International Journal of Management Technology
Vol.12, No 1, pp.67-76, 2025
Print ISSN: ISSN 2055-0847(Print)
Online ISSN: ISSN 2055-0855(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Synthetic Data for Payment Systems: AI-Powered Privacy-Preserving Testing

Prajwalkumar B. Bhatkar
Senior Lead Software Engineer at a Fortune 500 Bank, Richmond, Virginia, United States
Email - prajwal.bhatkar@gmail.com

doi: https://doi.org/10.37745/ijmt.2013/vol12n16776 Published February 23, 2025

Citation: Bhatkar P. B. (2025) Synthetic Data for Payment Systems: AI-Powered Privacy-Preserving Testing, International Journal of Management Technology, Vol.12, No 1, pp.67-76

Abstract: *In modern banking, ensuring that new payment systems operate accurately and securely requires extensive testing. However, testing with real-world data introduces privacy risks, and synthetic data offers a promising alternative. This paper explores the potential of Generative AI for producing realistic, privacy-compliant synthetic transaction data. The proposed approach addresses challenges such as data privacy, diverse dataset creation, and the ability to simulate rare or edge-case scenarios—thus enhancing the robustness of payment systems.*

Keywords: Synthetic data, payment systems, privacy, machine learning, generative AI

INTRODUCTION

Testing financial systems poses unique challenges because of the sensitive nature of transactional data. Ensuring that payment systems are secure, reliable, and efficient requires diverse datasets for simulation and testing. However, using real data raises significant privacy concerns and legal implications. This paper discusses how synthetic data generation can revolutionize the testing process by providing realistic, privacy-compliant data.

Recent advances in AI and machine learning have transformed various applications in the financial sector—from fraud detection and risk assessment to personalized banking services. As these systems grow more sophisticated, the need for comprehensive testing methodologies becomes paramount. Traditional methods relying on real or manually created data are limited by:

- **Privacy Risks:** Real data use can expose customer information and violate regulations (e.g., GDPR, CCPA).

- **Limited Data Diversity:** Manually generated datasets often fail to capture the complexity of real-world transactions.
- **Lack of Edge Case Representation:** Rare scenarios (e.g., fraud attempts or system outages) may be underrepresented, reducing system robustness.

Additionally, compliance with frameworks such as Basel III necessitates rigorous testing under conditions that real data may not fully provide.

LITERATURE/THEORETICAL UNDERPINNING

Recent developments in Generative AI have paved the way for generating synthetic data that is both realistic and privacy-preserving. Generative models learn from anonymized historical transaction data to produce datasets that mimic real patterns without exposing sensitive information.

- **Generative Adversarial Networks (GANs):**
  Utilize a generator and a discriminator in a competitive setting, yielding highly realistic synthetic data.
- **Variational Autoencoders (VAEs):**
  Learn compressed representations of data and generate new samples from a latent space.
- **Transformer-based Models:**
  Recent models (e.g., GPT-3) have demonstrated potential in generating structured financial transaction data.
- **Federated Learning:**
  This approach trains models on distributed datasets, addressing privacy concerns by keeping data localized.

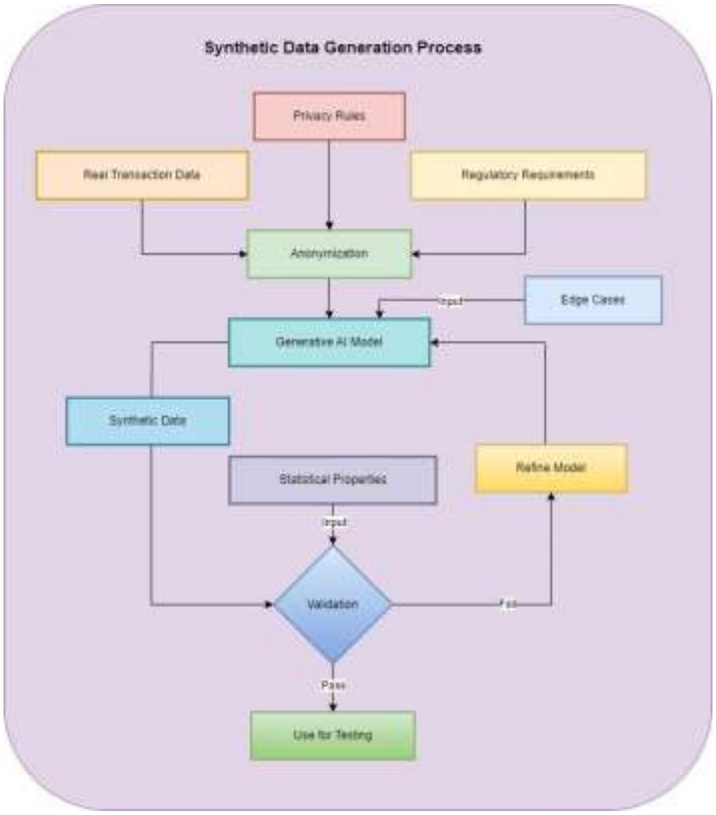

Diagram 1: Diagram showing Synthetic Data Generation Process

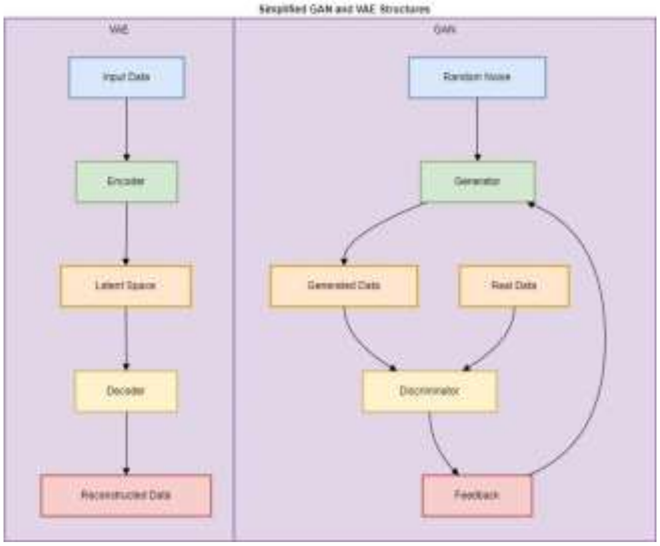

Diagram 2: Simplified GAN and VAE structure

METHODOLOGY

To address privacy concerns further, a federated learning approach can be employed where models are trained on distributed datasets without centralizing the data.

3.1 Producing Realistic, Privacy-Compliant Synthetic Transaction Data

- **Data Characteristics:**
  The synthetic data mimics real transaction patterns, reflecting features such as transaction frequency, amounts, merchant types, seasonal trends, and demographic segmentation.
- **Privacy Compliance:**
  Synthetic data is designed to be statistically similar to real data while avoiding the reproduction of any individual's transactional history. The generation process adheres to regulations such as GDPR and CCPA, allowing for risk-free testing.
- **Implementation Techniques:**
  Models (GANs, VAEs) are trained on anonymized data and integrated with privacy-preserving methods (e.g., differential privacy) to maintain data utility while protecting privacy.
- **Data Validation Techniques:**
  Validation is performed using statistical tests (e.g., Kolmogorov-Smirnov test) and by comparing machine learning model performance when trained on both real and synthetic datasets.

Creating Diverse Datasets for Edge Cases and Rare Scenarios

- **Scenario Generation:**
  Synthetic data can simulate specific conditions such as fraud, stress testing, or other rare events.
- **Edge Case Identification:**
  Anomaly detection techniques applied to real data help identify rare scenarios, which are then amplified in the synthetic dataset.
- **Customization and Adversarial Examples:**
  Testers can define parameters for the data generation process, combining multiple rare scenarios or incorporating adversarial examples to robustly challenge payment systems.

RESULTS/FINDINGS

Benefits of Synthetic Data for Testing

- **Enhanced Testing Coverage:**
  Synthetic data exposes systems to a wider range of scenarios, including rare events.
- **Improved Fraud Detection:**
  Amplified fraud patterns in synthetic data help in training systems to detect and prevent fraudulent activities.
- **Support for Continuous Integration:**
  On-demand synthetic data generation enables iterative testing in continuous integration pipelines.
- **Regulatory Compliance:**
  Testing environments remain compliant with privacy regulations since no real customer data is used.
- **Cost Reduction:**
  Reducing reliance on complex data anonymization and mitigating data breach risks results in lower costs.

Performance Benchmarks Using Distributional Similarity Analysis

Kolmogorov-Smirnov (K-S) Test:
Measures the maximum difference between the cumulative distribution functions (CDFs) of the real and synthetic datasets.

Table 1.The table shows Kolmogorov-Smirnov test result for sample dataset

Feature	**K-S Statistic (0-1)**
Transaction Amount	0.035
Transaction Time	0.012
Account Balance	0.047

A lower K-S statistic indicates a closer match between the two distributions.

Jensen-Shannon Divergence (JSD):
Quantifies the similarity between two probability distributions.

Table 2.The table shows Jensen-Shannon Divergence test result for sample dataset

Feature	**JSD (0-1)**
Transaction Amount	0.0054
Transaction Time	0.0021
Merchant Category	0.0148

These benchmarks demonstrate that the synthetic data closely mirrors the statistical properties of real data, with only minor deviations.

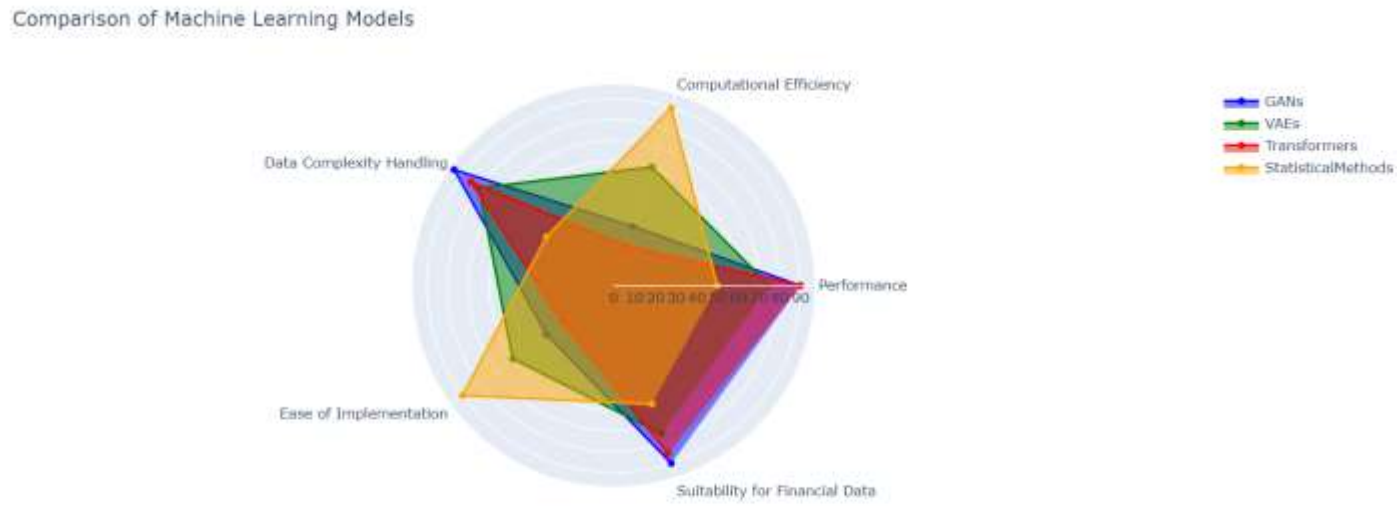

Diagram 3: Comparison of Machine Learning Models

DISCUSSION

The findings confirm that synthetic data can effectively mimic the statistical properties of real transactional data while ensuring privacy. However, several implementation challenges were identified:

- **Computational Resources:**
  High-quality data generation, especially using GANs, requires significant computational power. Cloud-based scaling and distributed training methods help mitigate this issue.
- **Data Quality and Consistency:**
  Maintaining the statistical integrity of synthetic data is complex. Rigorous validation processes and ensemble modeling techniques are necessary to ensure quality.
- **Privacy Guarantees:**
  Balancing data utility with privacy is critical. Incorporating differential privacy with carefully calibrated budgets ensures robust privacy protection.
- **Model Stability and Mode Collapse:**
  Training challenges such as instability and mode collapse in GANs can be addressed using techniques like spectral normalization, gradient penalty, or alternative architectures (e.g., Wasserstein GANs).
- **Regulatory Compliance:**
  Ensuring adherence to diverse regulatory requirements demands early engagement with authorities, detailed documentation, and flexible data pipelines.
- **User Acceptance and Trust:**
  Gaining stakeholder confidence in synthetic data requires clear comparative analyses with real data and gradual integration into testing frameworks.
- **Adapting to Evolving Transaction Patterns:**
  Continuous learning and periodic retraining ensure that synthetic data remains reflective of current transaction trends.
- **Integration with Existing Testing Frameworks:**
  Developing modular tools and robust APIs facilitates the seamless incorporation of synthetic data into current workflows.

Overall, while synthetic data shows great promise, slight deviations in distribution suggest that supplemental testing with limited real data may still be valuable for critical applications.

Implications to Research and Practice

Real-Life Case Studies

- **JPMorgan Chase’s Synthetic Data Program:**
  This program generates millions of realistic customer records, accelerating innovation while preserving privacy.

- **UK Financial Conduct Authority (FCA) TechSprint:**
  The TechSprint event demonstrated practical solutions for synthetic data generation in the financial sector, emphasizing regulatory acceptance.
- **Mastercard's Synthetic Fraud Detection Dataset:**
  By generating datasets that include both legitimate and fraudulent activities, Mastercard has improved fraud detection capabilities without exposing sensitive data.

Broader Implications

Synthetic data supports regulatory compliance, reduces testing costs, and drives innovation in payment systems. The successful implementation of synthetic data generation models encourages broader adoption in areas such as credit scoring, customer relationship management, and anti-money laundering systems.

CONCLUSION

Synthetic data generation offers a robust, privacy-preserving alternative for testing payment systems. The approach enhances testing coverage, improves fraud detection, supports continuous integration, and ensures compliance with privacy regulations—ultimately driving innovation while protecting customer data.

Future Research

Future work could explore:

- Extending synthetic data applications to other financial domains such as credit scoring, customer relationship management, or anti-money laundering.
- Comparative evaluations of different generative models and privacy-preserving techniques in various real-world scenarios.
- Integration of blockchain technology with synthetic data generation to provide an immutable, transparent record of the data generation process, thereby increasing trust and security.

REFERENCES

Dataset from IEEE-CIS fraud detection. Retrieved from https://www.kaggle.com/competitions/ieee-fraud-detection/data

Aziz, S., & Dowling, M. (2019). Machine Learning and AI for Risk Management. In Disrupting Finance. Palgrave Pivot, Cham, 33-50.

Basel Committee on Banking Supervision. (2017). Basel III: Finalising post-crisis reforms. Bank for International Settlements.

Goodfellow, I., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv:1312.6114.

Brown, T. B., et al. (2020). Language models are few-shot learners. arXiv:2005.14165.

McMahan, H. B., et al. (2017). Communication-efficient learning of deep networks from decentralized data. arXiv:1602.05629.

Massey Jr., F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46(253), 68-78.

Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. arXiv:1706.02633.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv:1412.6572.

Regulation (EU) 2016/679 (General Data Protection Regulation). (2016). Official Journal of the European Union.

Hittmeir, M., et al. (2019). Utility-preserving data synthesis for privacy-preserving data sharing. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 1293-1302.

JPMorgan Chase. (2021). Annual Report 2020.

Financial Conduct Authority. (2019). Global AML and Financial Crime TechSprint.

Di Castro, D., et al. (2019). Synthesizing labeled financial time series data: A case study with credit card fraud. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 1492-1499.

Zheng, X., et al. (2018). Blockchain-based secure data sharing system for the Internet of Vehicles. IEEE Internet of Things Journal, 6(3), 4671-4679.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

Abadi, M., et al. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

Isola, P., et al. (2017). Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5967-5976.

Mirsky, Y., et al. (2019). Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. NDSS Symposium 2019.

Cresci, S., et al. (2015). Fame for sale: Efficient detection of fake Twitter followers. Decision Support Systems, 80, 56-71.

Tak, J., & Song, Y. S. (2021). Generating Time-Series Data with Stable GAN. arXiv:2106.01389.

Raschka, S. (2020). Machine Learning and AI-Based Approaches for Synthetic Data Generation. ACM Computing Surveys, 53(3), 1-35.

Van Rossum, G., & Drake, F. L. (2009). The Python Language Reference Manual. Network Theory Ltd.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv:1802.05957.

Salimans, T., et al. (2016). Improved techniques for training GANs. Advances in Neural Information Processing Systems, 29, 2234-2242.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196.