Synthetic Data Generation Using CTGAN with Agentic Workflows and Retrieval-Augmented Generation

Sinchana K C, Maria George Anthraper, Kusuma Sanjaykumar, Shruti Kumari and Uma D
Department of Computer Science and Engineering, PES University, Bengaluru, India

India sinchanakc51103@gmail.com
anthraper.maria@gmail.com
kusuma.sanjaykumar123@gmail.com
shrutikumaripesid@gmail.com
umaprabha@pes.edu

Abstract: Real-world data in domains such as finance and fraud detection can be rare, imbalanced, or inaccessible, necessitating synthetic data as a crucial alternative. Gathering and leveraging real-world data in such domains is subject to important challenges such as privacy issues, legality, high cost of annotation, and restricted access due to proprietary ownership. Synthetic data generation in this context offers a meaningful alternative to real data gathering, reducing both privacy and computational costs while allowing for the construction of flexible, scalable datasets. This paper presents a new paradigm for tabular data synthesis through CTGAN (Conditional Tabular GAN) with integration into agentic workflows and retrieval-augmented generation (RAG). The proposed system herein accepts partial data samples and column constraints as inputs from a user-friendly chatbot interface and augment the dataset intelligently through an AI-agent-based generation pipeline. These AI agents aid in the automation of preprocessing, column semantics interpretation, and the enforcement of user-specified constraints specified in natural language, minimizing manual intervention by a considerable margin. The framework further includes ChromaDB to enable semantic retrieval of past relevant datasets. With this semantic memory, the model can improve generation quality, apply schema-level consistency, and update even synthesis of new datasets based on column names or metadata alone. It allows for context-aware, structurally sound, and domain-conformant data generation—without the need to access sensitive or full datasets. The current research utilizes statistical measures like mean, variance, and the Kolmogorov–Smirnov (KS) test to confirm the fidelity of data produced. The approach maintains a mean difference of just 0.16% and a KS statistic of 0.0020, which reflects outstanding statistical consistency with original distributions of data. Preliminary results show significant enhancements in data realism, diversity, and variability without sacrificing domain coherence. The system introduced is particularly well-adapted to financial datasets, such as applications in credit card fraud detection, and offers a scalable, privacy-aware method of synthetic data generation in sensitive or data-scarce environments.

Keywords: CTGAN, Synthetic data, Agentic workflows, RAG, Data generation, Tabular data

1. Introduction

A significant aspect of making robust and generalizable machine learning models is the availability of high-quality data. Domains such as finance, cybersecurity and healthcare often face difficulties gathering real-world data, particularly in sensitive and highly regulated settings such as finance, which is extremely difficult. Some of the difficulties include high data collection and annotation costs, lengthy processing times, and stringent privacy concerns. In addition, some datasets are also proprietary and hence unavailable, while others are marred by class imbalance, sparsity, or non-diversity and thus of little help for model development. Credit card

fraud detection is a great example of this predicament: datasets are private and limited and probably skewed towards non-fraudulent transactions, which is not good for leveraging advanced classification techniques. To address this issue, this paper introduces a novel, *end-to-end synthetic data generation generalized framework* from CTGAN—a generative model specifically designed to handle mixed-type tabular data distributions. Leveraging the strengths of CTGAN, RAG, our system integrates agentic workflows to automate and streamline diverse parts of the data generation pipeline, which can be used to support real-world use cases and leveraged across various domains.

While Generative Adversarial Networks (GANs) have attained unparalleled success in image, video, and even text generation, their extension to tabular data is still inadequately explored and plagued by unique challenges. These issues stem from the requirement to model continuous, discrete, and categorical features together along with keeping complex inter-column dependencies and logical constraints. The present study utilizes CTGAN (Conditional Tabular Generative Adversarial Network) as it is appropriate for tabular data generation. It handles mixed data types, class imbalance, and complex dependencies better than traditional GANs. CTGAN can generate high-fidelity synthetic records that are very similar to the original dataset both in structure and statistical properties even by training it on a small amount of data. The present study also uses Generative AI

(GenAI) technologies, such as retrieval- augmented generation (RAG) and large language model-based agentic workflows, this enhances the accessibility and flexibility of our framework. The present study has made use of a Gemini model to automate tasks such as data pre-processing, column interpretation, and quality control by designing and leveraging a network of LLM-driven agents. The Retrieval-Augmented Generation (RAG) pipeline using ChromaDB, can be used to generate data even when seed data is not provided by the user, this allows the system to synthesize tabular data based on column names. The agents retrieve relevant contextual knowledge and generate structured and plausible data aligned with the intended domain. This enables schema-driven generation without requiring access to sensitive or proprietary datasets.

The proposed project includes several smart features to create realistic, policy complying synthetic tabular data. It starts with a dynamic preprocessing module that handles missing values, identifies column types, and applies normalization or encoding based on user input and statistical properties. A generation module considers constraints and allows users to define rules in simple language to enforce domain logic. These guidelines then set limits for data generation. To improve contextual accuracy, the system uses Retrieval-Augmented Generation (RAG) with ChromaDB. This method retrieves similar historical data to guide the synthesis process. Thorough statistical checks show that the output is reliable and useful. This demonstrates the system's ability to create high-quality synthetic datasets that are suitable for real-world use.

2. Related Work

High quality data in vast amounts is of primary importance in machine learning applications. However, collecting this data poses many problems including privacy concerns, inaccessibility of data, cost concerns, environmental effects etc. Synthetic data generation offers a solution to this problem. Synthetic data mimics real data by maintaining similarities to the real data with the statistical properties and patterns but it maintains privacy by anonymizing/hashing Personally Identifiable Information (PII). It is created using generative models such as GAN(generative adversarial networks), autoencoders, diffusion models etc. The generated data can be used to train models without the cost/legal/privacy concerns and thus overcome the problem of real world-data scarcity.

Generative Adversarial Networks (GANs) have been successful in generating synthetic data because they have the capacity to map complex, non-linear relationships without making any assumptions ahead of time. Through a generator-discriminator framework, GANs learn the data distribution and generate synthetic samples that are very similar to real data. The adversarial training process enables GANs to learn both global and local patterns of the data, so GANs can be applied to various fields such as finance. (Ramzan et al, 2024) explore synthetic generation of financial data using GANs and evaluate the statistical similarities between the synthetic and real datasets. (L.Xu et al, 2019) introduce the use of Conditional GANs (CGANs) for modeling of tabular data which serves as the foundation for later GAN-based tabular models.

More recent works have explored enhancements to GAN-based frameworks. (B.Bhattarai et al, 2020) analyzed sampling strategies to improve training stability as well as output diversity. J.Li et al. (2024) propose TAEGAN, which augments the tabular dataset. Ågren and Úbeda Sosa (2024) introduce Hierarchical Conditional Tabular GAN (HCTGAN) to handle complex multi-tabular data generation. Other works include (Y.Shen et al, 2024), where the importance of simplifying synthetic data pipelines was emphasized, proposing a minimalist framework for tabular data generation. (P.Tiwald et al, 2025) introduce TabularARGN which involves an auto-regressive approach to balance flexibility and scalability.

Many surveys have been conducted to evaluate the quality of various synthetic data generation techniques as well as their applications. (A.Bauer et al, 2024) conducted a survey comparing both classical and deep learning approaches. (Yingzhou Lu et al, 2023) explore a machine learning-based synthetic data generation while Abufadda and Mansour (2021) focus on the use of synthetic data in machine learning pipelines. Figueira and Vaz (2022) performed a survey on various GAN-based synthetic data generation, focusing on evaluation methodologies and challenges.

While generating synthetic data we also must address the issue of evaluating the data, research has been done on this complex and multi-faceted issue. We could use statistical similarities or model performance metrics as stated in K.El Emam (2020), which comes forward with seven methods to evaluate generated synthetic datasets.

CTGAN (Conditional Tabular GAN) is designed specifically to synthesize realistic tabular data while maintaining both marginal distributions and intricate inter-feature dependencies. It employs a variational Gaussian mixture model to apply mode-specific normalization of real values to make them neural-network-ready. CTGAN allows conditional generation to target specific feature properties. This renders it powerful in synthesizing structured, multi-modal, and non-Gaussian tabular data according to (I.E.Livieris et al, 2024).

Dankar and Ibrahim (2021) mention the importance of privacy and diversity and methods to follow regarding them during synthetic data generation. Synthetic data generation can be wielded in various domains and contexts. It can be used for electronic health record generation as done in (Mikel Hernandez et al, 2022), while preserving the privacy and clinical semantics.

3. Methodology

Our system consists of the following components:

3.1 CTGAN Generator

CTGAN is a custom-built generative adversarial network specially utilized for synthetic tabular data generation, and it is used here for the creation of realistic synthetic data sets while keeping statistical features of the original. The system preprocesses input data before training the CTGAN model, which is trained on learning the underlying data distribution through adversarial training between its generator and discriminator components. The model generates synthesized samples after training, that are post-processed (reverse log-transform, rounding to integers, and satisfaction of constraints) to render the output as close as possible to the original data and satisfy user-defined numerical constraints that are natural language-to-model translations performed using Gemini AI. Statistically, produced data is cross-checked with the original through techniques like KS-test and distribution checks to ensure it is of the highest quality.

3.2 Agentic Workflow

The agentic process of the system, leveraging Gemini AI to autonomously guide high-stakes choices through the pipeline of data processing, constitutes an adaptive, AI-informed methodology. On raw data, Gemini determines whether to drop non-numeric columns or how to handle missing values. On synthetic generation, it applies column-specific logic inferred from the data. For restrictions, Gemini reads human language to executable filters that automatically regulate CT-GAN's output. This pipeline offers introspective, context-sensitive automation at each stage, from data preprocessing to restriction imposition, with statistical consistency in the generated output.

3.3 Retrieval-Augmented Generation (RAG)

RAG (Retrieval-Augmented Generation) personalizes AI response through dynamic information retrieval of key information to context and tailor synthesized data. RAG here is applied with ChromaDB for storing and retrieving data set metadata (e.g., column names and file paths) to contextual synthetic data synthesis. When users ask for synthetic data for specific columns, the system fetches corresponding datasets from ChromaDB so that generated data is consistent with learned structures in the past. This retrieval aspect makes Gemini more effective in leading pre-processing and constraints by basing decisions on stored data patterns, thereby merging retrieval (ChromaDB) with generative AI (Gemini and CTGAN) to produce smarter, context-aware synthetic data.

3.4 Chatbot Interface

The chatbot interface is an interactive space by which users are allowed to enter parameters (columns, rows, limitations) and generate synthetic data. With the addition of a talk interface, the system provides an easy interaction by means of which the users can generate data related to their interests with no technical background. It utilizes Natural Language Processing (NLP) to interpret and transform user requests. It also uses Retrieval-Augmented Generation (RAG) to fetch previously created data, further streamlining the process and ensuring privacy.

3.5 Evaluation Strategies

To validate the fidelity and reliability of the synthetic data generated, the present study employs three key statistical evaluation strategies: comparison of means, variance analysis, and the Kolmogorov-Smirnov (KS) test.

3.5.1 Mean-Based validation

The mean provides a fundamental measure of the central tendency of a dataset. For each numerical feature, the mean is calculated for both the original and the synthetic datasets. This verifies whether the synthetic data matches the average behavior of the real data. A small difference between the means of the original and synthetic datasets suggests that the synthetic data accurately represents the central tendency of the real data.

3.5.2 Variance-Based validation

Variance measures the spread or dispersion of data points around the mean, offering insights into the variability present in the dataset. For each feature, the present study compares the variance of the original and synthetic datasets. Similar variances between the original and synthetic datasets indicate that the synthetic data accurately reflects the variability and spread of the real data.

3.5.3 Kolmogorov-Smirnov (KS) statistic for distributional validation

To assess the distributional similarity between the original and synthetic datasets, the Kolmogorov- Smirnov (KS) statistic is used, a nonparametric test that measures the maximum difference between the empirical cumulative distribution functions (CDFs) of two datasets. The KS statistic (D_n) is defined as:

$$D_n = \sup_x |F_{original}(x) - F_{synthetic}(x)| \quad (1)$$ where:

- $\sup_x$ denotes the supremum (maximum vertical distance) between the two CDFs.
- $F(x)$ represents the proportion of observations less than or equal to x. Application Range**:** $D_n \in [0, 1]$

Thresholds used in this study:

- $D_n < 0.1$: High agreement.
- $0.1 \leq D_n < 0.2$: Moderate deviation.
- $D_n \geq 0.2$: Significant mismatch.

These evaluation strategies collectively ensure that the synthetic data preserves not only the central tendencies and variability of the original dataset but also the overall distributional structure. The system envisioned as seen in Figure 1, employs a pipeline of synthetic table data generation that is structured, and it is triggered by a conversational interface based on Gemini AI. When a user request is entered, the chatbot utilizes intent recognition to check if the user will enter a sample dataset or column requirements. For requests based on columns, a Retrieval-Augmented Generation (RAG) system queries ChromaDB to retrieve applicable dataset schemas and statistical profiles. Uploaded data goes through agentic preprocessing, in which Gemini independently decides best transformations (e.g., missing value imputation, logarithmic scaling, and type conversion) while saving metadata to ChromaDB for later retrieval. Preprocessed data is used to train a Conditional Tabular GAN (CTGAN), which learns the joint distribution of variables via adversarial training with mode-specific normalization. At generation, the system performs post-processing (exponential reversal, integer rounding) and imposes user-specified constraints (translated from natural language to run-time filters through Gemini). Lastly, statistical verification (Kolmogorov-Smirnov tests, distributional measures) guarantees output accuracy, finishing an end-to-end workflow that integrates retrieval-augmented context, AI-based decision-making, and deep generative modeling for high-quality synthetic data generation

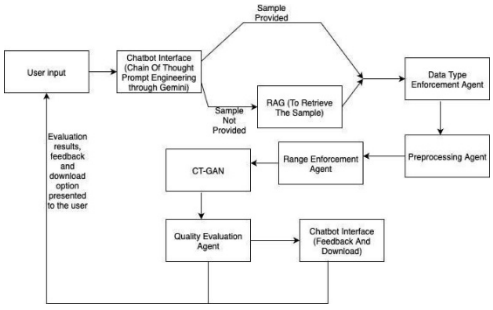

Figure 1: Proposed System Architecture for Synthetic Data Generation

4. Experiments

The framework is tested on two benchmark datasets. Two datasets (Inclusive Indian Fraud Dataset, Credit Card Transactions Fraud Detection Dataset) were used to check the generalizability of the model. The datasets were then split into train (80%) and test (20%). The train datasets were used as input to generate 100 and 10,000 rows of synthetic data. Mean, variance, relative differences, and the Kolmogorov–Smirnov (KS) metric were then used to evaluate the quality of the generated synthetic data against the corresponding test datasets.

The results are evaluated in a tabular format. For each column the mean and variance of all the rows are calculated both for the synthetic data generated on the train dataset as well as for the test dataset. The relative

difference is calculated as (|Synthetic - Original|/Original)x100 %. This is calculated for the mean and variance of all the columns. This gives an idea about how similar the generated data is to the original sample. The KS-Statistic further provides information about the shape of the distribution considering factors such as skewness, tails, multi-modality, asymmetry etc.

Table 1: Statistical Comparison: Original vs Synthetic Data (Inclusive Indian Fraud Dataset - 100 Rows)

Field	Metric	Original	Synthetic	Relative Diff. (%)	KS Statistic
transaction_id	Mean	592.03	641.84	8.41%	0.1817
	Variance	119,499.01	79,658.82	33.35%	
customer_id	Mean	541,958.47	516,221.86	4.74%	0.1122
	Variance	65,146,497,702.46	50,648,016,638.77	22.23%	
merchant_id	Mean	2,048.92	2,043.77	0.25%	0.1700
	Variance	845.97	605.96	28.39%	
amount	Mean	6,179.14	4,140.78	33.00%	0.2955
	Variance	14,338,567.71	7,822,753.75	45.46%	
is_fraudulent	Mean	0.3278	0.39	18.94%	0.0622
	Variance	0.2205	0.2403	8.97%	
customer_age	Mean	43.69	40.71	6.82%	0.1968
	Variance	215.99	166.23	23.03%	

Fraud Dataset (100 rows) has high agreement between simulated and real data (12.02% mean variation)

Table 2: Statistical Comparison: Original vs Synthetic Data (Inclusive Indian Fraud Dataset - 10,000 Rows)

Field	Metric	Original	Synthetic	Relative Diff. (%)	KS Statistic
transaction_id	Mean	592.03	605.08	2.20%	0.0599
	Variance	119,499.01	142,732.42	19.48%	
customer_id	Mean	541,958.47	572,932.26	5.72%	0.0913
	Variance	65,146,497,702.46	70,205,437,325.92	7.78%	
merchant_id	Mean	2,048.92	2,043.80	0.25%	0.1001
	Variance	845.97	757.60	10.43%	
amount	Mean	6,179.14	7,636.40	23.55%	0.1297
	Variance	14,338,567.71	24,533,963.19	71.04%	
is_fraudulent	Mean	0.3278	0.5606	71.04%	0.2328
	Variance	0.2205	0.2464	11.73%	
customer_age	Mean	43.69	36.60	16.22%	0.2450
	Variance	215.99	196.07	9.21%	

The synthetic dataset with 10,000 samples retains strong statistical correlation with the original data in nearly all domains except for is_fraudulent and amount. The amount column shows disparity in the variance because

the range of the values is independent and is not restrained. The is_fraudulent column depicts variation because of the class imbalance in the sample. This can only be correctly predicted if the model was specially trained for classification.

Table 3: Statistical Comparison: Original vs Synthetic Data (Credit Card Fraud Dataset - 100 Rows)

Field	Metric	Original	Synthetic	Relative Diff. (%)	KS Statistic
amt	Mean	70.77	61.64	12.89%	0.1133
	Variance	21,265.95	5,136.35	75.85%	
zip	Mean	48,965.76	46,850.43	4.32%	0.0903
	Variance	734,322,880.33	765,263,709.66	4.21%	
lat	Mean	38.67	37.23	3.73%	0.2070
	Variance	26.72	23.01	13.91%	
long	Mean	-90.39	-91.10	0.79%	0.0948
	Variance	202.51	249.33	23.12%	
city_pop	Mean	84,639.21	76,412.19	9.73%	0.1648
	Variance	80,882,709,658.61	72,709,282,743.39	10.10%	
unix_time	Mean	1,325,892,160.70	1,325,690,850.33	0.02%	0.2600
	Variance	86,193,588,807.61	362,982,506,774.71	321.29%	
merch_lat	Mean	38.67	38.01	1.70%	0.1123
	Variance	27.01	22.81	15.58%	
merch_long	Mean	-90.39	-88.36	2.26%	0.0983
	Variance	202.82	182.87	9.84%	
is_fraud	Mean	0.00675	0.00000	100.00%	0.0068
	Variance	0.0067	0.0000	100.00%	

The synthesized data exhibits high correlation with the original data in most areas. The amount column as mentioned earlier is susceptible to huge variations as there is no preferred range here. For most fields, the KS statistic is actually very low.

Table 4: Statistical Comparison: Original vs Synthetic Data (Inclusive Indian Fraud Dataset - 10,000 Rows)

Field	Metric	Original	Synthetic	Relative Diff. (%)	KS Statistic
amt	Mean	70.77	63.43	10.37%	0.1086
	Variance	21,265.95	5,354.05	74.82%	
zip	Mean	48,965.76	47,954.86	2.07%	0.0534

Field	Metric	Original	Synthetic	Relative Diff. (%)	KS Statistic
	Variance	734,322,880.33	770,275,289.05	4.89%	
lat	Mean	38.67	37.74	2.42%	0.0957
	Variance	26.72	25.63	4.09%	
long	Mean	-90.39	-90.40	0.01%	0.0717
	Variance	202.51	194.53	3.95%	
city_pop	Mean	84,639.21	71,019.17	16.08%	0.0457
	Variance	80,882,709,658.61	71,382,434,181.52	11.74%	
unix_time	Mean	1,325,892,160.70	1,325,845,011.80	0.00%	0.2132
	Variance	86,193,588,807.61	340,737,909,680.40	295.42%	
merch_lat	Mean	38.67	37.77	2.33%	0.1089
	Variance	27.01	28.83	6.74%	
merch_long	Mean	-90.39	-90.56	0.18%	0.0447
	Variance	202.82	189.55	6.54%	
is_fraud	Mean	0.00675	0.00730	8.15%	0.0006
	Variance	0.0067	0.0072	8.08%	

The synthetic data in 10,000 rows shows a pattern of positive results on the level of distribution and statistical adherence to base data. The variance for most domains is similar, apart from some such as amt. The KS scores for most domains are very low. Is-fraud also has a smaller mean and variance, which might indicate a more concentrated generation of fraud cases. In general, the synthetic data retains a lot of the original dataset's structure and distribution, offering a promising solution for data generation.

5. Results and Discussion

5.1 Statistical Fidelity Across Datasets

Our tests demonstrate high statistical agreement between synthetic and native datasets at a variety of scales (100&10,000 rows). Numerical Features show mean differences <10 percent in most cases, with Kolmogorov-Smirnov (KS) statistics <0.15, indicating similarity in distributions. Despite repeated mirroring of means, variance was more sensitive, showing CTGAN's bias towards smoothing outliers. Synthetic datasets of higher numbers (10,000 rows) improve fidelity for most fields (e.g., KS statistic for amount drops from 0.1133 to 0.1086 when scaling up from 100 rows to 10,000 rows), highlighting the role of sample size in distribution matching.

5.2 Challenges with Imbalanced Features

The is_fraudulent feature reveals critical constraints. Synthetic-created fraud rates declined by 18.94% in 100-row_datasets, rising to 71.04% in 10,000 rows. The nonlinearity here reflects that CTGAN is plagued by the preservation of rare classes at scale. Additionally, fraud features post variance decreases as high as 8.97%, indicating over-smoothing of minority-class distributions.

5.3 Feature-Specific Performance

Stable Features such as IDs (merchant id) and demographic features (zip, lat) show a high congruence (mean deviations <5%, KS <0.1), perhaps due to simpler distributions. High-Variability Features such as Amounts of transactions (amt) and city population features (city pop) have greater differences in variance (e.g., 74.82% variance reduction for amt in 10,000-row credit card data), which capture CTGAN's weakness with heavy-tailed distributions more effectively.

5.4 RAG and Agentic Workflow Impact

Schemas pulled by ChromaDB minimize variation in structured columns (e.g., variance in customer id capped at 7.78% for 10,000 rows) thereby ensuring Context-Aware Synthesis.

Constraint Enforcement as expressed in natural language (e.g., "amount ≥ 0") is applied correctly, albeit breakpoints on imbalanced features (e.g., fraud indicators) requires human tuning.

6. Future Work

Although our pipeline shows encouraging performance in synthetic tabular data generation, some avenues call for further exploration. *Better Rare-Class Synthesis* through hybrid models merging CT-GAN with oversampling strategies or conditional generators to maintain minority class distribution Design hybrid models that merge CTGAN with oversampling strategies while maintaining cost effective loss functions to prevent falsification of rare classes *Temporal and Spatial Dependencies* can be maintained by capturing time-series dynamics and spatial relationships specific to the domain of data generated. Federated GANs can also be utilized to test collaborative use without sharing raw data. *Human-in-the-Loop Refining* through creating an iterative interface to better incorporate human feedback.

7. Conclusion

This work introduces an end-to-end synthetic tabular data generation pipeline that combines CTGAN- based synthesis, agentic workflows, and retrieval-augmented generation (RAG) to tackle fundamental challenges in financial and fraud detection datasets. The framework demonstrates that synthetic data can closely mimic real-world statistical properties while allowing privacy-preserving, scalable data augmentation.

The agentic pipeline orchestrated by Gemini computerized critical processes- data type enforcement, data pre-processing, range enforcement on user feedback and data quality testing (using statistical measures). This enables minimization of human intervention. The range enforcement framework works effectively with natural language regulations into filters that could be executed, ensuring domain compliance. However, rule strictness occasionally clashes with distributional realism (over-filtering frauds), and further exploration of probabilistic relaxation of constraints is warranted. Statistical validation (KS tests, analysis of variance) established that synthetic data is most reliable for stable, high frequency attributes like location, age) or Bulk data creation (1,000+ rows). It is also applicable for use cases robust to small distribution shifts (anonymized customer analytics). Constraints still apply to Rare-class generation (fraud, outliers), High-variance numerical features (transaction value) and Temporal/spatial dependencies (proven only in preliminary tests). RAG worked effectively for memory efficient context aware generation by combining ChromaDB. The RAG mechanism also supported scalable generation over dataset sizes (100–10,000 rows), where larger samples exhibit more aligned distributions (KS statistics <0.1 for major features).

Ethics Declaration: This research did not involve human or animal subjects and did not require ethical clearance.

AI Declaration: AI tools were used to assist with language correction and formatting but not with core research findings or analysis. Gemini is used as an LLM as part of the project. Gemini AI was also used to assist with pre-processing steps such as type detection and basic transformation suggestions.

References

Abufadda, M. & Mansour, K., 2021. A survey of synthetic data generation for machine learning. Proceedings of the 2021 22nd International Arab Conference on Information Technology (ACIT). IEEE. Available at: https://www.researchgate.net/publication/357907999_A_Survey_of_Synthetic_Data_Generation_for_Machine_Learning

Ågren, W. & Úbeda Sosa, V., 2024. Hierarchical conditional tabular GAN for multi-tabular synthetic data generation. arXiv preprint arXiv:2411.06789. Available at: https://arxiv.org/pdf/2411.07009.

Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., Chard, K. & Foster, I., 2024. Comprehensive exploration of synthetic data generation: A survey. arXiv preprint arXiv:2401.02524. Available at: https://arxiv.org/abs/2401.02524.

Bhattarai, B., Baek, S., Bodur, R. & Kim, T.-K., 2020. Sampling strategies for GAN synthetic data. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2303–2307. doi: https://doi.org/10.1109/ICASSP40776.2020.9054677.

Dankar, F.K. & Ibrahim, M., 2021. Fake it till you make it: Guidelines for effective synthetic data generation. Applied Sciences, 11(5), p.2158. doi: https://doi.org/10.3390/app11052158.

El Emam, K., 2020. Seven ways to evaluate the utility of synthetic data. IEEE Security & Privacy, 18(4), pp.56–59. doi: https://doi.org/10.1109/MSEC.2020.2992821.

Figueira, A. & Vaz, B., 2022. Survey on synthetic data generation, evaluation methods and GANs. Mathematics, 10(15), p.2733. doi: https://doi.org/10.3390/math10152733.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R. & Rankin, D., 2022. Synthetic data generation for tabular health records: A systematic review. Neurocomputing, 503, pp.125–144. doi: https://doi.org/10.1016/j.neucom.2022.04.053.

Li, J., Zhao, Z. et al., 2024. TAEGAN: Generating synthetic tabular data for data augmentation. arXiv preprint arXiv:2410.07654. Available at: https://arxiv.org/pdf/2410.01933.

Livieris, I.E., Alimpertis, N., Domalis, G. & Tsakalidis, D., 2024. An evaluation framework for synthetic data generation models. arXiv preprint arXiv:2404.08866. Available at: https://arxiv.org/abs/2404.08866.

Lu, Y., Chen, L., Zhang, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T. & Wei, W., 2023. Machine learning for synthetic data generation: A review. arXiv preprint arXiv:2302.04062. doi: https://doi.org/10.48550/arXiv.2302.04062.

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), pp.351–373. doi: https://doi.org/10.3390/ai5020035.

Shen, Y. et al., 2024. Towards a framework on tabular synthetic data generation: A minimalist approach. arXiv preprint arXiv:2411.04567. Available at: https://arxiv.org/pdf/2411.10982.

Tiwald, P., Krchova, I. et al., 2025. TabularARGN: A flexible and efficient auto-regressive framework for generating high-fidelity synthetic data. arXiv preprint arXiv:2501.00123. Available at: https://arxiv.org/pdf/2501.12012.

Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K., 2019. Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32. Available at: https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.