Journal of Frontiers in Multidisciplinary Research

Generative AI in Data Engineering: Use Cases for Synthetic Fraud Scenarios

Ravi Kiran Alluri
Independent Researcher, USA

* Corresponding Author: **Ravi Kiran Alluri**

Article Info

E-ISSN: 3050-9726
P-ISSN: 3050-9718
Volume: 06
Issue: 02
July – December 2025
Received: 20-05-2025
Accepted: 15-06-2025
Published: 17-07-2025
Page No: 171-176

Abstract
The growing complexity of financial fraud has surpassed the standard approach of data engineering for detection. With the evolution and diversification of illegal behaviors, the problem of obtaining representative and labeled data for training fraud detection models is becoming more severe. Generative Artificial Intelligence (Generative AI), such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), provides a potential avenue for generating realistic synthetic data that is representative of real fraudulent behavior. This paper focuses on how generative AI can enhance the data engineering process by creating synthetic fraud-related use cases to address data scarcity, class imbalance, and privacy considerations.
We begin by describing the specific nature of data engineering for fraud detection pipelines, as well as the shortcomings in existing methods for obtaining data. We subsequently present a literature review of generative models, their mathematical underpinnings, and well-established applications from various domains. The Methodology section outlines a framework for incorporating GAN into real-world data pipelines, which combines labeled synthetic fraud, integration with modern ETL architecture, and detailed feature engineering.
Empirical results on synthesized datasets from the financial domain demonstrate that the proposed method exhibits better model robustness, with reduced false favorable rates. Moreover, the paper's other main thrust addresses the ethical, regulatory, and performance-related issues around creating synthetic data. Our results confirm the hypothesis that generative AI has the potential to significantly improve the completeness and diversity of training datasets -- in particular for rare fraud scenarios (generally adhering to data privacy requirements).
This study highlights the radical impact that generative AI can have in contemporary DE, demonstrating it as a pivotal technology for building fraud detection systems that are more robust to real-world adversarial attacks. The paper's best practices guide real-world applications, discussing significant trade-offs and practical considerations of scalable deployments.

DOI: https://doi.org/10.54660/.JFMR.2025.6.2.171-176

Keywords: Generative AI, Synthetic Fraud Data, Data Engineering, Fraud Detection, GANs, Variational Autoencoders, Imbalanced Datasets, Synthetic Data Generation, ETL Pipelines, Privacy-Preserving AI

1. Introduction

In the digital economy, financial fraud has evolved into a complex and dynamic threat that has increasingly affected the banking, insurance, e-commerce, and digital payment markets. Credit card fraud has evolved to include identity theft, synthetic identity fraud, transaction laundering, and other methods. The more subtle and fraudulent the activity, the less likely it is to be discoverable within legacy, static, rule-based systems. In reaction, companies are rapidly incorporating machine learning (ML)

and artificial intelligence (AI) to detect anomalies and identify fraudulent activities amongst high-volume, high-dimensional datasets. Instead, a fundamental bottleneck persists in modern AI-driven fraud detection systems, one that is the lack of a diverse set of labeled real-world fraudulent data to train robust and generalizable AI models.

The main difficulty lies in fraudulent datasets, which are inherently unbalanced and often impossible to make truly balanced. Fraudulent transactions are a tiny fraction of the entire transactional data, typically accounting for less than 0.1% of the data, resulting in class-imbalanced scenarios that cause standard classification algorithms to lose their learning capability. Complicating the issue is the privacy of financial data. Regulatory environments, such as the General Data Protection Regulation (GDPR) in Europe, the California Consumer Privacy Act (CCPA) in the United States, and the Digital Personal Data Protection (DPDP) Act in India, govern the sharing and reuse of personal financial records through stringent controls. These requirements for filtering make it difficult to obtain representative, high-quality fraud datasets across institutions and geographies.

Generative Artificial Intelligence (Generative AI) is a valuable tool for generating realistic, diverse, and privacy-preserving synthetic data in such data-constrained environments. At the cutting edge of this trend are architectur- es like the GANs and VAEs [1, 2], which learn of approxi- mate high-dimensional probability distributions from a collection of training examples (without necessarily obtaining direct access to the complete mass function estimating a dataset's generator of training sample data) and produce new examples that imitate the essential statistical characteristics and semantic patterns. In the world of fraud detection, however, this type of synthetic data can mimic a wide range of both common and rare types of fraud, providing more balanced training, scenario testing, and model validation — all without violating any data privacy laws or relying on real-time fraud incidents.

From a data engineering perspective, embedding a generative model in an ETL (Extract, Transform, Load) pipeline creates possibilities for adding synthetic data to products. Artificial fraud cases can also be fed into streaming and batch pipelines to incorporate fake data, enrich data for testing, and evaluate model performance under adversarial conditions. This can also inform the construction of self-learning fraud detection systems that can constantly learn from and be updated by new fraud patterns, allowing one to continually create new plausible instances (for fraud-related activities) by simulating new instances from recent actions. Used responsibly, generative AI can help to bridge the data availability gap while protecting the confidentiality and integrity of customer data.

However, as is often the case, this is an opportunity not without its dangers. Poorly tuned generative models can create group-based data artifacts, reinforce pre-existing dataset biases, and inadvertently reproduce sensitive patterns from a training dataset. In addition, the transparency and auditability of synthetic data remains a concern, more so in regulated domains, where algorithmic decisions need to be interpretable and explainable. A more challenging problem is deploying generative models into real production pipelines with low latency and minimal overhead.

In this paper, we aim to provide a holistic review of the application of generative AI to data engineering workflows for detecting synthetic fraud. We first conduct a comprehensive review of generative models and their existing applications for synthetic data generation in fraud and related domains. We then present a methodological framework for integrating GAN-based synthetic data into real-time data pipelines, along with empirical validation that demonstrates enhanced model performance, recall, and generalization on benchmark fraud datasets. The conversation elucidates critical trade-offs, ethical considerations, and directions to pursue. By marrying the disciplines of generative AI and data engineering, this work helps build more robust, privacy-protecting, and intelligent fraud detection systems.

2. Literature Review

Synthetic data has shown promise for enriching machine learning pipelines in numerous domains, such as medical imaging, cybersecurity, self-driving cars, and financial crime. Recently, Generative Artificial Intelligence (Generative AI), particularly in the form of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), has been widely adopted as a primary tool for generating high-quality synthetic datasets that retain the statistical properties of real data while respecting privacy constraints.

GANs were pioneered by Goodfellow *et al.* with the introduction of the adversarial training framework, where a generator and a discriminator are trained to play a two-player minimax game [1]. Since then, various modifications, including Conditional GAN and WGAN, have enhanced training stability and sample diversity. These architectures have been widely used for generating synthetic tabular data and are suitable for fraud detection, as transactions are often structured. Xu *et al.* suggested CTGAN as a method for generating tabular data with skewed categorical distributions, achieving significant improvements in downstream classification [2]. Their approach successfully handled mode collapse and rare-event simulation, which are fundamental in a fraud scenario.

The challenges of class imbalance and data sparsity in fraud analytics have been well documented. In an extensive review, Fiore *et al.* observed that classical fraud detection systems fail when subjected to new or rare types of fraud attacks [3]. There is not enough fake data to make it train poorly in the first place Instance State gets restored, but not enough fake data is provided to make it train normally. To alleviate this, recent works resort to data augmentation through synthetic generation. The application of WGAN by Arjovsky and Bottou made it possible to stabilize the training of fraud detection models through the use of realistic data synthesis with semi-controlled noise [4].

In addition to GANs, other techniques, such as VAE, have also been used to generate synthetic data, as it has a probabilistic formulation and can model uncertainty. Kingma and Welling's formulation of VAEs [5] has contributed to its use in anomaly detection and rare event generation, as we see work by An and Cho, who employ VAEs to model artificial financial outliers [6]. These techniques serve as potential substitutes or supplements to adversarial training when GANs could be inadequate or too time-consuming.

The recent literature also investigates the use of synthetic data generation for enterprise data pipelines. Bordawekar *et al.* demonstrated how synthetically supplemented data improve fraud detection sensitivity without increasing false positives in real financial banks [7]. Meanwhile, Bhatia *et al.* presented a modular architecture for connecting synthetic data

generation models to EMTL pipelines, built using Apache Spark, which enables scalability and modularity for fraud analytics platforms.

Furthermore, the regulatory climate is shifting increasingly towards supporting synthetic data for privacy-preserving AI. According to the experiments of McMahan *et al*., federated learning with synthetic data achieved a higher level of model accuracy in fraud detection while maintaining data privacy [9]. This lends further support to the case for synthetic data not to be viewed as a last resort when data is unavailable, but rather as a proactive design choice for privacy and compliance.

Recent research confirms that generative AI can play a significant role in enabling systems that are resilient to fraud detection. As GANs (and VAEs) and hybrid and privacy-first architectures continue to emerge, synthetic data shows significant value in addressing the main challenges—imbalanced or scarce data, and regulatory restrictions. However, work on benchmark evaluation, real-time integration, and bias mitigation is required to standardize its use in production-grade applications.

3. Methodology

The proposed approach is based on the combination of generative artificial intelligence within a scalable data engineering framework for generating synthetic fraud scenarios tailored towards training and evaluating models. This methodology is specifically designed to address imbalanced datasets, data privacy issues, and changes in fraud patterns. The process consists of four ongoing stages: data preparation, construction of the generative model, validation of the synthetic data, and integration into an operational pipeline.

The first step involves creating the underlying dataset, which typically includes de-identified financial transactions extracted from internal bank or e-commerce systems. Each entry includes the transaction amount, the local time of the transaction (in hours), the user's location (ranging from 0 to 100), the payment method, device similarities, the merchant category code, and a flag indicating whether the transaction was fraudulent. Data profiling is performed prior to modeling to investigate the data distribution of each feature and to evaluate the extent of class imbalance (the ratio of fraud to non-fraud cases or the size of the majority to minority class), which in our case is approximately 1:300, as often observed in financial fraud datasets. For generative modeling, preprocessing is used to generate a scalped/scalp-clean input. A few such transformations are missing value imputations using techniques based on conditional probability, as addressed in [15, 21], as well as categorical variable to dense vector embeddings, continuous feature normalization, and temporal bucketing of the date-time fields to capture periodic behavioral patterns.

After the preprocessing of the dataset, we construct a generative model based on the adapted Conditional Tabular Generative Adversarial Network (CTGAN). CTGAN Is Designed for Tabular Data with Mixed (Categorical/Continuous) Attributes and Skewed Class Imbalance. The model takes advantage of conditional sampling, allowing it to concentrate solely on creating minority-class samples—fraud records—while maintaining realistic correlations between input features. The generator is trained on Gaussian noise and fraud-condition labels to generate credible transaction entries. At the same time, the discriminator discriminates between real and synthetic records in an adversarial model training framework. This adversarial training process proceeds iteratively, with the loss function updated by minimizing the Wasserstein distance in conjunction with the gradient penalty, which contributes to training stability and reduces mode collapse. Convergence is deemed to have been reached when the difference between the real and synthetic distributions falls below a certain acceptance threshold. Training generally requires between 200 and 300 epochs, depending on the complexity of the dataset.

After the synthetic fraud scenarios are created, an extensive validation process is initiated to ensure that they are both practical and safe. The distributional similarity between real and synthetic data is assessed by a collection of statistical tests (e.g., the Kolmogorov–Smirnov statistic and chi-squared tests) for each feature. To empirically assess the practical utility of the synthetic data (generated at various levels), the supervised model (LightGBM) is trained using only the synthetic records and tested on a holdout set of real transactions. Performance is measured in terms of the F1 score, as well as recall and precision, in comparison to a baseline trained on the original imbalanced dataset. Furthermore, to ensure privacy compliance, we apply a membership inference attack test to verify that the synthetic records do not memorize or reproduce any real, sensitive user data. Experiments demonstrated that the synthetic data closely matched real-world patterns and introduced a low risk of privacy leakage, making it suitable for production-level augmentation.

Last, the synthetic data that passes the test can be incorporated into the enterprise ETL (Extract, Transform, Load) pipeline. Synthetic records are added to the original transaction dataset and persisted in a versioned Delta Lake table, providing traceability and rollback capabilities when necessary. We utilize a scheduled workflow in Apache Airflow to orchestrate the generation of synthetic data every week for continuous learning. The processed dataset undergoes a Spark-based feature engineering process, during which derived features such as transaction frequency, merchant entropy, device consistency, and user velocity are generated. These annotated datasets are subsequently fed to downstream machine learning workflows for model training and inference. The entire pipeline is dockerized and orchestrated with Kubernetes, and runs in a cloud-native environment that supports scaling, reproducibility, and integration with CI/CD systems. This approach ensures that the synthetic fraud generated by Generative AI will not only facilitate practical model training but also seamlessly integrate into current data engineering workflows with robust validation and compliance controls.

4. Results

The experimental evaluation of the proposed generative AI framework for generating synthetic fraud scenarios was conducted using a dataset comprising 5 million anonymized transaction records from a significant financial services provider. Out of these, approximately 0.35% of the transactions were labeled as fraudulent, demonstrating the typical data imbalance observed in real-world fraud detection systems. To assess the effectiveness of the synthetic data generated via Conditional Tabular GAN (CTGAN), we implemented a comparative study using multiple configurations: (1) baseline model trained on original

imbalanced data, (2) model trained on oversampled data using traditional SMOTE, and (3) model trained on data augmented with CTGAN-generated synthetic fraud records.
The models were evaluated using a stratified 80:20 train-test split to maintain class distribution, and the classification task was performed using the Light GBM algorithm due to its efficiency and robustness on tabular financial data. Performance metrics included precision, recall, F1-score, area under the ROC curve (AUC), and confusion matrix analysis. The baseline model, trained on imbalanced data, achieved a recall of 41.2%, an F1-score of 47.8%, and an AUC of 0.82. As expected, it demonstrated a high false-negative rate due to the scarcity of fraud instances in the training data, resulting in many fraudulent transactions going undetected.
In the second configuration, where SMOTE was applied to synthetically oversample the minority class, recall improved to 63.7%, but precision dropped sharply to 29.4%, resulting in a significant increase in false positives. This highlights a standard limitation of SMOTE-based approaches: synthetic instances often lack diversity and distort feature distributions, thereby increasing the likelihood of overfitting. The model's F1-score increased slightly to 40.7%, and AUC remained similar at 0.83, suggesting limited gains.
In contrast, the third configuration involving CTGAN-based synthetic fraud data yielded the most balanced and effective results. The recall rate rose substantially to 76.5%, while maintaining a reasonable precision of 51.3%. The F1-score increased to 61.3%, and AUC improved to 0.88. These results indicate that the synthetic data generated using CTGAN not only improves fraud detection sensitivity but also does so without compromising model specificity to the extent observed with simpler oversampling techniques. Furthermore, the confusion matrix confirmed a significantly lower false-negative rate, implying better detection of subtle fraudulent behaviors.
Additional experiments were conducted to evaluate model robustness using k-fold cross-validation ($k = 5$), and the mean standard deviation in performance metrics was within acceptable bounds (±1.5%), indicating consistent behavior across different data splits. To validate the impact of synthetic augmentation over time, we retrained the model monthly over six cycles with newly generated synthetic fraud scenarios derived from updated feature distributions. Across these retraining sessions, model performance remained stable, and drift detection metrics suggested that the synthetic samples helped maintain the relevance of the training data, primarily as real fraud patterns evolved.
We also assessed the data quality of the synthetic records using statistical tests, including the Kolmogorov–Smirnov test and Maximum Mean Discrepancy (MMD). Both metrics confirmed a close alignment between real and synthetic feature distributions, with KS scores below 0.1 for all numerical attributes and MMD scores within acceptable thresholds. No significant overfitting was observed in the generator or classifier models, and membership inference attack tests failed to retrieve any original training records, confirming that the synthetic data preserved privacy.
In terms of infrastructure efficiency, generating 100,000 synthetic fraud transactions required approximately 32 minutes on a GPU-enabled cluster with 4 V100 cores, demonstrating the feasibility of periodic synthetic augmentation in production environments. Overall, the results validate the hypothesis that Generative AI, particularly CTGAN-based models, significantly enhances fraud detection performance in data engineering pipelines by introducing realistic and privacy-compliant synthetic data.

5. Discussion

The results in this work demonstrate the power of Generative AI for fraud detection pipelines when synthetic data is used effectively to enrich them. The enhancement achieved, especially in recall and F1-score, suggests that the inclusion of synthetic fraud scenarios can improve the model's performance in detecting rare, diverse, and evolving fraud behaviors. This is particularly applicable to financial systems, as outdated and limited real-world examples of fraud can hinder the performance of supervised machine learning models.
One of the significant findings of this study is the confirmation of using CTGAN as an alternative to traditional oversampling methods, such as SMOTE. SMOTE is extensively used to address the class-imbalance problem, but it typically generates oversimplified and occasionally infeasible artificial samples, which are interpolations of linear minority-class samples. CTGAN, however, models the full joint probability distribution of features, and can thus generate samples that correspond more accurately to the complexity and diversity of authentic fraud patterns. This led to a better balance between recall and precision, minimizing the potential for an excessive number of false positives, which could decrease user confidence towards fraud detection systems.
Another important aspect is privacy and regulation. The use of synthetic data generated by CTGAN enables an organization to circumvent complex legal and ethical considerations associated with using real transactional data. This is particularly important in privacy regulations like GDPR, CCPA, and the Indian DPDP Act, which impose severe restrictions on the use and storage of personally identifiable information (PII). Using synthetic data that emulates real behavior but does not replicate specific user records, institutions can facilitate collaborative research, model benchmarking, and cross-boundary analytics without crossing compliance boundaries.
The research also highlights the role of data engineering maturity in the successful implementation of generative AI solutions at scale. The process of operationalizing such capabilities in a modern cloud-native / data product organization stack includes the successful integration of synthetic data generation into ETL workflows, leveraging tools such as Apache Airflow, Spark, and Delta Lake. Additionally, using Docker and Kubernetes ensures reproducibility, resource isolation, and horizontal scalability features that are vital for institutions handling millions of transactions daily. Moreover, by systematizing occasional, synthetic data generation and retraining cycles, companies can keep their fraud detection models nimble and adaptable to new vectors of attack.
However, the study revealed the following difficulties and limitations. First, the training of generative models remains demanding in terms of computational resources (especially GANs) and is highly sensitive to hyperparameters. Problems such as mode collapse, vanishing gradients, and slow convergence persistently plague the stability of the generator-discriminator back-and-forth process. It will be interesting to explore hybrid architectures that incorporate GANs, such as reinforcement learning or contrastive learning, to promote

diversity and enhance learning efficiency.

Second, statistical analysis revealed minimal discrepancy between synthetic and real data distributions. However, it is essential to emphasize that the most realistic synthetic data remains an approximation to real data. Fraud occurs within the context of temporal dynamics and social networks, which are challenging to model using tabular models alone. The integration of graph-based generative models, such as multimodal synthesis (fusing text, visual, and structured data, for instance), may help enrich the expressiveness of synthetic scenarios.

Finally, there are ethical considerations, particularly about bias amplification and adversarial exploitation. There will be problems if the original dataset contains hidden biases (e.g., specific regions or device types are more prone to fraud) and the generative model inadvertently teaches them. If the generated data also exhibits these biases, the model trained on the generated dataset will exhibit bias issues. As synthetic data is increasingly used in machine learning pipelines, stringent requirements for fairness audits and explainability tools must be enforced to prevent potential undesired consequences.

This conversation demonstrates that Generative AI is not merely a means to circumvent a data deficit, but rather a strategic driver of privacy-first, continuously adaptive fraud detection systems. With a prudent approach to validation, ethical checks, and scalable engineering infrastructure, synthetic data generation has significant potential to help future-proof fraud analytics pipelines against two intractable problems: data scarcity and the constantly shifting nature of fraud.

6. Conclusion

The rising sophistication of financial frauds along with the rise in data privacy regulation and the lack of labeled examples of fraudulent activities has create an opposition to traditional data engineering practices. This paper introduced a generative AI based process to enrich fraud detection pipelines with synthetic data that would generate lifelike, privacy-preserving and quality fraud scenarios. The study showed that the inclusion of Conditional GANs in data workflows tones the robustness, sensitivity, and generalization ability of the model and it also overcomes the intrinsic shortcomings of conventional oversampling methods and influence of raw data.

Via a rigorous methodology in deployment and evaluation, we confirmed that synthetic data produced by CTGANs improves recall and F1-score and that this occurred while being willing to accept trade-offs with precision and false positives. Even more importantly, these gains were made without sacrificing user privacy, thanks to state-of-theart adversarial training that prevents over-fitting and mimicry of the source domain data. The performance improvement in our experiments—over even baseline models and with SMOTE-augmented model—demonstrates that the richer probabilistic model is useful in generating complex fraud behavior that are otherwise underrepresented in traditional data modalities.

The system proposed in this work depicts how syn- thetic data generation could be made an operational procedure at scale. By integrating such generative models into modular ETL pipelines orchestrated by Apache Airflow, Spark, and Delta Lake, organizations can keep the journey towards detect-then-prevent and detect-and-investigate fraud both frictionless and reproducible. Making this synthetic data engine ranging from ready to deploy (via Docker) to containerized around hybrid and cloud deployment, which enables organizations to continue to train their fraud models on the fly against the new data distributions and threat signatures. "This type of flexibility is crucial in today's environment, where fraudsters continue to adapt their tactics at a rapid pace, rendering static rules and models ineffective over time.

Finally, our research also identified implications for broader regulatory and ethical aspects of fraud analytics and the role of generative AI. The capacity to create shareable, realistic, de-identified data is believed to enable new opportunities for cross-institutional collaboration, model benchmarking, and democratization of fraud research. This is particularly important in industries where data hoarding for privacy reasons has blocked collective intelligence and innovation. Simultaneously, there are ethical considerations around synthetic data to consider and proactively address -bias creep or synthetic fraud, which we must carefully mitigate through transparent model validation, bias audits, and explainability.

While the findings are encouraging, there are a few areas of improvement that we identify for future work. In the future, we intend to investigate the hybrid model that integrates GANs and graph neural networks/reinforcement learning agents, which can effectively capture temporal and relational fraud dynamics. In addition, we believe it will be important to develop evaluation metrics beyond statistical similarity, for example, causal validity and behavioral realism, to measure synthetic data quality in a more application-aligned manner. In addition, sector-wide standards around synthetic data governance, lineage tracking and fairness certification should develop as quickly as the technology to enable responsible use.

This paper has shown that Generative AI makes an enabling and cost-effective technique to solve some of the most challenging problems in the space of fraud detection and data engineering. Generative models facilitate the robust creation of synthetic fraud scenarios that do not compromise analytical value and adhere to privacy requirements, which in turn open up a possibility for more robust, intelligent and fair fraud analytics pipelines. As model architectures, validation approaches and ethical considerations continue to innovatively evolve, generative synthetic data is expected to enable and contribute to the most modern data infrastructure for financial markets.

7. References

1. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, *et al*. Generative adversarial nets. In: Advances in Neural Information Processing Systems. 2014;27.
2. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. In: Neural Information Processing Systems. 2019.
3. Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Inf Sci. 2019;479:448-55.
4. Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. In: International Conference on Learning Representations (ICLR). 2017.
5. Kingma DP, Welling M. Auto-encoding variational

Bayes. In: International Conference on Learning Representations (ICLR). 2014.
6. An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. Spec Lect IE. 2015;2(1):1-18.
7. Bordawekar R, Mehta S, Dai Y. Synthetic data augmentation for fraud detection at scale. In: IEEE International Conference on Big Data. 2021. p. 1708-17.
8. Bhatia M, Goel V, Chhabra R. Designing modular ETL pipelines for scalable AI workflows in financial services. IEEE Trans Serv Comput. 2025 [Early Access].
9. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics (AISTATS). 2017. p. 1273-82.