

Synthetic Financial Data Generation Using Generative AI: Privacy, Security and Model Reliability in Modern Banking

M Ravi Kumar

Research Scholar
Department of Management
Osmania University, Hyderabad

Abstract:

The rapid digital transformation of the banking sector has created an increasing demand for large, diverse and high-quality financial datasets for artificial intelligence, fraud detection, credit-risk assessment and financial modelling. However, the use of real customer data creates significant privacy, security and regulatory challenges. Synthetic financial data generated using Generative Artificial Intelligence (GenAI) provides an emerging solution by producing artificial datasets that preserve important statistical characteristics of real financial data without directly exposing individual customer records. This paper examines the role of Generative AI in synthetic financial data generation, with particular emphasis on privacy protection, security, data utility and model reliability in modern banking. Generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models and Large Language Models (LLMs) can generate synthetic transaction and customer datasets for model development and testing.

However, synthetic data does not automatically guarantee privacy, as models may reproduce sensitive information or create disclosure risks. Therefore, privacy evaluation, data-quality assessment, bias detection and model validation are essential. The paper proposes a framework combining data generation, privacy assessment, security controls, utility evaluation and human governance. The present research states that synthetic financial data can support responsible AI adoption in banking when supported by strong privacy, security, validation and regulatory governance mechanisms.

Keywords: Generative AI, Synthetic Data, Banking, Financial Data, Privacy, Cybersecurity, GAN, Data Security, Model Reliability, Artificial Intelligence.

Introduction

Modern banking increasingly depends on data for fraud detection, credit assessment, customer analytics, risk management and regulatory compliance. However, financial datasets contain highly sensitive information such as account details, transaction histories, customer identities and financial behaviour. Sharing or using such data for research and AI development can therefore create privacy and security concerns.

Synthetic data offers an alternative approach. Instead of directly using real customer records, AI models can generate artificial observations that reproduce important characteristics of the original dataset. The Financial Conduct Authority identified synthetic data as a potential privacy-enhancing technology for financial services, particularly for data sharing, fraud controls, model testing and bias mitigation.

Generative AI has expanded the possibilities for synthetic data generation. GANs, VAEs, diffusion models and LLM-based approaches can generate increasingly realistic financial records.

For Indian banks, this issue is particularly relevant because financial institutions are undergoing rapid digital transformation while handling large volumes of sensitive customer information. RBI's regulatory work has emphasized data governance, privacy, security, algorithmic transparency and responsible use of technology in financial services.

RBI's Working Group on Digital Lending, for example, highlighted concerns surrounding collection and use of sensitive financial and personal information and recommended transparency and explainability for algorithmic systems.

Objectives of the Study:

- To examine the role of Generative AI in synthetic financial data generation.
- To analyse privacy and security benefits of synthetic financial data in banking.
- To examine the reliability and utility of AI-generated financial datasets.
- To identify risks associated with synthetic-data generation.
- To examine the Indian regulatory perspective concerning financial data privacy and responsible AI.

Research Methodology

The study follows a conceptual and exploratory research methodology based on secondary data.

The analysis focuses on four dimensions:

Privacy → Security → Data Utility → Model Reliability

Generative AI and Synthetic Financial Data

Synthetic financial data refers to artificially generated information designed to reproduce relevant characteristics of real financial datasets.

For example, a real banking dataset may contain:

Customer ID → Account Type → Transaction Amount → Date → Merchant → Location → Transaction Type

A generative model can learn statistical relationships between these variables and create new artificial records.

The resulting dataset should ideally preserve: statistical distributions, correlations, transaction patterns, customer segments, fraud characteristics.

Generative models increasingly provide sophisticated approaches to this problem. Research on financial synthetic-data generation has evaluated GANs, diffusion models and variational approaches according to fidelity, privacy, efficiency and downstream usefulness.

Technologies for Synthetic Financial Data

Generative Adversarial Networks: GANs consist of two competing models:

- Generator → creates synthetic data
- Discriminator → evaluates whether data resembles real data

Through repeated training, the generator attempts to produce increasingly realistic observations.

GANs can be used for transaction-data generation, fraud-data augmentation, credit-risk modelling and financial simulation.

Variational Autoencoders: VAEs learn a probabilistic representation of financial data and generate new observations from the learned latent space.

They can be useful where banks need synthetic datasets with controlled statistical characteristics.

Diffusion Models:

Diffusion models generate data through an iterative process of adding and removing noise. They have emerged as another approach for high-quality synthetic financial data generation.

Large Language Models:

LLMs can generate structured or semi-structured financial information and can assist with synthetic textual datasets such as: customer-service conversations, financial reports, compliance documents, investigation records.

Privacy Protection:

Privacy is one of the strongest motivations for synthetic financial data. Synthetic datasets can reduce the need to distribute actual customer records to researchers, technology providers or model-development teams. However, synthetic data should not automatically be considered completely anonymous.

A poorly designed generative model could memorize parts of the training dataset. This creates potential risks such as membership inference, data reconstruction, model inversion, attribute disclosure.

Therefore, privacy testing must be performed before synthetic data is released or used externally.

Model Reliability

The usefulness of synthetic data depends on whether the generated dataset accurately represents the characteristics of real financial data. Three major dimensions should be evaluated:

- Fidelity- Does the synthetic dataset resemble the statistical properties of real data?
- Privacy- Does the synthetic dataset prevent disclosure of individual records?
- Utility- Can models trained using synthetic data achieve useful performance when applied to real-world problems?

Indian Banking Perspective

India's banking sector is rapidly adopting AI, cloud computing, digital payments and advanced analytics. This creates substantial opportunities for synthetic data.

Indian banks could use synthetic financial datasets for fraud-detection model development, credit-risk modelling, stress testing, cybersecurity testing, employee training, software testing, financial research, AI model validation.

At the same time, financial data requires strong governance. RBI's work on digital lending has highlighted concerns around sensitive personal and financial information, data security, algorithmic transparency and explainability. The RBI's broader technology and data-governance framework also emphasizes the importance of data quality, information architecture and controlled use of financial information. Thus, synthetic data can potentially help Indian banks balance AI innovation with customer privacy.

Advantages

Synthetic financial data can provide several advantages:

- Improved privacy by reducing direct exposure of customer records.
- Greater data availability for AI research.
- Reduced data-sharing barriers between financial institutions and technology partners.
- Fraud-data augmentation where genuine fraud cases are limited.
- Model testing without exposing production customer data.
- Faster AI development through accessible datasets.
- Regulatory testing in controlled environments.

Challenges:

Despite its potential, synthetic financial data faces several challenges:

Insufficient realism,

Privacy leakage,

Model bias,

Rare-event representation,

Difficulty reproducing complex transaction networks,

Model drift,

Computational cost,

Regulatory uncertainty,

Lack of common evaluation standards.

The evidence also shows that current generative approaches may struggle to reproduce complex financial structures perfectly. Therefore, banks should not assume that synthetic datasets are automatically equivalent to real-world datasets.

Conclusion

Generative AI-based synthetic financial data generation represents an important emerging opportunity for modern banking. It can provide financial institutions with realistic datasets for AI development, fraud detection, risk modelling, software testing and research while reducing direct exposure to sensitive customer information.

However, synthetic data is not inherently private or reliable. Its effectiveness depends on the quality of the generative model, privacy protection mechanisms, validation procedures and governance framework. Banks should therefore evaluate synthetic data using a combination of privacy, security, fidelity, utility and fairness measures.

For Indian banking institutions, synthetic financial data could become an important component of responsible AI adoption. Its successful implementation requires alignment between technological innovation, RBI expectations, data-protection requirements, cybersecurity controls and human oversight. The future banking environment is therefore likely to move toward a model in which real data is protected, synthetic data supports innovation, and AI systems are continuously validated and governed.

REFERENCES:

1. Naidu, A. (2024). Privacy-preserving synthetic banking data generation using generative adversarial networks. *Journal of Computational Analysis and Applications*, 32(2), 787–798.
2. Karst, F. S., Chong, S.-Y., Antenor, A. A., Lin, E., Li, M. M., & Leimeister, J. M. (2024). Generative AI for banks: Benchmarks and algorithms for synthetic financial transaction data. SSRN.
3. Sun, Y., et al. (2024). Reinforcement prompting for financial synthetic data generation. *The Journal of Finance and Data Science*, 10, 100137. <https://doi.org/10.1016/j.jfds.2024.100137>
4. Financial Conduct Authority. (2024). Report: Using synthetic data in financial services. Financial Conduct Authority.
5. Reserve Bank of India. (2021). Report of the Working Group on digital lending including lending through online platforms and mobile apps. Reserve Bank of India.
6. Reserve Bank of India. (2010). Report on information management in the Reserve Bank of India. Reserve Bank of India.
7. Pires, O. M., Nooblath, M. Q., Silva, Y. A. C., da Silva, M. H. F., Galvão, L. Q., & Albino, A. S. (2024). Synthetic data generation with hybrid quantum-classical models for the financial sector. *The European Physical Journal B*, 97, 178.

8. European Commission, Joint Research Centre. (2024). Digital financial services: Synthetic data ensures compliance with confidentiality requirements. European Commission.