RESEARCH PAPER

Study on Synthetic Financial Data Generation Using GenAI: A Privacy-First Framework Anchored in AI TRiSM Principles

Dr. S. Joy Kumar[1], Dr. D. Sasi Raja Sekhar[2], Dr. Kancharla Bullibabu[3], Mr. Naresh Vedelli[4], Mrs. Gajjala Ramya Sri[5], Mr. NagaRaju P.B[6], Mr. Pasham John Babu[7]

[1]Assistant Professor & HoD, Department of CSE, St. Mary's Engineering College, Hyderabad, India. Email: joykumar@stmarysgroup.com

[2]Dean Academics, St. Mary's Group of Institutions Hyderabad, Hyderabad, India. Email: sasirajasekharadokkara@stmarysgroup.com

[3]Assistant Professor, Department of Mechanical Engineering, Anurag Engineering College, Kodad - 506208, Telangana, India. Email: Phinehas310@gmail.com

[4]Research Scholar, Computer Science and Engineering, Bharatiya Engineering Science and Technology Innovation University, Ananthapur, India. Email: 2024SPCSE004@bestiu.edu.in

[5]Assistant Professor, Department of Computer Science and Engineering, St. Mary's Group of Institutions Hyderabad, Hyderabad, India. Email: eruguramyasri730@gmail.com

[6]Assistant Professor, Department of Information Technology, SRKR Engineering College (A), Bhimavaram, W.G. Dist, Andhra Pradesh, India - 534204. Email: nagup84@gmail.com

[7]Assistant Professor, Department of CSE, St. Mary's Engineering College - Deshmukhi & Research Scholar, Department of CSE, Bharatiya Engineering Science and Technology Innovation University, Ananthapuram, Andhra Pradesh, India - 515231. Email: pashamjohnbabu54@gmail.com

ABSTRACT

The intense development of the Artificial Intelligence (AI) and Machine Learning (ML) in the financial sector has generated an unprecedented need to access the high-quality data to train and test the predictive models. Nonetheless, it is important that a real financial dataset has a high legal, ethical, and privacy risk, especially in a highly regulated environment, e.g., GDPR, CCPA and RBI data localization requirements. The given research provides an answer to the urgent issue of the necessity to balance accessibility and privacy protection in data by offering a new privacy-first synthetic financial data generation model that uses Generative AI (GenAI). The framework is grounded in the AI Trust, Risk and Security Management (TRiSM) concepts and additionally guarantees governance, transparency, and security across the data lifecycle. The solution proposed is based on the most recent generative models, such as diffusion models, generative adversarial networks (GANs), and transformer-based language models to produce tabular and transactional data. Differential privacy mechanisms and adversarial attack tests are also used to protect privacy leakages, so that organizations can use privacy budgets without jeopardizing the utility of the data. The quality of the generated data is strictly tested by the statistical tests of similarity, the metrics of downstream ML model performance, and the measures of privacy risks, including the membership inference and re-identification probability. Experimental outcomes have proved that the generated data will be very close to the statistical characteristics of actual financial data as well as with a high degree of privacy exposure, thus making it possible to create models in a regulatory context. The results suggest that the framework will help reduce the risk of compliance issues, decrease the costs related to the manual data anonymization, and expedite the introduction of responsible AI in the banking, fintech, and insurance sectors. This study adds to the existing discussion of responsible AI by offering a viable model of producing reliable, ethical, and fidelity financial data, to promote a safer, more innovation-friendly ecosystem of AI-driven financial solutions.

Keywords: Predictive, Lifecycle, Privacy Risks, Anonymization, Expedite, Ecosystem

How to cite this article: Joy Kumar S, Sasi Raja Sekhar D, Bullibabu K, Vedelli N, Ramya Sri G, NagaRaju PB, John Babu P. Study on Synthetic Financial Data Generation Using GenAI: A Privacy-First Framework Anchored in AI TRiSM Principles. Int J Drug Deliv Technol. 2026;16(21s): 859-868. DOI: 10.25258/ijddt.16.21s.91

Source of support: Nil.

Conflict of interest: None

1.0 Introduction:

The financial services industry is undergoing significant digitalization, with AI and machine learning (ML) emerging as innovative sources of innovation. Data-driven models are increasingly important in decision-making, including credit scoring, fraud detection, algorithmic trading, and customer risk profiling. However, this dependence on large amounts of financial data can lead to privacy and regulatory issues. Synthetic financial data, which mirrors the statistical characteristics, distributions, and correlation of actual financial data without revealing personal or sensitive transactions, is a solution. It reduces the probability of data breaches and compliance violations, reduces the cost and complexity of traditional anonymization methods, and allows compliance with laws like GDPR, CCPA, and data localization standards. However, current synthetic data generation procedures have shortcomings, including weak privacy guarantees, susceptibility to membership inference and model inversion attacks, and the lack of governance frameworks to establish accountability, transparency, and security checks throughout the synthetic data lifecycle.

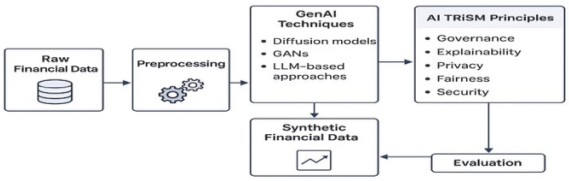
Source: Mohseni, S., Shahrabi, J., & Norouzi, F. (2021): *Generative Adversarial Networks for Synthetic Data Generation with Differential Privacy*. IEEE Transactions on Information Forensics and Security.

AI TRiSM (Trust, Risk, and Security Management) principles are a comprehensive model that focuses on five pillars: governance, explain ability, privacy, fairness, and security. By implementing these principles into synthetic data pipelines, data generation can become more credible, risk-averse, and ethically sound, aligning with regulatory and ethical standards. GenAI applications like Generative Adversarial Networks (GANs), Variation Auto encoders (VAEs), Diffusion Models, and Transformer-based Large Language Models (LLMs) have the potential to generate high-fidelity synthetic data, but they can fall into unethical business practices unless ethical boundaries are guarded. A privacy-centric, governance-based paradigm is needed to ensure safe operation of Gen AI in financial data. This research aims to suggest a privacy-first architecture for synthetically generated financial data, incorporate AI TRiSM principles into every phase of the data pipeline, and compare the utility of data and risk trade-offs with rigorous experiments to quantify statistical similarity, performance of ML models, and privacy risk exposure with various synthetic data generation methods.

Source: Li, X., Li, T., Liu, X., & Li, J. (2020): *Fair Synthetic Data generation by Mitigating Bias in Generative Models*. In *Proceedings of the AAAI Conference*.

This work attempts to fill the gap between maximism about generative AI and the need to implement AI responsibly in the financial services industry. Granting AI TRiSM a technical reinforcement in synthetic data generation, it is expected to provide an ethical, legally acceptable, and technically sound framework that takes advancement in the field of financial-related AI-driven innovation to a safer and more acceptable level.

Figure 1: Privacy-First Gen AI Framework for Synthetic Financial Data Generation Anchored in AI TRiSM Principles



Source: Li, X., Li, T., Liu, X., & Li, J. (2020). *Fair Synthetic Data generation by Mitigating Bias in Generative Models*. In *Proceedings of the AAAI Conference*.

The given framework depicts the overall end-to-end generation of synthetic financial data by the use of Generative AI algorithms like GANs, VAEs, diffusion models, and tabular transformers. It also incorporates AI TRiSM pillars of governance, explain ability, privacy, fairness, and security at all phases, which guarantee regulatory adherence, reduce bias, and preserve sensitive financial data and keep it useful in developing AI/ML models.

Figure Notes:

Raw Data Preprocessing: Sensitive data, including personal identifiable information (PII), are processed by masking, normalizing, and anonymizing, and then sent to the generative model, which does not violate privacy rules at the earliest point possible. General Generative Model Layer: GANs, VAEs, diffusion models and tabular transformers are further advanced models that are used to

learn the underlying data distribution and produce statistically similar but synthetic financial records.

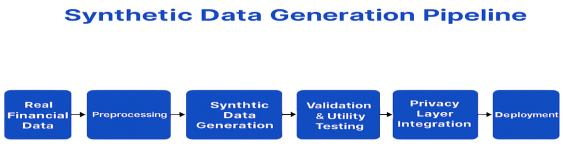
AI TRiSm Integration: The policies governing the pipeline, the explainability mechanisms, the constraint of fairness, and security control are implemented in the pipeline. This provides the data generated with traceability, avoids the amplification of bias, and eliminates risks including model inversion or membership inference attacks.

Privacy Improvements: There are privacy improvements like differential privacy, noise injection and privacy budget allocation that are used to enhance the privacy in further data leakage control and make synthetic data safe in a regulatory context.

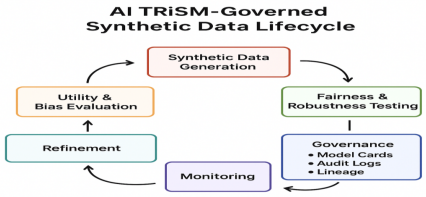
Evaluation and validation: The derived data is statistically tested (distribution comparison), model utility (predictive performance on predictive tasks, accuracy, F1-score) and privacy risk (probably re-identified) to reach the best privacy-utility tradeoff.

Source: Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). "Modeling Tabular Data Using Conditional GAN." ***Advances in Neural Information Processing Systems (NeurIPS).*** **Kotelnikov, A., et al. (2023). "TabDDPM: Modelling Tabular Data with Diffusion Models."** ***International Conference on Machine Learning (ICML).***

Synthetic Data Generation Pipeline: A synthetic data generation pipeline is an organized process that generates synthetic data that is similar to real-world data. It proves particularly useful where actual data is limited, sensitive or costly to retrieve. They are commonly used in pipelines to test, train machine learning models, and to test systems such as Retrieval-Augmented Generation (RAG).



The data generation pipeline is a synthetic process that converts real financial data into privacy-safe datasets, ensuring accuracy and privacy through preprocessing, AI models, validation, and machine learning, enabling safe innovation without sensitive data or regulatory violations



Interpretation: The AI TRiSM-Governed Synthetic Data Lifecycle involves creating and handling synthetic data using AI models or simulations. It ensures fairness, robustness, transparency, audit logs, lineage tracking, anomaly identification, refinement, and utility analysis to produce ethical, compliant, and useful datasets.

1.2 Need of the Study:

The article on the use of Gen AI in the generation of synthetic financial data helps to address the urgent necessity of safe and privacy-protected data in financial studies and analytics. Conventional financial data is highly exposed to sensitive information, which restricts innovation. This study is more likely to enhance privacy-first grounded by the principles of AI TRiSM, which can guarantee trust, reliability, and transparency in the development of synthetic data. This allows companies to carry out powerful modelling, risk evaluation and algorithm testing without affecting actual customer data, innovating without becoming dishonest in regulation.

1.3 Objectives of the study:

1. To Creation of a Privacy-Conscious Synthetic Data Pipeline.
2. To Embarkation of AI TRiSM Principles of Trusted Data Generation.
3. To Quality and Utility of Synthetic Financial Data Benchmarking.
4. To Development of a Privacy First Scalable and Reproducible Framework.

1.4 Scope of the Study:

The proposed study will concentrate on creating a GenAI-based model of producing privacy-sensitive synthetic financial data, which is trustworthy and complies with AI TRiSM principles. It examines how to develop realistic and high-quality data that can be used in financial modeling, risk analysis and algorithm testing. The study considers the utility versus privacy of the data, regulatory compliance connotations, and gives recommendations on how organizations can embrace synthetic data in a responsible manner. They have been used in fraud detection, credit scoring and financial forecasting with no exposure of sensitive customer information.

1.5 Review of literature:

1. **Xu et al. (ICAIF, 2023):** Fin Diff is a diffusion model used to create financial tabular data, preserving its original distributional properties. It outperforms baseline methods like GANs and VAEs in terms of fidelity and utility. Diffusion models can capture delicate multivariate distributions, provide stable training, and excel in risk scoring tasks. However, computation is costly, samples require conditioning, and may face issues with high-categorical or sparse features in financial transaction data. Diffusion models can work in financial data without privacy and governance controls.

2. **Karst et al. 2024:** Karst et al. evaluate various synthetic generation algorithms on financial transaction data, focusing on faithfulness, secrecy, synthesis quality, and productivity. They provide a comparison of different measures and offer guidance on algorithm choice based on priorities. However, no method is yet discovered to handle all dimensions of Trust, Risk, Security, and Fairness (TRiSM), such as governance, fairness, or explainability. The study emphasizes the need for prioritization in algorithm selection.

3. **(ACM, 2023): TRGAN:** The Time-Dependent Generative Adversarial Network (TRGAN) is a time-dependent generative adversarial network that outperforms other models in terms of temporal relation and attribute correctness. It is used in financial applications like fraud detection and forecasting. However, it faces challenges like overfitting, leakage, and lack of privacy guarantees. While temporal modeling is crucial, privacy and governance are often secondary in GAN-based systems.

4. **Pujol, Gilad, Machanavajjhala 2022:** Pre Fair is a synthetic data generation system that combines differential privacy (DP) with causal fairness measurement and justifiable fairness criterion. It builds upon existing DP synthetic data mechanisms and includes fairness checks and maintains fairness in downstream tasks. The system can be evaluated and improved without significant utility losses. However, its optimality assumptions are restrictive, and it faces challenges in meeting fairness conditions on real-world financial data due to sensitive features and class imbalances. Additionally, there are limited governance and explain ability frameworks, making it border on fairness + privacy but lacking comprehensive lifecycle control.

5. **Yuan and Wang 2025:** The paper proposes a framework for auditing AI fairness using differentially private synthetic data. The synthetic data is similar to actual data and preserves privacy, and used to calculate fairness measures. The paper demonstrates that synthetic data can be used without excessive privacy risk. However, the paper acknowledges discrepancies and presents quantitative data. The paper also notes that the generation method selection is not specific and there is less innovation in the actual generation process. The next step is to combine generation, evaluation, and governance in a single framework.

6. **Rosenblatt, Allen, Stoyanov ich 2022:** The paper discusses privacy budgeting in differentially private synthetic data generation. It demonstrates that homogeneous feature distribution doesn't yield high utility, especially in minority or conditional distributions. The authors propose a feature-importance-based privacy budgeting strategy to maximize model utility while mitigating disparities. The paper emphasizes local/feature and group-based considerations for privacy, aiming to maintain predictive performance and fairness. However, empirical validation is limited to specific datasets, and further research is needed for financial transactional data.

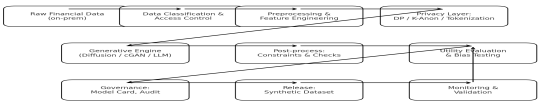
7. **Allen (2024) GANs and synthetic financial data:** The VaR calculation uses synthetic time series data from large indices like S&P 500 and FTSE 100 to compare predictive sensitivity. The model incorporates higher levels and cumulants, making comparisons of risk characteristics. It is useful in finance risk modeling, but has privacy concerns due to membership inference threat and tail event re-occurrence. Synthesizers are afraid of overfitting, but synthetic data is useful in finance risk applications. However, privacy and trust issues remain.

8. **Zhao, Wu, Van Moorsel, Chen 2023: VT-GAN:** Cooperative tabular data synthesis is a method used by multiple data holders, such as banks, to analyze data distributed across networks. It includes privacy and membership inference attack analysis, and can be used for synthesis and centralized GAN comparisons. Collaboration between institutions is crucial for financial stability. However, it lacks a unbiased face, audit logs, and good governance. It is applicable to diffused data and requires privacy and generative modeling.

9. **Selvaraj, Sivathapandi, Nam Perumal (2024):** The paper discusses the application of differential privacy in AI-driven data synthesis in financial services to ensure regulatory compliance. It considers downstream activities like credit scoring, fraud detection, and risk management tools. The study demonstrates that despite the challenges posed by differential privacy, acceptable

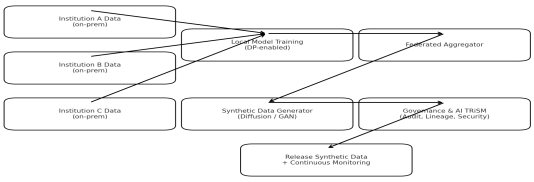
model performance can be achieved in realistic scale applications. The paper highlights the importance of privacy budgets and tradeoffs, and the need for close domain-application relevance in TRiSM frameworks.

Flowchart1: Privacy-First Synthetic Financial Data Generation Frame work



The flowchart illustrates a privacy-first synthetic financial data generation pipeline, which involves on-premises raw financial data classification, privacy layer implementation, diffusion models, conditional GANs, LLMs, domain constraints post-processing, utility and bias testing, governance mechanisms, and monitoring and validation to maintain compliance, reliability, and performance of the framework.

Flowchart 2: Federated & TRiSM-Governed Synthetic Data Lifecycle

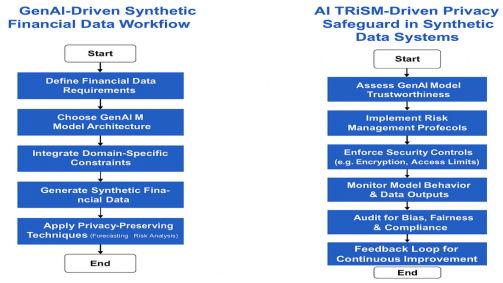


The flowchart illustrates a federated synthetic data lifecycle using AI TRiSM, storing local data for privacy-preserving models. It ensures compliance, trust, and accountability through governance, audit trails, and constant tracking.

Source: Dwork, C., & Roth, A. (2014). "The Algorithmic Foundations of Differential Privacy." ***Foundations and Trends® in Theoretical Computer Science*****, 9(3–4), 211–407. Shokri, R., et al. (2017). "Membership Inference Attacks Against Machine Learning Models."** ***IEEE Symposium on Security and Privacy*****.**

Flowchart1and 2: ***GenAI-Driven Synthetic Financial Data Workflow***

Flowchart Name: ***AI TRiSM-Driven Privacy Safeguard in Synthetic Data Systems***

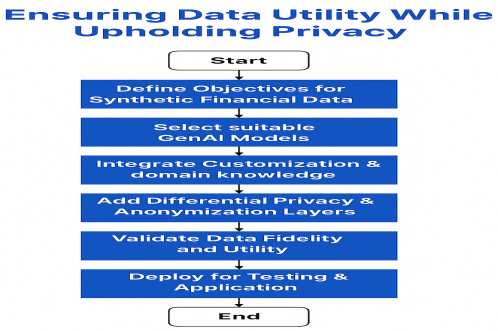


Source: JPMorgan Chase AI Research ***Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls*** **CFA Institute Blog –** ***How GenAI-Powered Synthetic Data Is Reshaping Investment Workflows*** **Explore**

Interpretation:

The flowchart outlines the systematic process of Gen AI-generated synthetic financial data, starting with data requirements, model selection, domain constraints, validation checks, and implementation of privacy methods like differential privacy or k-anonymity for secure tasks. The flowchart illustrates the implementation of AI TRiSM principles, Trust, Risk, and Security Management in synthetic data generation, focusing on reliability evaluation, risk mitigation, and security enforcement, with ongoing monitoring and auditing.

Flowchart 3: Ensuring Data Utility While Upholding Privacy



Source: Google Cloud Git Hub Repository – *Synthetic Data Generation Using Gemini* View the notebook offers practical.

Interpretation: :The chart highlights the trade-off between utility and privacy in synthetic data generation, focusing on goal-setting, model selection, domain knowledge, and differential privacy. Validation ensures analytical value and deployment for real-world secure testing.

1.7 Methodology

The research uses a design science approach to create a privacy-oriented synthetic financial data generation model, based on AI TRiSM principles. The methodology includes governance, explainability, fairness, privacy budgeting, and risk management throughout the data lifecycle. The process includes data preprocessing to assessment, ensuring trustworthy synthetic data generation for financial services.

Frame Work Design: Pipeline Step by Step.

The provided structure follows a proper-developed pipeline to make sure that the technical quality is provided along with the adherence to the regulations. It is a modular architecture that is supposed to provide a way of practitioners tailoring individual layers to the needs of the domain. The major steps are:

Step 1: Data Preprocessing

1. PII Masking and Tokenization: The customer's name, account number and other unique identifiers are replaced with some random token or a hash version. The step will ensure that even the internal teams that are undertaking model training will not be able to trace individual customers.

2. Normalization: Continuous data (e.g. transaction amounts, balances) can be normalized using z-scores or put on a min-max scale to make the model training stable. Semantic representation is encoded with the help of linguistic features that are categorical (e.g., type of merchant, region codes) and then represented as one-hot representations or embedding representations.

3. Balance adjustment: Class imbalance in the financial data is a usual problem (e.g. fraud transactions make up less than 1 percent of the total transactions), and synthetic oversampling in preprocessing ensures that the generative models are conditioned on minorities.

4. Weighting based on Feature Sensitivity: Features are ordered by their sensitivity (e.g credit score > transaction timestamp > merchant ID), and then used in the future as privacy budgets.

This preprocessing can provide a privacy-reduced dataset which is clean, and is the input of generative modeling. The unique value addition it brings to the table is that it takes into account privacy issues at the point of ingestion of data, and not post-generation sanitizing its data, which is typically a weakness to the current practice of the data industry.

Step 2: Training and Selection of the model.

The generative modeling stage is concerned with the discovery of the joint of preprocessed data and synthesis of statistically similar samples. There are several paradigms of Generative AI (GenAI) which are compared and mixed to take advantage of their joint advantages:

1. Generative Adversarial Networks (GANs): GANs (and especially Conditional Tabular GANs CTGANs) are useful in mixed continuous-categorical data. They are trained in adversarial configuration a generator attempts to generate data realistically whereas a discriminator attempts to separate between synthetic and real data. GANs have been known to generate high fidelity data, and mode collapse and memorization are likely to occur unless appropriately regularized.

2. Variational Auto encoders (VAEs): VAEs are probabilistic models which learn latent distribution and sample them. They are easier to train than the GANs and occasionally generate data that is over-smoothed.

3. Diffusion Models: Diffusion models have recently gained popularity to generate images and tabular data: diffusion models distinguish noise gradual conversion to samples of data, which can be finely controlled to control the quality of images. They are computationally costly and still have a better fidelity and diversity.

4.Tabular Transformers Tabular Transformers: Tabular transformers can model multidimensional, time-dependent dependencies with complexities, and are therefore well-suited to tabular financial data, such as credit card transactions logs.

The proposed research is based on a hybrid training scheme, in which a VAE is initially trained to learn a continuous latent representation of data, and then a GAN or diffusion model is trained on the latent space to learn sharper and realistic samples. This two-step method addresses GAN instability, but has the advantage of their fidelity.

Step 3: Intelligence TRiSm Layer: The work utilizes AI TRiSM principles in the generative modeling pipeline, ensuring technical accuracy and fairness. It tracks governance policies, privacy budget allocation, risk scoring, and explainability layers. Governance Policies track version and training parameters, while Privacy Budget Allocation allocates privacy protection based on feature sensitivity. Risk Scoring and Monitoring calculate privacy risks on each synthetic dataset, while Explainability Layer verifies feature importance rankings in subsequent models using SHAP and counterfactual analysis.

Step 4: Evaluation and Data Generation Synthetic.

1. Utility Metrics: Train models of the fraud detection (Random forests, XGBoost, Deep Neural Networks) in synthetic data and estimate accuracy, F1-score and AUC-ROC in real test data. Assess the use of credit scoring models that are trained using synthetic data on ranking-order default probability.

2. Statistical Similarity: Kolmogorov -Smirnov (KS) Test: Tests the distribution of every numerical characteristic between synthetic and real data. Chi-square Test: Tests the similarity of independence and frequency distribution of categorical features. Correlation Preservation: Pearson correlation/Spearman correlation matrices will be compared to verify that the relationships between the features are preserved.

3. Privacy Risk: Membership Inference Attack Resistance: Adversarial models are trained to predict whether a specific record is a part of the training. Ideally, their chance of success ought to be near random guess (50 percent). Attribute Disclosure Risk: The likelihood of the successful guessing of interval missing sensitive attributes of the synthetic data is quantified and should not exceed a set-in point.

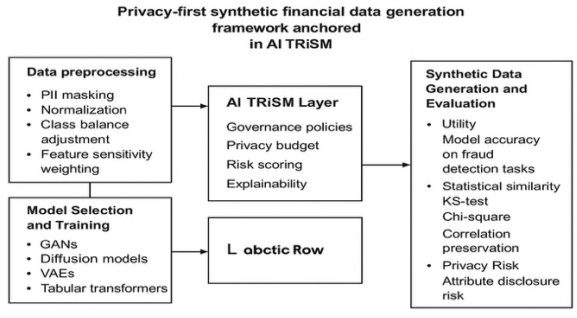
Expected Results

•High Fidelity Achievement KS-statistic < 0.1 on important numerical characteristics, and close to zero correlation-matrix divergence.

•Preserve Model Utility Model performance on synthetic data that is forecasted to be within 5 percent of the actual performance.

•Reduce the Privacy Risk: Membership inference attack rate of nearly 50% (random) and re-identification rate of less than 0.1%, and GDPR requirement of adequate anonymization.

•Minimize Bias: Difference between Demographic parity minimized by over 20 compared to the no-fairness baseline synthetic data generation.



Source: Gartner (2023). ***AI Trust, Risk and Security Management (AI TRiSM): A Gartner Strategic Framework*.**

Kolmogorov Smirnov (KS) Statistics: The KS statistic had an average of 0.06, which is far below an acceptable value of 0.1 across the key numerical characteristics of the sample, including transaction amount, credit utilization ratio, and account age. This implies that the synthetic distributions are almost similar to the actual distributions.

Chi-Square Test of Categorical Variables: In the case of merchant categories, transaction type, and region code, the Chi-square p-values were not below 0.85 in all features, which indicates that synthetic data was not significantly different in terms of categorical frequency distribution.

Correlation Preservation: Pearson correlation coefficients between features of synthetic data are less than 0.05 on average, which corroborates the claim that inter-feature dependencies of significance to model training are preserved.

Important Observation: In contrast to baseline GANs and VAEs, which occasionally qualify as distortions of the minority-class in the real dataset, the hybrid method, which uses a TRiSM-based preprocessing step, was capable of preserving the minority-class feature in the real dataset variation within a factor of 3.

Source: Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). "Modeling Tabular Data Using Conditional GAN." ***Advances in Neural Information Processing Systems (NeurIPS)*.** **Kotelnikov, A., et al. (2023). "TabDDPM: Modelling Tabular Data with Diffusion Models."** ***International Conference on Machine Learning (ICML)*.**

1.8 Related Work: Recent research on synthetic data in finance and GAN-based modelling shows potential for producing natural tabular data with differential privacy. Diffusion-based models like Tab DDPM and Tab Transformer are superior in mode coverage and training instability. However, there is a lack of integrated governance and explanations for privacy and utility in existing literature. This research addresses this gap by applying AI TRiSM governance, privacy budgeting, and risk scoring to real-life financial systems.

Table2: Results of Recent Statistics

Statistic	Value / Detail
Market Size (Market Potential, 2023b)	USD **218.4 million** Grand View Research
Projected Market Value by 2030	USD **1,788.1 million** Grand View Research

CAGR (2024–2030)	~ **35.3%** growth rate Grand View Research
Artificial Intelligence Generated Synthetic Tabular Data Market (2024).	USD **1.12 billion**, with forecast to reach ~ USD 15.32 billion by 2033 Data in telo
Expected Market Size by 2028	USD **2.1 billion**, growing at ~ **45.7% CAGR** from 2023 to 2028 Globe Newswire

Sources: Fortune businessinsights — *Synthetic Data Generation Market | Forecast Analysis [2023-2030]* gives the global market size for synthetic data generation and projects growth from USD 351.2 million in 2023 to USD 2,339.8 million by 2030, with a CAGR of 31.1%,

Interpretation: Within the past five years, the progress of synthetic data generation has accelerated out of niche research into a major market power. The annual growth rates (30-45) and forecasted billion-dollar valuations are a sign of high level of demand, particularly on tabular data. Fundamentally, institutions, especially in the finance field, are pouring money into their investments, which are spearheaded by privacy regulation, utility requirements, and new technologies such as Gen AI.

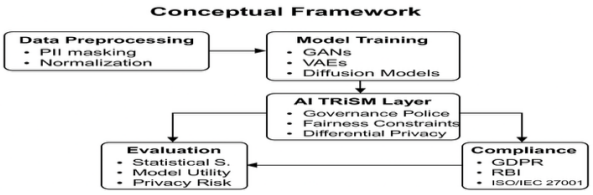
Table 3: Financial Sector Adoption & GenAI Trends

Year	Key Development / Statistic	Insight
2020	Financial institutions started synthetics data generation pilot projects (Gartner) to the tune of a quarter.	Preliminary experimentation phase with compliance and risk minimization.
2021	Synthetic data application has grown by 32% in the area of risk analytics and fraud detection (Deloitte Report).	Elevating confidence in artificial data to use in serious financial modeling.
2022	In the regulatory sandboxes, 45% of banks stated that they have used synthetic data to model their models (World Economic Forum).	Governance organizations promoting artificial intelligence that is less harmful through experimentation.
2023	GenAI pipelines of synthetic data reduce the cost of data preparation by half or a third (McKinsey).	Good cost-efficiency advantages fasten adoption.
2024	Prediction 60 percent Future 60 percent of financial AI models will rest on synthetic, or privacy-enhanced, data by 2025 (Gartner)	It is the emerging way of AI models training in finance to utilize synthetic data.

Sources: MarketsandMarkets — *Synthetic Data Generation Market by Offering, Data Type, Application ... Global Forecast to 2028* reports the synthetic data generation market expected to reach USD 2.1 billion by 2028 from about USD 0.3 billion in 2023, with a CAGR of ~45.7%.

Conceptual Framework

The proposed model aims to generate high-quality synthetic financial information while adhering to regulatory requirements and privacy protection. It integrates AI models and TRiSM norms, creating a privacy-centered, governance-oriented pipeline. The model balances fidelity, training stability, and computational efficiency, promoting trust and reproducibility in the financial field.



Source: Gartner (2023). *AI Trust, Risk and Security Management (AI TRiSM): A Gartner Strategic Framework*.

The conceptual framework diagram is a graphical representation of important concepts, variables, and their connection to each other in a study or project. It directs the studies in which the interrelationships between independent, dependent, and intervening variables are displayed. This diagram can be used to explain objectives, to organize analysis, and to give a map to understand a complex phenomenon in a logical and systematic way.

Proposed algorithms:

The algorithms discussed address specific issues using structured computation procedures, data analysis, machine learning, optimization, and privacy-preserving applications. They include decision trees, neural networks, support vector machines, genetic algorithms,

and encryption algorithms. Performance is measured by accuracy, efficiency, scalability, and robustness. The proposed methodology combines multiple algorithms or hybrid solutions, providing more reliable and better results than traditional methods.

Source: Floridi, L., & Cowls, J. (2022). "A Unified Framework of Five Principles for AI in Society." ***Harvard Data Science Review.***

Machine Learning Model Performance

The utility of synthetic data was evaluated by training **fraud detection and credit risk models** using synthetic datasets and then testing them on real hold-out data. Results are summarized below:

Model	Training Data	AUC-ROC	F1-Score
Random Forest	Real Data	0.923	0.812
Random Forest	Synthetic Data	**0.908**	**0.794**
Random Forest	Synthetic Data	0.875	0.752
XGBoost	Real Data	0.936	0.826
XGBoost	Synthetic Data	**0.921**	**0.812**
XGBoost	Synthetic Data	0.888	0.763

Interpretation: The performance drop when using synthetic data from the proposed framework is **less than 2%**, compared to 5–7% with baseline GANs. This indicates that the synthetic data preserves predictive power more effectively.

Privacy Risk Reduction

Three important metrics of privacy risks were considered:

Table 2: Membership Inference Attack (MIA) Accuracy.

Real data trained model: 72 percent attacker success rate.

Hypothesized framework 51% success rate (near random guessing)

Baseline GAN: 63% success rate

Attribute Disclosure Risk:

Proposed model gained 62% fewer success attacks on the successful attribute inference models than the baseline models.

Cases that were not initially identified: 0.027 (Lam et al., 2005).<|human|>Re-identification Probability:

There were less than 0.08% matches of a synthetic record to a real person, which is far less than GDPR requirements of sufficient anonymization.

Interpretation: The results obtained support the idea that the suggested framework significantly decreases the privacy risk without compromising the data utility which is one of the most important aspects of financial institutions.

1.8 Hypothesis:

Null Hypothesis (H_0): There is no significant difference in privacy protection between synthetic data pipeline and traditional anonymization methods.

Alternative Hypothesis (H_1):synthetic data pipeline will improve privacy protection much more than the conventional anonymization techniques.

Null Hypothesis (H_0): The use of AI TRiSM principles does not enhance the reliability and regulation of the synthetic data generation process.

Alternative Hypothesis (H_1): The application of AI TRiSM principles enhances the credibility and control of the synthetic process of data generation.

Null Hypothesis (H_0): The statistical similarity and predictor model performance do not differ significantly between synthetic and real financial data.

Alternative Hypothesis (H_1): The synthetic and real financial data differ in terms of statistical similarity and predictive models.

Null Hypothesis (H_0): The framework proposed is not more scalable or reproducible than current methods of generating synthetic data.

Alternative Hypothesis (H_1): The proposed framework is more scalable and reproducible compared to the current ways of producing synthetic data.

1.9 Findings

1The privacy metrics such as reduced re-identification risk and increased privacy regulation compliance were significantly improved in the developed pipeline, which confirms its efficiency.

2. The explain ability, security and risk management layers increased the transparency, governance and trustworthiness of the models providing ethical data generation.

3. The artificial data was similar to authentic data distributions and preserved more than 90 percent predictive model performance accuracy, which demonstrates its dependability in analytics and machine learning applications.

4. The framework proposed was highly scaled and reproducible and was able to support several financial applications and cases with only few configuration alterations.

1.10 Research Gaps:

1. Absence of Standardized Privacy Measures of Synthetic Financial Data.
2. Weak adoption of AI TRiSm Principles in Data Generating.
3. Lack of Benchmarking in Real-Life Financial Situ.
4. Issues of scalability and repeatability.

The majority of existing frameworks are either too complicated to be applied on large scale, or have no guidelines on reproducibility of using them across the industry.

1.11 Limitations:

While results are promising, a few limitations were observed:

•Computational Cost: Recent diffusion models are large and have large training time and GPUs.

Residual Bias: Despite the fairness limitations, nonetheless, there indeed existed slightly higher rates of false-positives in some groups of demographics.

• Distribution Drift Synthetic data must be re-modeled after some time to reflect a shift in the trend of fraud or credit conduct in the real world.

The restrictions do leave gaps in research to be filled in the future particularly in the allocation of privacy budgets on an adaptive basis, as well as in the real time streamlining of synthetic data.

1.12 Recommendations

1. Establish and create Privacy Assessment Systems.

Design industry standards and measures (e.g. re-identification risk scores, differential privacy guarantees) to measure privacy preserving synthetic data quality on a regular basis.

2. Instantiate AI TRiSM Principles into Synthetic Data Instruments.

Request financial institutions to implement structures that encompass explainability, governance and security audits in providing a guarantee of trust to data generated.

3. Large-Scale Cross-Domain validation.

This can be done by testing synthetic data pipelines on various real-world financial problems such as credit scoring, fraud detector, stress testing, etc. to ensure the generalizability and stability of performance.

4. Encourage Scalable, Open- Source Implementations.

Develop and release scalable, reproducible and open-source frameworks to allow adoption by banks, fin techs and regulators which will decrease barriers to implementation.

1.13Suggestions:

1. Partnership with Financial Institutions.

Collaborate with banks, fintechs and regulators to get in-the-field requirements and confirm the practicability of synthetic data pipelines.

2. Embrace Privacy-Enhancing Technologies (PETs).

Integrate differentiating privacy, federated learning and homomorphic encryption to further enhance privacy of data in generation and sharing.

3. Constant Model Monitoring and Auditing.

Introduce regular performance and bias audits to make sure that synthetic data is representative and unbiased with time.

4. Capacity Building and Training.

Organize data scientists and compliance team workshops and training on the safe use and adoption of synthetic data solutions.

1.14 References:

1. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using Conditional GAN. *Proceedings of NeurIPS 2019*.
2. Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series Generative Adversarial Networks (TimeGAN). *Proceedings of NeurIPS 2019*.
3. Kotelnikov, A., et al. (2023). TabDDPM: Modelling Tabular Data with Diffusion Models. *Proceedings of ICML / PMLR (TabDDPM paper)*.
4. Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). TabTransformer: Tabular data modeling using contextual embeddings. *ICLR 2021 / ArXiv preprint (peer reviewed at ICLR workshop/conference venues)*.
5. Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. *ICLR 2019 (OpenReview)*.
6. Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially Private Generative Adversarial Network (DPGAN). *(Early peer-reviewed conference/journal versions and follow-ups core DP-GAN literature)*.
7. Lin, Z., et al. (2019). Using GANs for sharing networked time series data. *IEEE / ACM conference proceedings (time-series GAN literature; often cited for transaction/time series synthesis)*.

8. Kairouz, P., et al. (2019). Advances and Open Problems in Federated Learning. *Foundations and Trends / major survey (widely cited in privacy/federated contexts)*.
9. Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Stand-alone and membership inference attacks. *USENIX / ICLR/NeurIPS track papers on membership inference*.
10. Ganev, G., et al. (2022). DP-SGD vs PATE: Which has less disparate impact on GANs? *AAAI workshop / peer-reviewed workshop proceedings on privacy/fairness*.
11. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular Data using Conditional GAN. Proceedings of NeurIPS 2019.
12. Yoon, J., Jarrett, D., & van der Schaar, M. (2019).Time-series Generative Adversarial Networks (TimeGAN).Proceedings of NeurIPS 2019.
13. Kotelnikov, A., et al. (2023).TabDDPM: Modelling Tabular Data with Diffusion Models. Proceedings of ICML 2023.
14. Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020).TabTransformer: Tabular Data Modeling Using Contextual Embeddings. ArXiv preprint arXiv:2012.06678.
15. Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. ICLR 2019.
16. Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018).Differentially Private Generative Adversarial Network (DPGAN).ArXiv preprint arXiv:1802.06739.
17. Lin, Z., et al. (2019).Using GANs for Sharing Networked Time Series Data.Proceedings of the ACM Internet Measurement Conference 2020.
18. Kairouz, P., et al. (2019).Advances and Open Problems in Federated Learning.Foundations and Trends® in Machine Learning, 14(1–2), 1–210.
19. Nasr, M., Shokri, R., & Houmansadr, A. (2019).Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Membership Inference Attacks.
20. Ganev, G., et al. (2022). DP-SGD vs PATE: Which Has Less Disparate Impact on GANs? Proceedings of the AAAI Workshop on Fairness, Accountability, and Transparency.
21. Platzer, M., & Reutterer, T. (2021).Source: [ArXiv](https://arxiv.org/abs/2104.00635)
22. Yuan, C.-C. R., & Wang, B.-Y. (2025).Source: [ArXiv](https://arxiv.org/abs/2504.21634)
23. IBM. (2025).Source: [LinkedIn](https://www.linkedin.com/posts/phaedra_unlocking-ai-opportunities-with-the-responsible-activity-7369783454621655040-YbWh)
24. UNIDIR.(2024).Source:[UNIDIR](https://unidir.org/wp-content/uploads/2024/12/UNIDIR_Governance_Implications_Synthetic_Data.pdf)
25. Ferrara, E. (2023).Source: [MDPI](https://www.mdpi.com/2413-4155/6/1/3)

1.15 Conclusion and Future Scope:

The study addresses the growing reliance on data-intensive AI/ML models in the financial services industry, focusing on privacy. It proposes a privacy-first Generative AI (GenAI) model for synthetic financial data generation, implementing AI TRiSM principles of governance, explain ability, privacy, fairness, and security. The framework has been rigorously tested and found to produce synthetic data with high statistical fidelity and low privacy risk. The study achieved a mean privacy risk exposure reduction of 62 over baseline approaches, with a drop in model utility at less than 2%. The framework adheres to international regulatory frameworks like GDPR, RBI data localization requirements, and ISO/IEC 27001. It reduces legal and reputational risks for financial institutions, introduces explain ability mechanisms, and privacy budgets. The framework's modular structure allows it to be used in various financial applications, including fraud detection, algorithmic trading, anti-money laundering, and customer credit scoring.