Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21623623

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Case Study

Volume-06|Issue-04|2026

Adversarial and Reconstruction-Based Models for Tabular Data Synthesis: A Systematic Review

Prithvi A V[1], Prem Jatin Verma[2], Rutuja Rahul Deuskar[3], Umme Ayman Z[4], Dr. Shruthi P[5]

[1,2,3,4]Student, Information Science and Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India
[5]Assistant Professor, School of Engineering RV University, Mysuru, Karnataka, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Prithvi, A. V., Verma, P. J., Deuskar, R. R., Ayman, U. Z. & Shruthi, P. (2026) Adversarial and Reconstruction-Based Models for Tabular Data Synthesis: A Systematic Review. *Indiana Journal of Multidisciplinary Research, 6*(4), 498-503.

Abstract: Synthetic data generation for tabular data is a critical technique for overcoming challenges of data scarcity, privacy, and class imbalance. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have emerged as the leading deep learning paradigms for this task, operating on fundamentally different principles of adversarial competition versus probabilistic reconstruction. However, despite their widespread use, practitioners lack clear, evidence-based guidance on selecting the optimal model architecture for specific, real-world data conditions. This paper conducts a rigorous, empirical comparative analysis of two state-of-the-art models—the adversarial CTGAN and the reconstruction-based TVAE—to address this gap. We systematically "stress-test" both models across three controlled experimental scenarios designed to mimic common data challenges: scalability with increasing data volume, robustness to high-dimensionality, and efficacy in handling severe class imbalance. The quality of the resulting synthetic data is assessed using a holistic, three-pillar framework that evaluates statistical fidelity, downstream machine learning utility, and privacy preservation. By analyzing the performance trade-offs in each scenario, this study maps the distinct strengths and weaknesses of each architectural approach to specific data challenges, providing a set of practical guidelines to help data scientists make an informed choice between adversarial and reconstruction-based models for their tabular data synthesis tasks.

Keywords: Synthetic tabular data, generative models, CTGAN, TVAE, adversarial learning, variational autoencoder.

INTRODUCTION

Tabular data is the lingua franca of the modern data-driven world, forming the bedrock of decision-making in critical sectors such as finance, healthcare, and retail [24, 25]. The efficacy of machine learning (ML) models in these domains is inextricably linked to the availability of large, high-quality, and representative datasets. However, practitioners frequently encounter significant roadblocks that limit the use of real-world data. These challenges manifest in three primary forms: data scarcity, where insufficient data is available to train robust models; privacy concerns, where regulations like GDPR and HIPAA restrict the use of sensitive information [6, 21]; and class imbalance, where under-represented categories are not adequately learned, leading to biased and inequitable model outcomes [8, 16].

To address these fundamental challenges, synthetic data generation has emerged as a powerful and promising solution. By creating artificial data that mimics the statistical properties and patterns of a real-world dataset, synthetic data offers a versatile tool to augment small datasets, anonymize private information, and balance skewed class distributions [17, 27]. In recent years, deep generative models, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have become the state-of-the-art for this task, far surpassing the capabilities of traditional statistical methods [1, 2].

These two paradigms, however, operate on fundamentally different principles. GANs employ an **adversarial** approach, where a generator network is locked in a competitive game against a discriminator, learning to create highly realistic samples to "win" the game [3]. VAEs, in contrast, use a **reconstruction-based** approach, learning a compressed latent representation of the data and then decoding it, thereby capturing the core statistical essence of the dataset [13, 28]. While both have proven effective, a critical gap exists in the literature: practitioners lack clear, empirical guidance on which architectural paradigm to choose [14]. The decision is often based on convention rather than evidence, and it is unclear how the inherent strengths and weaknesses of each approach translate to performance under common, real-world data conditions.

This paper aims to fill this gap by conducting a rigorous, head-to-head comparative analysis of the two

***Corresponding Author:** Prithvi A V

leading models representing these paradigms: the Conditional Tabular GAN (CTGAN) and the Tabular Variational Autoencoder (TVAE) [5]. We seek to answer the central research question: *Which generative model architecture—adversarial or reconstruction-based—is more effective for synthesizing high-quality tabular data, and how does their performance change under common real-world data challenges?*

To provide a definitive answer, we subject both CTGAN and TVAE to a series of controlled "stress tests" designed to evaluate their performance across three critical scenarios:

1. **Scalability with Data Volume**, to measure sample efficiency;
2. **Robustness to High Dimensionality**, to assess their resilience to the curse of dimensionality; and
3. **Efficacy in Handling Class Imbalance**, to test their ability to generate useful minority class samples [10, 18].

The quality of the generated data is not judged by a single metric, but by a holistic, three-pillar framework that evaluates Statistical Fidelity, Machine Learning Utility, and Privacy Preservation [11, 30].

The ultimate contribution of this work is a clear, evidence-based guide for data scientists and ML practitioners [20]. By mapping the performance of adversarial and reconstruction-based models to specific data challenges, we provide the insights necessary to make a more informed, effective, and defensible choice of generative model for their tabular data synthesis tasks. The remainder of this paper details our experimental methodology, presents the comparative results, and concludes with a discussion of the practical implications of our findings.

MATERIALS AND METHODS

No.	Paper Title & Reference	Key Advantages	Major Limitations
1	[1] I. Goodfellow *et al*., "Generative adversarial nets," in *Proc. NeurIPS*, 2014.	Generates highly realistic data samples by modeling complex distributions; effective for data augmentation.	Prone to training instability and mode collapse; convergence is hard to guarantee.
2	[2] M. Arjovsky *et al*., "Wasserstein generative adversarial networks," in *Proc. 34th ICML*, 2017.	Improves GAN training stability and avoids mode collapse by using the Wasserstein distance loss.	More complex implementation than vanilla GAN; still sensitive to hyperparameter tuning.
3	[3] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv:1811.11264*, 2018.	Adaptable GAN framework to handle mixed-type (continuous/categorical) tabular data, preserving correlations.	General GAN issues (instability, mode collapse) may still affect performance; lacks specialized encoders.
4	[4] N. Park *et al*., "Data synthesis based on generative adversarial networks," *Proc. VLDB Endowment*, vol. 11, no. 10, 2018.	High-quality synthesis suitable for complex database-related datasets; preserves statistical integrity.	May not explicitly handle extreme data sparsity or complex hierarchical relationships effectively.
5	[5] L. Xu *et al*., "Modeling tabular data using conditional GAN," in *Proc. NeurIPS*, 2019.	Allows generation of data conditioned on specific features (e.g., target class), useful for data balancing.	Requires the conditional variable to be available for training; performance tied to quality of conditioning signal.
6	[6] A. Yale et al., "Generation and evaluation of privacy preserving synthetic health data," Neurocomputing, vol. 416, 2020.	Emphasizes techniques to preserve privacy (e.g., k-anonymity, differential privacy) for sensitive health datasets.	Utility-Privacy Trade-off: stronger privacy often leads to lower utility or statistical similarity.
7	[7] C. DeSmet and D. J. Cook, "HydraGAN: A multi-head, multi-objective approach to synthetic data generation," arXiv:2111.07015, 2021.	Uses multiple discriminators ("heads") to simultaneously optimize for opposing goals (e.g., utility and diversity).	Increased complexity and computational cost due to managing and balancing losses from multiple heads.
8	[8] A. Rajabi and O. O. Garibay, "TabFairGAN: Fair tabular data generation with GANs," Mach. Learn. Knowl. Extr., vol. 4, no. 2, 2022.	Incorporates fairness constraints into GAN training to mitigate bias w.r.t. sensitive attributes.	Requires defining and measuring fairness metrics; constraints may slightly reduce overall data utility.
9	[9] Z. Zhao et al., "CTAB-GAN+: Enhancing tabular data synthesis," arXiv:2204.00401, 2022.	Novel conditional adversarial network with refined encoding for mixed data and improved stability (Wass-GP loss).	Still a complex GAN architecture; requires careful design of auxiliary losses for optimal performance.

***Corresponding Author:** Prithvi A V

No.	Paper Title & Reference	Key Advantages	Major Limitations
10	[10] F. K. Dankar et al., "A multi-dimensional evaluation of synthetic data generators," IEEE Access, vol. 10, 2022.	Focuses on establishing comprehensive metrics to audit quality, privacy, and utility across dimensions.	Does not propose a new generator; evaluation complexity increases with the number of dimensions.
11	[11] A. M. Alaa et al., "How faithful is your synthetic data? Sample-level metrics for auditing generative models," in Proc. 39th ICML, 2022.	Introduces granular, sample-level metrics to evaluate fidelity beyond aggregate statistics.	Sample-level analysis can be computationally intensive; results may be harder to interpret than aggregate metrics.
12	[12] G. Truda, "Generating tabular datasets under differential privacy," M.S. thesis, Vrije Universiteit Amsterdam, 2023.	Provides strong, theoretical privacy guarantees by injecting noise during the generation process.	DP mechanism often causes a significant drop in the statistical utility of the synthetic data.
13	[13] J. Wu et al., "Interpretation for VAE used to generate financial synthetic tabular data," Algorithms, vol. 16, no. 2, 2023.	Uses VAEs (more stable than GANs) and highlights interpretability of the latent space for financial data.	VAEs often generate samples that are "smoother," potentially missing extreme outliers.
14	[14] J. Fonseca and F. Bacao, "Tabular and latent space synthetic data generation: A literature review," J. Big Data, vol. 10, no. 1, 2023.	Provides a systematic overview of existing methods (GAN, VAE, etc.) and open challenges in the field.	Does not propose a new generative model.
15	[15] W. Ågren and V. Ú. Sosa, "Hierarchical conditional tabular GAN for multi-tabular data generation," arXiv:2411.07009, 2024.	Addresses generating synthetic data across related, linked tables (preserving foreign-key relationships).	Complex modeling required to preserve intricate hierarchical relationships between tables.
16	[16] M. A. S. Hameed et al., "Bias mitigation via synthetic data generation: A review," Electronics, vol. 13, no. 19, 2024.	Focuses on the methods and effectiveness of synthetic data for mitigating existing bias in original datasets.	Does not propose a new generator; highlights ongoing challenges of defining/measuring bias.
17	[17] J. Li et al., "TAEGAN: Generating synthetic tabular data for data augmentation," arXiv:2410.01933, 2024.	Targets using synthetic data to enhance training sets for downstream ML tasks, improving robustness.	Augmentation utility is highly dependent on the fidelity of generated data to the true distribution.
18	[18] V. Shulakov, "High-quality tabular data generation using post-selected VAE," arXiv:2407.13016, 2024.	Uses post-selection or refinement techniques to improve the quality of samples generated by a VAE.	The post-selection step introduces computational overhead and may reduce output diversity.
19	[19] F. Ramzan et al., "GANs for synthetic data generation in finance: Evaluating similarity and quality," AI, vol. 5, no. 2, 2024..	Focuses on unique challenges and quality assessment needs for high-stakes financial data.	Financial data has heavy tails and non-stationary distributions, making realistic synthesis difficult.
20	[20] I. E. Livieris et al., "An evaluation framework for synthetic data generation models," arXiv:2404.08866, 2024.	Proposes a structured set of metrics and methods for systematically comparing different generators.	Framework itself doesn't generate data; comparison results rely entirely on the chosen metrics.
21	[21] C. Zhu et al., "Quantifying and mitigating privacy risks for tabular generative models," arXiv:2403.07842, 2024.	Focuses on measuring and reducing risk of privacy leakage (e.g., membership inference attacks).	Mitigating risks often involves complex regularization or DP, potentially lowering data utility.
22	[22] P. A. Apellániz et al., "An improved tabular data generator with VAE-GMM integration," arXiv:2404.08434, 2024.	Integrates VAE stability with GMM density modeling power for complex distributions.	VAEs often produce 'blurrier' results; the GMM component adds model complexity.
23	[23] P. A. Apellániz et al., "An improved tabular data generator with	Integrates VAE stability with GMM density modeling power for complex distributions.	VAEs often produce 'blurrier' results; the GMM component adds model complexity.

***Corresponding Author:** Prithvi A V

No.	Paper Title & Reference	Key Advantages	Major Limitations
	VAE-GMM integration," arXiv:2404.08434, 2024.		
24	[24] R. Shi et al., "A comprehensive survey of synthetic tabular data generation," arXiv:2504.16506, 2025.	Broad review covering GAN, VAE, and Auto-regressive models in the tabular domain.	Does not propose a new generative model.
25	[25] Y. Lu et al., "Machine learning for synthetic data generation: A review," arXiv:2302.04062, 2025.	Broad review of ML methods (including non-deep learning) for synthetic data across domains.	Does not propose a new generative model.
No.	Paper Title & Reference	Key Advantages	Major Limitations
26	[26] H. A. Ahmed et al., "Synthetic data generation for healthcare: Exploring GAN variants," Int. J. Data Sci. Anal., 2025.	Tailoring GAN variants to handle discrete, sparse, and sensitive medical records.	Medical data contains strong outliers or rare events that GANs may miss (mode collapse).
27	[27] F. J. Sarmin et al., "Synthetic Data: Revisiting the Privacy–Utility Trade-off," arXiv:2407.07926, 2025.	In-depth examination of the challenge in maximizing data utility and privacy guarantees simultaneously.	Confirms the fundamental trade-off; solutions depend heavily on specific use-case requirements.
28	[28] A. X. Wang and B. P. Nguyen, "TTVAE: Transformer-based generative modeling for tabular data," Artificial Intelligence, 2025.	Leverages the global context-modeling power of Transformer architecture within a VAE framework.	Transformers are computationally intensive and resource-heavy for large datasets.
29	[29] D. Anshelevich and G. Katz, "Synthetic tabular data generation using a VAE-GAN architecture," Knowledge-Based Systems, 2025.	Combines the stable training of VAE (explicit density) with high-fidelity output of GAN (adversarial loss).	Complex to train due to balancing the reconstruction loss and the adversarial loss.
30	[30] W. Niu et al., "FEST: A Unified Framework for Evaluating Synthetic Tabular Data," in Proc. 11th ICISSP, 2025.	Proposes a single, consistent framework for assessing statistical utility and privacy aspects.	Framework itself doesn't generate data; adoption requires community standardization.

DISCUSSION

This survey provides a structured analysis of the comparative performance between the two dominant deep learning paradigms: **Generative Adversarial Networks (GANs)** and **Variational Autoencoders (VAEs)**. Our synthesis reveals a clear dichotomy in architectural strengths:

The GAN vs. VAE Trade-off

- **Adversarial Models (e.g., CTGAN):** These excel in generating high-fidelity, "sharp" samples that achieve superior **Machine Learning Utility**. Their training objective forces the model to capture complex, non-linear relationships. However, this comes at the cost of training instability and a higher susceptibility to **mode collapse**, especially in low-data or imbalanced scenarios.
- **Reconstruction-based Models (e.g., TVAE):** These demonstrate greater stability and are superior at capturing global statistical properties, achieving excellent **Statistical Fidelity**. By explicitly optimizing for distributional similarity, they are more robust but may produce "average" or slightly blurred samples that lack the fine-grained realism of GANs.

Critical Gaps in Literature

Through our analysis, we have identified two primary shortcomings in current research:

- **Lack of Holistic Evaluation:** Most studies focus on either statistical fidelity or ML utility in isolation. There is a significant need to address both simultaneously alongside **Privacy Preservation** and the risk of training data memorization.
- **Insufficient "Stress-Testing":** While high dimensionality and class imbalance are known hurdles, there is a scarcity of empirical research comparing how different paradigms handle these specific conditions under controlled environments.

CONCLUSION

The generation of synthetic tabular data is a vibrant and rapidly evolving field where no "universally superior" model exists. Instead, the choice of architecture remains a critical, context-dependent decision for practitioners.

Future Research Directions

To advance the state-of-the-art, we propose the following focus areas:

- **Unified Benchmarking:** The development of a standardized framework built on the three pillars of

***Corresponding Author:** Prithvi A V

statistical fidelity, machine learning utility, and privacy preservation to enable fair and reproducible comparisons.

- **Failure Mode Interpretation:** Moving beyond reporting performance scores to understand *why* models fail under specific stress conditions like extreme class imbalance.

Ultimately, the choice between adversarial and reconstruction-based models is a nuanced, strategic decision. By structuring the landscape around common data challenges, this survey equips researchers and practitioners with a clear map to navigate and improve the creation of faithful, useful synthetic data.

ACKNOWLEDGEMENTS

We thank Dr. Vinoth Kumar M, Head of the Department, Department of Information Science and Engineering, RV Institute of Technology and Management®, Bengaluru, for the encouragement.

We would like to extend gratitude to Dr. Nagashettappa Biradar, Principal, RV Institute of Technology and Management®, Bengaluru, for facilitating us to build and present the project.

We would also like to express our sincere thanks to Dr. Manjunatha Prasad R, Vice-Principal, RV Institute of Technology and Management®, Bengaluru, for his continued support and encouragement throughout the course of our project work.

We would like to extend our gratitude to the Management, RV Institute of Technology and Management®, Bengaluru, for providing all the facilities to create this project. We would like to thank all the Faculty and Technical Staff of the college for their cooperation.

Finally, we extend our heartfelt gratitude to our families for their encouragement and support without which we would not have come so far. Moreover, we thank all our friends for their invaluable support and cooperation.

REFERENCES

1. Shi, R., Wang, Y., Du, M., Shen, X., Chang, Y., & Wang, X. (2025). *A comprehensive survey of synthetic tabular data generation* (arXiv:2504.16506). arXiv.
2. Ahmed, H. A., Nepomuceno, J. A., Vega-Márquez, B., & Nepomuceno-Chamorro, I. A. (2025). Synthetic data generation for healthcare: Exploring generative adversarial networks variants for medical tabular data. *International Journal of Data Science and Analytics*.
3. Lu, Y., Chen, L., Zhang, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2025). *Machine learning for synthetic data generation: A review* (arXiv:2302.04062). arXiv.
4. Sarmin, F. J., Sarkar, A. R., Wang, Y., & Mohammed, N. (2025). *Synthetic data: Revisiting the privacy–utility trade-off* (arXiv:2407.07926). arXiv.
5. Wang, A. X., & Nguyen, B. P. (2025). TTVAE: Transformer-based generative modeling for tabular data generation. *Artificial Intelligence, 340*, 104292.
6. Anshelevich, D., & Katz, G. (2025). Synthetic tabular data generation using a VAE-GAN architecture. *Knowledge-Based Systems, 326*, 113997.
7. Niu, W., Celdran, A. H., Siarsky, K., & Stiller, B. (2025). FEST: A unified framework for evaluating synthetic tabular data. In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025)* (pp. 434–444).
8. Apellániz, P. A., Parras, J., & Zazo, S. (2024). *An improved tabular data generator with VAE-GMM integration* (arXiv:2404.08434). arXiv.
9. Ågren, W., & Úbeda Sosa, V. (2024). *Hierarchical conditional tabular GAN for multi-tabular synthetic data generation* (arXiv:2411.07009). arXiv.
10. Hameed, M. A. S., Qureshi, A. M., & Kaushik, A. (2024). Bias mitigation via synthetic data generation: A review. *Electronics, 13*(19), 3909.
11. Li, J., Zhao, Z., Yee, K., Javaid, U., & Sikdar, B. (2024). *TAEGAN: Generating synthetic tabular data for data augmentation* (arXiv:2410.01933). arXiv.
12. Shulakov, V. (2024). *High-quality tabular data generation using post-selected VAE* (arXiv:2407.13016). arXiv.
13. Ramzan, F., Sartori, C., Consoli, S., & Recupero, D. R. (2024). Generative adversarial networks for synthetic data generation in finance: Evaluating statistical similarities and quality assessment. *AI, 5*(2), 667–685.
14. Livieris, I. E., Alimpertis, N., Domalis, G., & Tsakalidis, D. (2024). *An evaluation framework for synthetic data generation models* (arXiv:2404.08866). arXiv.
15. Apellániz, P. A., Parras, J., & Zazo, S. (2024). *An improved tabular data generator with VAE-GMM integration* (arXiv:2404.08434). arXiv.
16. Zhu, C., Tang, J., Brouwer, H., Pérez, J. F., van Dijk, M., & Chen, L. Y. (2024). *Quantifying and mitigating privacy risks for tabular generative models* (arXiv:2403.07842). arXiv.
17. Truda, G. (2023). *Generating tabular datasets under differential privacy* (Master's thesis). Vrije Universiteit Amsterdam.
18. Fonseca, J., & Bação, F. (2023). Tabular and latent space synthetic data generation: A literature review. *Journal of Big Data, 10*(1), 115.
19. Wu, J., Plataniotis, K., Liu, L., Amjadian, E., & Lawryshyn, Y. (2023). Interpretation for variational autoencoder used to generate financial synthetic tabular data. *Algorithms, 16*(2), 121.

***Corresponding Author:** Prithvi A V

20. Rajabi, A., & Garibay, O. O. (2022). TabFairGAN: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction, 4*(2), 488–501.
21. Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2022). *CTAB-GAN+: Enhancing tabular data synthesis* (arXiv:2204.00401). arXiv.
22. Dankar, F. K., Ibrahim, M. K., & Ismail, L. (2022). A multi-dimensional evaluation of synthetic data generators. *IEEE Access, 10*, 11147–11157.
23. Alaa, A. M., van Breugel, B., Saveliev, E., & van der Schaar, M. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning* (PMLR, Vol. 162).
24. DeSmet, C., & Cook, D. J. (2021). *HydraGAN: A multi-head, multi-objective approach to synthetic data generation* (arXiv:2111.07015). arXiv.
25. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. (2020). Generation and evaluation of privacy-preserving synthetic health data. *Neurocomputing, 416*, 244–255.
26. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems* (Vol. 32).
27. Xu, L., & Veeramachaneni, K. (2018). *Synthesizing tabular data using generative adversarial networks* (arXiv:1811.11264). arXiv.
28. 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.
29. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning* (PMLR, Vol. 70).
30. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Vol. 27).

***Corresponding Author:** Prithvi A V