

Generative AI for Synthetic Data Generation in Machine Learning: A Review and Evaluation Framework

Shravani Mahadev Karvande

Department of Artificial Intelligence and Machine Learning Engineering, Navsahyadri Group of Institutes, Faculty of Engineering, Pune, Maharashtra, India

Published October 2026

Abstract

Machine learning models need large, representative, labeled datasets, but real data are often scarce, imbalanced, costly to label, or restricted by privacy rules. Generative AI can produce synthetic data that keeps the statistical structure of a source dataset without reusing its records, although memorization remains a risk. This paper reviews four generator families: generative adversarial networks, variational autoencoders, diffusion models, and language-model-based tabular generators. It organizes evaluation along three axes: fidelity to real data, downstream utility for model training, and privacy leakage. A worked example shows how synthetic minority samples can rebalance a skewed dataset. The paper then proposes an evaluation protocol that combines train-on-synthetic, test-on-real experiments, differential-privacy accounting, and membership-inference testing. No measured performance result is claimed. The contribution is a structured comparison and an evaluation plan for researchers and practitioners.

Keywords: synthetic data generation; generative adversarial networks; variational autoencoders; diffusion models; tabular data; data augmentation; privacy-preserving machine learning

1 Introduction

Machine learning models learn from examples, so their quality depends on the data behind them. In many domains that data is scarce, imbalanced, costly to label, or restricted by privacy rules. Chen et al. [1] note that the shortage of annotated medical data has led to growing use of synthetic data in healthcare.

Generative AI offers one way to ease these limits. A generative model learns the distribution of a dataset and can draw new samples from it. Generative adversarial networks (GANs) [2], variational autoencoders (VAEs) [3], and diffusion models [4] are the main families used for this purpose. All three were first developed for images, and each has since been adapted to tables and time series.

Tabular data is harder to model than images. Its columns mix numerical, categorical, and ordinal features, and many datasets are small. CTGAN [5] adapted conditional GANs to mixed-type tables.

TabDDPM [6] applied diffusion models to tabular data and reported better results than GAN- and VAE-based baselines. GReaT [7] instead serializes each row as text and samples from an autoregressive language model. Jordon et al. [8] describe the main reasons for using synthetic data, including privacy-preserving sharing and augmentation of small datasets.

This paper reviews these methods and organizes their evaluation around three questions: whether the records look realistic, whether models trained on them perform well on real data, and whether they reveal information about real individuals.

The contribution is a structured comparison and an evaluation plan. No new generator is proposed, and no measured performance is claimed. Section 2 describes the sources and the research gap.

Section 3 compares the methods, presents an illustrative example, and states the evaluation dimensions. Sections 4 and 5 cover peer review and planned revisions, and Section 6 concludes.

2 Identify, research and collect idea

The topic came from a practical problem. Projects often need more training data than they can collect, or data they cannot share. Synthetic data is proposed for both cases, but the claims made for it are not always tested in the same way.

The sources reviewed fall into three groups. The first covers generator architectures [2]–[7]. The second covers evaluation, including the train-on-synthetic, test-on-real protocol that Esteban et al. [9] used for medical time series. The third covers privacy and risk: differentially private generators such as PATE-GAN [10], the formal guarantee of differential privacy [11], membership inference attacks [12], and recursive training effects [13]. Oversampling with SMOTE [14] serves as the classical baseline.

The research gap is practical rather than absolute. Results are often reported on different datasets, with different preprocessing and metrics, which makes direct comparison difficult. Reports often emphasize one or two of these dimensions rather than all three.

This paper addresses that gap by organizing the evaluation around fidelity, utility, and privacy. It also states what each test can and cannot show. The idea was collected from the literature above, not from new experiments.

3 Write down your studies and findings

3.1 Generation pipeline and generator families

A typical pipeline has five steps: preprocess the real data, fit a generator, sample new records, check the samples, and use them for training or sharing. For tables, preprocessing matters because numerical columns may need scaling and categorical columns need encoding.

Four generator families appear most often in the literature.

- **GANs** [2] train a generator against a discriminator that tries to tell real records from generated ones. CTGAN [5] adapts this idea to mixed-type tables. Training can be unstable, and the generator may cover only part of the data distribution, a failure called mode collapse.
- **VAEs** [3] learn an encoder and a decoder linked through a latent space. Training is usually more stable, but samples are often smoother than the real data.
- **Diffusion models** [4] learn to reverse a gradual noising process. TabDDPM [6] applies this idea to tables with mixed feature types. Sampling needs many steps, which raises compute cost.

- **Language-model generators** use an autoregressive language model. GReaT [7] fine-tunes such a model on text-serialized rows. It requires minimal preprocessing, but training and sampling are expensive.

3.2 Evaluation dimensions

Synthetic data should be judged on three dimensions, and each can pass or fail on its own.

Fidelity asks whether synthetic records resemble real ones. Common checks compare per-column distributions, for example with the Kolmogorov–Smirnov statistic for numerical columns, and compare the correlations between columns.

Utility asks whether a model trained on synthetic data performs well on real data. Esteban et al. [9] evaluated synthetic time series with this train-on-synthetic, test-on-real (TSTR) protocol. The result should be compared with a train-on-real, test-on-real (TRTR) baseline on the same real test set. The utility loss is then

$$\Delta_{\text{util}} = A_{\text{TRTR}} - A_{\text{TSTR}} \quad (1)$$

where A is the same score, such as accuracy or F1, for both models. A gap near zero means the synthetic data preserved the information this model needs, for this task.

Privacy asks whether the generator reveals information about training records. Distance-to-closest-record checks flag synthetic records that nearly copy a real one. Membership inference attacks test whether an attacker can tell training members from non-members [12]. On a balanced set, an attack accuracy near 0.5 means the attacker does no better than guessing. Chen et al. [1] warn that generative models used carelessly in healthcare may create re-identification risk. Formal guarantees come from differential privacy [11], which limits how much any single record can change the output. Differentially private generators such as PATE-GAN [10] offer such guarantees, at a utility cost that must be measured. Empirical tests can reveal leakage, but they cannot prove its absence.

3.3 Comparison with existing approaches

Table 1 compares the approaches. SMOTE is not a generative model; it is included as a classical baseline. The entries are qualitative and summarize design trade-offs reported in the cited works. They are not measured rankings.

Table 1. Comparison of synthetic data approaches.

Approach	Strengths	Limitations	Representative work
SMOTE oversampling	Simple and fast; no deep training	Interpolates between existing minority records; may create unrealistic points; no formal privacy guarantee	[14]
GAN	Flexible; often realistic samples	Unstable training; mode collapse; sensitive to tuning	[2], [5], [10]
VAE	Stable training; structured latent space	Smoother, less sharp samples	[3]
Diffusion model	High sample quality; stable training	Many sampling steps; higher compute cost	[4], [6]

Approach	Strengths	Limitations	Representative work
Language-model generator	Minimal preprocessing for mixed types	High training and sampling cost; results depend on text encoding	[7]

3.4 Illustrative example and evaluation plan

Consider a binary classification dataset with 1,000 records: 950 in the majority class and 50 in the minority class. A classifier that always predicts the majority class reaches 95% accuracy but detects no minority cases. Accuracy alone hides this failure, which is why Section 3.2 uses several measures.

To balance the classes by adding synthetic minority records, the number needed is

$$N_{\text{syn}} = N_{\text{maj}} - N_{\text{min}} = 950 - 50 = 900 \quad (2)$$

So 900 synthetic minority records bring the minority class to 950, giving 1,900 records in total. SMOTE [14] produces the same count by interpolation, and a generator produces it by sampling. This example is arithmetic only.

Three findings follow from the review. They are not experimental results.

1. Generator choice is a trade-off among sample quality, training stability, and compute cost.
2. Fidelity and utility do not show privacy. A generator can score well on both and still leak training records.
3. Every study should include a real-data baseline (TRTR) and a simple baseline such as SMOTE.

A measured study should follow these steps:

1. Use at least two public tabular datasets with known class imbalance, such as the UCI Adult income dataset.
2. Fix random seeds, repeat each run, and report the mean and spread.
3. Report fidelity, utility (TSTR and TRTR), and privacy results together.
4. Include SMOTE and at least one differentially private generator as baselines.

4 Get peer reviewed

Before submission, the manuscript should be reviewed by the project guide or a subject-matter expert. The reviewer should check that each described method matches its cited source, that the cited papers support each claim, and that the limitations are stated clearly.

No review has been completed for this draft, and this paper makes no claim of peer review. Reviewer name, review date, and comments will be added here after the review.

5 Improvement as per reviewer comments

No reviewer has commented on this draft yet. The changes below are anticipated from common review concerns. They are planned, not completed.

1. Add a measured experiment on at least two public tabular datasets, reporting fidelity, utility, and privacy results with their spread across repeated runs.

2. Add SMOTE and a differentially private generator as baselines, so the trade-offs are measured rather than described.
3. Discuss recursive training, in which generated data is reused to train later models. Shumailov et al. [13] report that this can cause model collapse, in which the tails of the original distribution disappear.
4. State privacy claims precisely. Empirical tests such as membership inference can reveal leakage but cannot prove privacy.
5. Document data licensing and the permitted uses of generated data.

6 Conclusion

Generative AI can produce useful synthetic data for machine learning, especially where real data is scarce, imbalanced, or sensitive. GANs, VAEs, diffusion models, and language-model generators each trade off sample quality, training stability, and compute cost.

Synthetic data is not automatically private or useful. Its value depends on fidelity, downstream utility, and privacy tests reported together. This paper provides a structured comparison and an evaluation plan. Future work should replace that plan with measured results on public datasets.

Acknowledgments

The author thanks Prof. Pawar A. N. and the Department of Artificial Intelligence and Machine Learning Engineering at Navashyadri Groups of Institute Faculty of Engineering for academic guidance and support during this work.

References

- [1] Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*. 2021;5:493–497. <https://doi.org/10.1038/s41551-021-00751-8>
- [2] Goodfellow I, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2014. p. 2672–2680.
- [3] Kingma DP, Welling M. Auto-encoding variational Bayes. In: *Proceedings of the International Conference on Learning Representations (ICLR)*; 2014. *arXiv:1312.6114*.
- [4] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2020. p. 6840–6851.
- [5] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2019.
- [6] Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. TabDDPM: Modelling tabular data with diffusion models. In: *Proc. 40th Int. Conf. Machine Learning (ICML)*; 2023. p. 17564–17579.
- [7] Borisov V, Seßler K, Leemann T, Pawelczyk M, Kasneci G. Language models are realistic tabular data generators. In: *Proc. Int. Conf. Learning Representations (ICLR)*; 2023.
- [8] Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, et al. Synthetic data—what, why and how? *arXiv:2205.03257 [Preprint]*. 2022.
- [9] Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv:1706.02633 [Preprint]*. 2017.
- [10] Jordon J, Yoon J, van der Schaar M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In: *Proc. Int. Conf. Learning Representations (ICLR)*; 2019.
- [11] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: *Proc. Theory of Cryptography Conf. (TCC)*; 2006. p. 265–284.

-
- [12] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: *Proc. IEEE Symp. Security and Privacy (S&P)*; 2017. p. 3–18.
 - [13] Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. *Nature*. 2024;631:755–759. <https://doi.org/10.1038/s41586-024-07566-y>
 - [14] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16:321–357.