Vol. 2, No. 2, pp.: 1-4, 2024

Synthetic Tabular Data Generation: A Benchmark of Six GAN-Based Methods on Financial Datasets

Nadia Petrova[1,*], **Samuel Adewale**[2], **Björn Lindström**[3]

[1] Laboratory for Financial Data Science, University of Zurich, Zurich, Switzerland, 8001

[2] Department of Statistics, University of Cape Town, Cape Town, South Africa, 7701

[3] School of Economics, Stockholm University, Stockholm, Sweden, 10691

[*] *nadia.petrova@df.uzh.ch*

Article Information	
Received	15 January 2024
Accepted	30 June 2024
DOI	https://doi.org/10.63646/datamind.2024.020201

Abstract

Synthetic data generation for tabular financial datasets presents a distinctive set of challenges relative to image or text synthesis: heterogeneous column types (continuous, categorical, temporal, binary), highly non-Gaussian marginal distributions characteristic of financial variables, complex conditional dependencies including temporal autocorrelations, and stringent privacy requirements driven by financial regulation. This paper presents a systematic benchmark of six GAN-based synthetic tabular data generation methods — CTGAN, TVAE, CopulaGAN, TableGAN, CTAB-GAN+, and REaLTabFormer — across three financial datasets: a retail credit application dataset, an institutional trade order dataset, and a customer transaction dataset. We evaluate fidelity, utility, and privacy across twelve metrics including Wasserstein distance, train-on-synthetic-test-on-real (TSTR) accuracy, statistical feature similarity, and membership inference attack success rate. CTAB-GAN+ achieves the best overall fidelity and utility balance, but no single method dominates across all metrics. We identify a systematic tradeoff between privacy protection and distributional fidelity that is more severe in financial data than reported benchmarks on general tabular datasets, and we discuss the implications for regulated financial data sharing.

Keywords: ***synthetic data; tabular data generation; GAN; financial data; privacy; data augmentation; CTGAN; benchmark***

1. Introduction

The demand for synthetic financial data has grown faster than the technical capacity to produce it reliably. Data science teams inside financial institutions regularly face the situation of having too little labelled data for rare events (fraud, default, large market moves) while being prohibited by regulatory and commercial confidentiality constraints from sharing the real data they have with researchers, vendors, or even internal teams outside specific access perimeters. Synthetic data generation promises to resolve this tension, and in specific applications it has delivered on that promise. In others, the synthetic data is subtly but consequentially different from the real data in ways that invalidate downstream modelling conclusions.

The particular characteristics of financial data make this harder than synthesising, say, health records or public census tables. Financial variables have heavy tails, asymmetric distributions, and

long-range temporal dependencies. The conditional structure between variables is complex — the relationship between transaction amount, merchant category, and fraud probability is not well-captured by factored approximations. And the privacy bar is high: financial records are among the most sensitive categories of personal data, and demonstrating privacy protection with respect to sophisticated membership inference attacks is a genuine challenge.

2. Methods Evaluated

We benchmark six methods spanning three architectural families. CTGAN and TVAE represent the original benchmark from Xu et al. (2019): CTGAN uses mode-specific normalisation and a conditional generator to handle mixed-type columns; TVAE uses a variational autoencoder with the same column encoding. CopulaGAN augments CTGAN with an explicit Gaussian copula step that pre-processes continuous columns to remove marginal non-Gaussianity before GAN training, then inverts this transformation in synthesis. TableGAN uses a convolutional generator operating on row-ordered feature representations. CTAB-GAN+ introduces additional conditioning on class-conditional and long-tail distributions, specifically designed for imbalanced categorical targets common in fraud detection. REaLTabFormer uses a transformer encoder over tabular rows — a more recent departure from the CNN/MLP-based architectures that dominate this space.

3. Datasets and Evaluation Framework

Dataset 1 is a retail credit application dataset (N = 150,000 rows, 42 features, 8.3% default rate) drawn from a European consumer bank (anonymised, used under data processing agreement). Dataset 2 is an institutional trade order dataset (N = 800,000 rows, 28 features, continuous timestamps) from an exchange data vendor. Dataset 3 is a customer transaction dataset (N = 2.1M rows, 15 features, 0.4% fraud rate) from a payment processor (anonymised, GDPR-compliant access). Each dataset is split 80/10/10 train/validation/test. Methods are trained on the training split; all evaluations are conducted on the held-out real test set.

Table 1. *Benchmark results across fidelity, utility, and privacy dimensions. Scores are normalised to [0,1] where 1 is best. An asterisk (*) indicates the best score in each column. Averages are unweighted.*

Method	Wasserstein ↓	KS Stat ↓	TSTR Acc ↑	MI Attack AUC ↓	Avg Score
CTGAN	0.72	0.68	0.71	0.58	0.68
TVAE	0.68	0.71	0.69	0.61	0.67
CopulaGAN	0.75	0.74	0.73	0.57	0.70
TableGAN	0.65	0.62	0.66	0.63	0.64
CTAB-GAN+	0.81*	0.79*	0.82*	0.55	0.74*
REaLTabFormer	0.78	0.77	0.80	0.52*	0.72

Wasserstein = 1 − normalised Wasserstein distance (lower better raw, higher score = better). KS Stat = 1 − average KS statistic across continuous features. TSTR Acc = AUROC on real test set for model trained on synthetic data. MI Attack AUC = 1 − membership inference attack AUROC (lower MI AUC raw = better privacy). Higher score always = better.

[Figure 1 — Radar plots comparing CTAB-GAN+ and REaLTabFormer across six evaluation dimensions (fidelity, marginal distributions, conditional structure, temporal consistency, utility, privacy) for each of the three financial datasets. The plots reveal that REaLTabFormer's advantage in privacy comes at a consistent

cost in temporal consistency on the trade order dataset.]

Figure 1. *Radar plots comparing CTAB-GAN+ and REaLTabFormer across six evaluation dimensions (fidelity, marginal distributions, conditional structure, temporal consistency, utility, privacy) for each of the three financial datasets. The plots reveal that REaLTabFormer's advantage in privacy comes at a consistent cost in temporal consistency on the trade order dataset.*

4. Key Findings

4.1 Fidelity-Privacy Tradeoff

The clearest finding from our benchmark is that the methods offering the strongest privacy protection — measured by resistance to membership inference attacks — systematically produce lower-fidelity synthetic data, as measured by distributional similarity metrics. REaLTabFormer achieves the best membership inference AUC across all three datasets, but this is accompanied by the largest KS statistic degradation on the transaction dataset's temporal autocorrelation structure. This tradeoff is more pronounced in financial data than in published benchmarks on general tabular datasets (e.g., UCI repository benchmarks), which we attribute to the more complex conditional dependency structure of financial records.

4.2 Performance on Imbalanced Targets

For the fraud detection use case on Dataset 3 (0.4% fraud rate), all methods produce synthetic data with inflated fraud rates relative to the real distribution — a well-known failure mode of conditional GANs on severe class imbalance. CTAB-GAN+ has specific architectural provisions for this case (conditional vector weighting) and shows the smallest inflation (0.4% → 0.9% synthetic fraud rate). The other methods inflate to between 1.8% and 4.3%. When TSTR models trained on this inflated synthetic data are evaluated on real data, the precision-recall characteristics are meaningfully distorted, with false positive rates elevated by 0.8–2.1 percentage points at standard operating thresholds. For fraud detection systems where false positive cost is substantial, this distortion is practically significant.

5. Conclusion

No single GAN-based method dominates across all financial data synthesis use cases. CTAB-GAN+ is our recommendation for applications where distributional fidelity is the primary concern; REaLTabFormer is preferable when privacy guarantees must be demonstrable against membership inference attacks. For temporal financial data, neither method fully resolves the challenge of maintaining long-range temporal autocorrelations under synthesis. The fidelity-privacy tradeoff we document is more acute in financial data than in domain-agnostic benchmarks, which we argue should motivate domain-specific evaluation standards for financial synthetic data.

References

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

Engelmann, J., & Lessmann, S. (2021). Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications, 174, 114582. https://doi.org/10.1016/j.eswa.2021.114582

Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: Effective table data synthesizing. ACML 2021. http://proceedings.mlr.press/v157/zhao21a.html

Solatorio, A. V., & Dupriez, O. (2023). REaLTabFormer: Generating realistic relational and tabular data using transformers. arXiv preprint arXiv:2302.02041. https://doi.org/10.48550/arXiv.2302.02041

Triastcyn, A., & Faltings, B. (2020). Federated generative privacy. IEEE Intelligent Systems, 35(4), 50–57. https://doi.org/10.1109/MIS.2020.2993532

Hittmeir, M., Mayer, R., & Ekelhart, A. (2019). Utility and privacy assessments of synthetic data for regression tasks. IEEE BigData 2019, 5764–5772. https://doi.org/10.1109/BigData47090.2019.9005979

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. Proceedings of the VLDB Endowment, 11(10), 1071–1083. https://doi.org/10.14778/3231751.3231757

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. CSF 2018, 268–282. https://doi.org/10.1109/CSF.2018.00027

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042