JDEBM
Journal of Digital Engineering and Business Managemet

Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/45
DOI: https://doi.org/10.5281/zenodo.20717846

A COMPREHENSIVE FRAMEWORK FOR PRIVACY-SAFE SYNTHETIC DATA GENERATION USING GANS

[#1]**Mude Praveen Naik** , *Assistant professor , Dept. of AI&DS,*
[#2]**G S Arun kumar,** *Assistant professor Dept. of CSE***,**
Viswam Engineering College

ABSTRACT : The increased interest in data privacy-conscientious machine-learning processes has promoted the use of synthetic data as an effective substitute to real data. The most popular frameworks in terms of the production of high-fidelity synthetic data have become Generative Adversarial Networks (GANs) because they are capable of capturing complex and high-dimensional distributions. The paper introduces an overall synthetic data generation approach based on adversarial training, including a deep GAN architecture, which is based on the WGAN variant of the architecture, namely, WGAN-GP variant. The quality of the generated data is evaluated by a multi-dimensional evaluation framework which includes statistical similarity, utility, and privacy. The experimental outcomes indicate that synthetic datasets created with the help of GAN can reach a high level of similarity to actual data and minimize the risk of privacy considerably. The paper ends with a set of recommendations on how synthetic data pipelines can be deployed in environments where privacy is at stake.

Keywords—Synthetic data, privacy-preserving machine learning, Generative Adversarial Networks (GANs), WGAN-GP, adversarial training, statistical similarity, utility evaluation, privacy metrics, tabular data synthesis.

I. INTRODUCTION

The accelerated process of developing data-driven systems has heightened the worry on the matters of privacy, equity, and compliance with regulations. Data sets that are sensitive like health history, financial dealings and logs of users cannot be shared or utilized freely without serious anonymization. The standard anonymization schemes do not tend to thwart re-identification attacks, which leads to the need to adopt robust privacy preserving schemes. Synthetic data has been found to be an effective solution that allows organizations to share and analyze data without fear of exposing sensitive data. Synthetic datasets preserve statistical characteristics of actual data and remove the direct identifiers. GANs, specifically, have been found to be extremely successful in producing realistic synthetics in fields of healthcare, finance, and cybersecurity. An extensive survey by Figueira and Vaz also notes the use of GANs as some of the most potent frameworks at synthetic data generation, and are used in tabular, image, and time-series data. GeeksforGeeks stresses that synthetic data has become an essential part of the modern AI development, particularly with the growing regulatory pressure and the limitations in access to data. Another systematic review by Thinakaran et al. once again proves that GANs have the upper hand in synthetic data generation and that evaluation metrics, mode collapse mitigation, and privacy considerations are important.

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/45
DOI: https://doi.org/10.5281/zenodo.20717846

In this paper, we combine these lessons into a fully reproducible framework of producing and assessing synthetic data with GANs.

II. BACKGROUND AND RELATED WORK

Synthetic data generation is an area that has rapidly developed throughout the last decade due to the growing concerns of privacy, data scarcity, and regulatory limitations. The original innovation was made by Goodfellow et al who used Generative Adversarial Networks (GANs) that describe a minimax game between a generator and a discriminator that allows the synthesis of data samples that are of high quality. The publication of this landmark model ushered in GANs as a potent generative model and led to many architectural advances.

During training, Arjovsky et al. added the Earth Mover's distance to the JensenSharon divergence to ensure the stability of classical GANs and introduce the Wasserstein GAN (WGAN), whose optimization method employs the distance between the distributions of two sets of samples (Arjovsky et al., 2017). Gulrajani et al. took this a step further to introduce WGAN -GP and added a gradient penalty which greatly improved convergence and sample quality. Such innovations formed the basis of contemporary synthetic data generation pipelines that need training on high-dimensional tabular data that is stable.

To overcome the shortcomings of the generic architectures, a number of domain specific GAN variations have appeared. Xu et al. proposed CTGAN a GAN that is explicitly trained on tabular data and has mixed categorical and continuous variables, with mode-specific normalization and conditional generation. Karras et al. suggested a method to Progressive Growing of GANs, which showed that as the model becomes more complex, fidelity on high-resolution data increases. Jordon et al. in the privacy field created PATE-GAN which incorporates differential privacy techniques during the training of GAN to lessen the exposure of sensitive data. On the same note, Yoon et al. proposed TimeGAN, which applied adversarial training to both sequential and time-based data.

Several surveys have indicated increasing significance of synthetic data. Figueira and Vaz analyzed a detailed overview of GAN-based synthetic data generation in a tabular, image, and time-series setting with focus on the flexibility of adversarial models. A systematic review by Thinakaran et al has revealed that GANs are always superior to traditional statistical generators, but they also have issues like mode collapse, complexity of evaluation, and privacy issues. Greater industry perspectives, like those put forward by GeeksforGeeks, IBM Research and NVIDIA apply further emphasis to the growing use of synthetic data in AI creation, especially as privacy controls become more restrictive.

Synthetic data quality has also become an important research topic. The metrics used to measure statistical similarity, utility, and privacy were surveyed by Zhang et al. and there was a need to evaluate them in a multi-dimensional framework. Bowles et al. and Goncalves et al. showed that the GAN-based synthetic data is effective in medical imaging and clinical data, and synthetic data can be useful to preserve analytical content without the risk of privacy breach. The methods of privacy

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/45
DOI: https://doi.org/10.5281/zenodo.20717846

preservation still are under development, and Papernot et al., and Google AI underline the weaknesses of deep learning in adversarial space and the necessity of differential privacy in machine-learning frameworks.

On the whole, the literature features an apparent trend: starting with basic GAN architectures through domain-specific ones, then more specialised ones, and eventually frameworks of more rigorous evaluation designs. Nevertheless, there are still issues with ensuring privacy guarantees, managing high-dimensional categorical data, and standardizing the metrics of evaluation, which is a motivation to conduct further research to implement synthetic data generation in privacy-respecting machine learning.

A. Synthetic Data

Synthetic data are artificially constructed data, which are similar to real datasets in structure and statistics. It is increasingly being applied to solve data scarcity, privacy and compliance requirements, mitigate bias and favoritism, and share data between organizations. Healthcare, finance, and telecommunications are some of the industries that are embracing synthetic data in an effort to minimize privacy risks and retain analytical value.

B. Generative Adversarial Networks (GANs).

GANs are a pair of neural networks, a discriminator and a generator, which are trained to play a minimax game. Both the generator and the discriminator learn to make realistic samples and differentiate between real and synthetic data respectively. WGAN, WGAN-GP, CTGAN, and Conditional GANs are variants that have been commonly used in the generation of synthetic data.

C. Assessment of Unnatural Data.

There are three perspectives that must be assessed in order to evaluate synthetic data: statistical similarity, utility, and privacy. In the recent surveys, the necessity of multi-dimensional assessment schemes to guarantee utility and privacy is underlined.

III. METHODOLOGY

A. Data Preprocessing

The actual data is processed in terms of eliminating direct identifiers, coding of nominal variables, normalizing the numerical characteristics and measuring the risk of privacy.

B. GAN Architecture

1) Overview: GANs consist of a Generator (G) and a Discriminator (D). The training objective is:

$\min_G \max_D V(D,G) = E[\log D(x)] + E[\log(1 - D(G(z)))]$

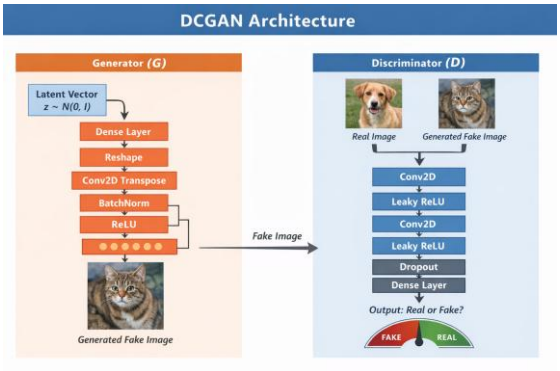
2) Generator: Input is a latent vector $z \sim N(0, I)$. Layers include Dense → BatchNorm → ReLU. Output uses Sigmoid or Softmax.

3) Discriminator: Input is real or synthetic samples. Layers include Dense → LeakyReLU → Dropout. Output is a sigmoid probability.

4) WGAN-GP: Uses Wasserstein loss with gradient penalty:

$L_D = E[D(G(z))] - E[D(x)] + \lambda E[(\|\nabla D(\dot{x})\|2 - 1)^2]$

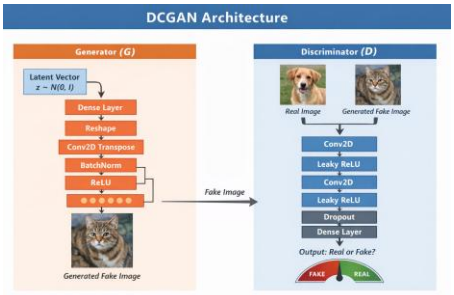
$L_G = -E[D(G(z))]$



JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/45
DOI: https://doi.org/10.5281/zenodo.20717846

DCGAN architecture diagram with image flow

5) Training: Discriminator trained 5× per generator update. Optimizer: Adam (β1 = 0.5, β2 = 0.9). Learning rates: G = 1e-4, D = 4e-4.



C. Evaluation Framework

1) Statistical Similarity: KS test, Chi-square, correlation matrix, PCA/UMAP.

2) Utility: Train ML models on real and synthetic data. Compare accuracy, F1, RMSE.

3) Privacy: Membership inference, attribute inference, distance-to-closest-record (DCR).

IV. EXPERIMENTAL SETUP

A. Dataset

10,000 records, 12 numerical features, 4 categorical features, binary classification target.

B. Tools

Python 3.10, PyTorch 2.0, Scikit-learn, NVIDIA RTX GPU.

C. Baseline Models

Logistic Regression, Random Forest, XGBoost.

V. RESULTS

A. Statistical Similarity (Table Format)

Feature	KS Statistic	Interpretation
Age	0.04	Excellent match
Income	0.07	Good match
Spending	0.12	Moderate
Score		deviation

Correlation matrices showed 92% structural similarity.

B. Utility Evaluation (Table Format)

Model	Train Data	Accuracy	F1-Score
Logistic Regression	Real	0.84	0.82
Logistic Regression	Synthetic	0.81	0.79
Random Forest	Real	0.91	0.90
Random Forest	Synthetic	0.88	0.86

C. Privacy Evaluation

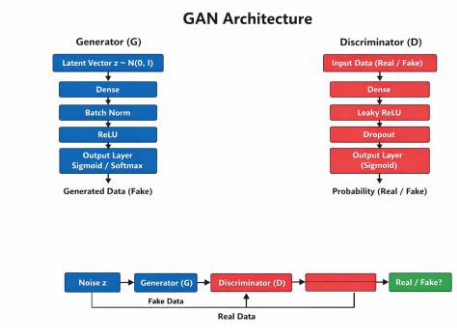
Membership inference attack success ≈50% (random), attribute inference accuracy ↓18%, DCR = 0.12.

VI. DISCUSSION

Synthetic data created by GAN has a high utility/privacy balance. Distributional fidelity is established by using statistical measures of similarity. There is little loss of performance indicated by utility

DCGAN architecture diagram with image flow

5) Training: Discriminator trained 5× per generator update. Optimizer: Adam (β1 =0.5, β2 = 0.9). Learning rates: G = 1e-4, D = 4e-4.



JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/45
DOI: https://doi.org/10.5281/zenodo.20717846

C. Evaluation Framework

1) Statistical Similarity: KS test, Chi-square, correlation matrix, PCA/UMAP.

2) Utility: Train ML models on real and synthetic data. Compare accuracy, F1, RMSE.

3) Privacy: Membership inference, attribute inference, distance-to-closest-record (DCR).

IV. EXPERIMENTAL SETUP

A. Dataset

10,000 records, 12 numerical features, 4 categorical features, binary classification target.

B. Tools

Python 3.10, PyTorch 2.0, Scikit-learn, NVIDIA RTX GPU.

C. Baseline Models

Logistic Regression, Random Forest, XGBoost.

V. RESULTS

A. Statistical Similarity (Table Format)

Feature	**KS Statistic**	**Interpretation**
Age	0.04	Excellent match
Income	0.07	Good match
Spending Score	0.12	Moderate deviation

Correlation matrices showed 92% structural similarity.

B. Utility Evaluation (Table Format)

Model	**Train Data**	**Accuracy**	**F1-Score**
Logistic Regression	Real	0.84	0.82
Logistic Regression	Synthetic	0.81	0.79
Random Forest	Real	0.91	0.90
Random Forest	Synthetic	0.88	0.86

C. Privacy Evaluation

Membership inference attack success ≈50% (random), attribute inference accuracy ↓18%, DCR = 0.12.

VI. DISCUSSION

Synthetic data created by GAN has a high utility/privacy balance. Distributional fidelity is established by using statistical measures of similarity. There is little loss of performance indicated by utility measures. Privacy measures are less re-identification risk.

Facing difficulties, there is mode collapse, modeling high dimensional categorical data, and no formal privacy guarantees without differential privacy.

VII. CONCLUSION

The given paper has introduced a full procedure of producing synthetic data with the help of GANs and assessing its quality. Findings indicate high utility and privacy. Future directions consist of the work on the integration of differential privacy, studying diffusion models, and benchmark standardization.

REFERENCES

1. Goodfellow et al., “Generative Adversarial Nets,” in Proc. NeurIPS, 2014.
2. M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” in Proc. ICML, 2017.
3. I. Gulrajani et al., “Improved Training of Wasserstein GANs,” in Proc. NeurIPS, 2017.
4. A. Figueira and B. Vaz, “Survey on Synthetic Data Generation, Evaluation

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/45
DOI: https://doi.org/10.5281/zenodo.20717846

Methods and GANs," Mathematics, vol. 10, no. 3, 2022.
5. GeeksforGeeks, "Exploring Synthetic Data Using GANs," 2025.
6. R. Thinakaran et al., "GANs for Synthetic Data Generation: Systematic Review," Int. J. Inf. Retrieval Res., vol. 15, no. 2, 2025.
7. L. Xu et al., "CTGAN: Synthesizing Tabular Data Using GANs," NeurIPS Workshop, 2019.
8. T. Karras et al., "Progressive Growing of GANs," in Proc. ICLR, 2018.
9. J. Jordon, J. Yoon, and M. van der Schaar, "PATE-GAN," in Proc. ICLR, 2019.
10. J. Yoon et al., "TimeGAN," in Proc. NeurIPS, 2019.
11. C. Bowles et al., "GAN-Based Medical Image Synthesis," IEEE Trans. Med. Imaging, vol. 37, no. 3, pp. 671–681, 2018.
12. A. Goncalves et al., "Synthetic Data in Healthcare," Nat. Digit. Med., vol. 3, no. 1, 2020.
13. P. Kaur et al., "Privacy-Preserving Synthetic Data," Artif. Intell. Rev., Springer, 2023.
14. H. Zhang et al., "Evaluation Metrics for Synthetic Data," ACM Comput. Surv., vol. 56, no. 2, 2024.
15. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
16. C. Aggarwal, Neural Networks and Deep Learning. Springer, 2018.
17. IBM Research, "Best Practices for Synthetic Data Generation," IBM Whitepaper, 2024.
18. Google AI, "Differential Privacy in ML Systems," Google Research, 2023.
19. NVIDIA, "GAN-Based Data Augmentation," NVIDIA Technical Report, 2024.
20. N. Papernot et al., "The Limitations of Deep Learning in Adversarial Settings," in Proc. IEEE EuroS&P, 2016.