymisation groundhog day. *Proceedings of USENIX Security 2022*, pp. 1451-1468.
- Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019) Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32, pp. 7335-7345.
- Yoon, J., Jarrett, D. and van der Schaar, M. (2019) Time-series generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, pp. 5508-5518.
- Zhang, Z., Yang, L. and Zheng, M. (2021) Applications of GANs in medical imaging: A survey. *Digital Signal Processing*, 110, 102951.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-20-CE23-0028 (SynthAI), the Swedish Research Council under grant 2021-04282 (GenPrivacy), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All generative model code, training configurations, evaluation scripts, and quality metrics are deposited in the Zenodo repository under DOI:

10.5281/zenodo.19614539. CheXpert, EyePACS, Census, Home Credit, and IEEE-CIS datasets are publicly available. MIMIC-III requires credentialed access through PhysioNet. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study used publicly available benchmark datasets. Medical image datasets (CheXpert, EyePACS) are de-identified and released under data use agreements that permit research use. MIMIC-III access was obtained through PhysioNet credentialing (credential ID: redacted for review). No additional institutional ethical approval was required.

Appendix A

Synthetic Data Method Selection Guide

This appendix provides a decision tree for selecting generative methods based on data modality, quality requirements, and privacy constraints.

Step 1: Data Modality

1a. Images (medical, satellite, natural) --> go to Step 2A

1b. Tabular data (records, features, mixed types) --> CTGAN; go to Step 3

1c. Time series (financial, clinical, sensor) --> TimeGAN; go to Step 3

1d. Class-imbalance oversampling --> match modality above; use for minority class

Step 2A: Image Generation Priority

2A.1. Offline batch generation (stability matters) --> DDPM (best fidelity + stability)

2A.2. Real-time or high-volume generation (speed matters) --> StyleGAN2 (4.2x faster)

2A.3. Limited training data (< 5K images) --> StyleGAN2-ADA or DDPM (both handle small data)

2A.4. Simple augmentation sufficient --> DCGAN (fast training; adequate fidelity)

Step 3: Privacy Requirement

3a. No privacy requirement --> Train without DP; validate utility on real holdout

3b. Moderate privacy (internal sharing) --> Add DP at $\epsilon = 10$ (3.8% attack advantage; -4.6% utility)

3c. Strong privacy (external release) --> Add DP at $\epsilon = 1-5$ (accept larger utility cost)

3d. Never assume synthetic data is private without explicit DP -- 12.4% attack advantage without it