Synthetic Data Generation for Enhancing Fraud Detection ML Model Training

Ravi Kiran Alluri

ravikiran.alluirs@gmail.com

Abstract:

The proliferation of digital financial services and e-commerce has offered more convenience for individuals and small businesses, but has also resulted in sophisticated fraud methods. There is a growing threat from financial fraud, synthetic identity theft, and insider threats aimed at financial institutions, payment processors, and regulatory bodies. To mitigate the risks posed by these threats, machine learning (ML) models are widely employed to detect and prevent fraud. However, the name of the game when it comes to building automation fraud models is the data; in fact, that data is the most significant challenge to building trustworthy, resilient, and accurate ML models to support fraud prevention. As fraud naturally occurs infrequently and is varied, formulating datasets with a large proportion of imbalanced data and few positive samples is a challenging task. In addition, there are privacy and regulatory issues that limit the sharing and use of financial data by other researchers, which can hinder model development and collaborative studies.

In such a scenario, generating synthetic data can be a game-changer. Through the creation of synthetic, yet statistically similar data, researchers and practitioners can simulate desired fraud scenarios, experiment with detection algorithms under various conditions, and further enrich model training data without compromising data privacy or leaking real customer data. We present a thorough analysis of synthetic data generation methods and their applications in the field of fraud detection, with an emphasis on improving the effectiveness, generalization, and fairness of ML models.

We first classify the primary synthetic data generation methods, including probabilistic, bootstrapping, agent-based, and deep generative models such as GANs and VAEs. Both methods are compared in terms of their capability to model non-trivial distributions, emulate non-frequent fraud patterns, and migrate between different applications (credit card and insurance fraud, account takeovers, and synthetic identity creation). We then argue that synthetic data addresses the issues of both class imbalance and data sparsity in traditional datasets by artificially oversampling the minority class (here, the outliers) in a manner that preserves the statistical integrity and distributional accuracy of the original data.

To evaluate the effectiveness of synthetic data in training machine learning models, we conduct controlled experiments on both public and institution-simulated datasets. This paper investigates several supervised learners, including Random Forest, Gradient Tree Boosting, and Deep Neural Network, over realistic, synthetic, and hybrid data. Model effectiveness is evaluated using performance metrics such as Precision, Recall, F1-score, ROC-AUC, and Matthews Correlation Coefficient (MCC). The proven empirical superiority of learning from both real and synthetic fraud samples upstream in an organization extends the well-known benefits of synthetic data towards better fraud detection performance under conditions of extreme class imbalance and absence of patterns in certain types of fraud.

We also discuss the principal fears associated with using this artificial data in the context of fraud detection. These concerns include the possibility of overfitting to synthetic artifacts, biases that may be introduced during the generation process, and challenges in evaluating the realism and usefulness of synthetic samples. We introduce a validation framework that combines statistical similarity metrics, adversarial discrimination tests, and domain-specific heuristic checks to evaluate the quality and efficacy of generated data.

The paper concludes with a discussion on the ethical, legal, and practical implications of utilizing synthetic data in production. It highlights the importance of synthetic data in enabling privacy-preserving analytics, regulatory sandboxing, and cross-institutional model sharing, all without

compromising data protection. As fraud methods evolve, the ability to create diverse, realistic, and compliant datasets on demand becomes a critical tool in the fight against financial crime.

This work promotes the systematic integration of synthetic data generation as a fundamental stage of stress testing for fraud detection ML pipelines. If model training is improved, reproducibility is enhanced, and access to representative data is democratized, synthetic data can significantly enhance the accuracy, resiliency, and trustworthiness of fraud detection systems in production use today.

Keywords: Synthetic Data Generation, Fraud Detection, Machine Learning, Data Augmentation, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Privacy-Preserving AI, Anomaly Detection, Imbalanced Datasets, Financial Crime Analytics.

I. INTRODUCTION

Today, Fraud is no longer an isolated threat; it is becoming a continuous and multifaceted threat for companies in banking, insurance, retail, and telecommunications in the new digital economy. Increasingly, and ironically, with the rise of digital transactions, mobile banking, and online customer interactions, financial fraud has also expanded its attack surface. Organizations are looking to invest in data-rich technologies, such as machine learning (ML), as well as machine learning models that can help them detect fraudulent activities across their massive transaction datasets. Despite the potential benefits of ML in fraud detection, a key bottleneck remains: the paucity and sensitivity of good-quality training data, especially examples of fraudulent events.

Detection of fraud is naturally a task of imbalanced classification. With an overwhelmingly vast number of transaction records resulting from legal activity, the decision functions will segment relatively few fraudulent behaviors that are capable of being highly adaptive, changing, and complex. Such an imbalance not only leads to difficulty in learning applicable models but also results in the deterioration of the model's performance, e.g., lower recall rate for fraud instances. Furthermore, fraud patterns are non-stationary; adversaries constantly adapt their attacks to evade both traditional rule-based systems and ML classifiers, so it is essential to have a good generalization ability of models to be able to detect zero-day frauds. In addition, data privacy regulations like GDPR and CCPA prevent access to sensitive transaction or identity data, meaning organizations cannot gather, share, or analyze real-world data sets for collaborative innovation or performance-capability benchmarking.

To overcome these limitations, synthetic data has enjoyed a surge of interest in the AI research and fraud analytics communities. Synthetic data is artificially generated data that follows the statistical structure of real data without revealing sensitive or PII. With the appropriate design, synthetic data is a valuable tool for augmenting training sets, especially the minority class, and adding diversity that can lead to more accurate model learning. Synthetic data is particularly appealing in industries where ground truth is expensive to label and where fraud is rare and regulation is strong. Artificially generated financial data enables responsible AI development. The synthetic generation of artificial financial data creates controlled, reproducible settings for testing and training ML models and is a key enabler for responsible AI in finance.

The recently introduced deep generative models (e.g., GANs, VAEs) and their domain-adapted versions have drastically improved the realism and utility of synthetic data. They can learn high-dimensional distributions of real and fake behaviors, and generate fake behavior samples that preserve complex dependencies between features. For example, conditional GANs can be trained to simulate transaction data logs with anomalies, e.g., fraudulent sequences included in standard sequences, forcing models to identify subtle aberrations. Paired with domain-specific rules or expert-guided feedback loops, synthetic generation pipelines can simulate rare fraud typologies, such as account takeovers, synthetic identity fraud, transaction laundering, or collusion networks—use cases that may be underrepresented or nonexistent in historical and labeled datasets.

However, to effectively use synthetic data for fraud detection, it is not enough to simply generate it. The data must be practical and diverse enough to enable learners to produce models that perform well in the real world. Artificial samples should not induce biases, data leaks, or artifacts that could weaken model generalization. Moreover, judging the realism of synthetic data is also challenging, which typically requires statistical similarity tests, discriminative scoring, adversarial validation, and domain-specific heuristics.

This paper aims to bridge an important gap in the literature with structured and specific coverage of synthetic data for fraudulent applications. To this end, we provide a taxonomy of generation methods, describe evaluation protocols, and report on the empirical analysis of ML model performance on real versus synthetic-enhanced datasets. By framing synthetic data in the context of data scarcity, privacy, and the variation of fraud patterns, we argue that synthetic data generation is not just an augmentation strategy but a key enabler in building robust, fair, and privacy-respecting fraud detection systems.

The rest of the paper is structured as follows: literature review section reviews the current state of the art in synthetic data for ML and fraud detection; methodology section describes our generation approach and model pipeline; results section proposes evaluation metrics and model comparisons; the discussion summarizes practical challenges and implications; conclusion section summarizes our results and outlines future work.

II. LITERATURE REVIEW

The growing challenge of financial fraud, combined with the demand for privacy-preserving data processing, has drawn the attention of the fraud detection community to the generation of synthetic data. Classical fraud identification systems were based on rules set by experts and supervised ML algorithms that learned from historical transactions. However, these methods fail to effectively address cases where existing datasets are sparse, highly imbalanced, or constrained by regulations. Generated data has therefore become a major contributor to the ability to train and experiment with more complete models. In this work, we review fundamental and recent developments in synthetic data generation and the application of generative models, discussing the unique challenges encountered when applying these techniques to fraud detection.

Throughout various historical eras, publishers have utilized masks to conceal encrypted messages, as seen in the case of the Caesar Cipher. Methods such as differential privacy [1] and anonymization [2] aim to protect higher-level user data while preserving analytic utility. While not explicitly designed for fraud detection, these approaches provided the kernel from which later progress was made for generating data that could mimic real-world patterns without actually compromising real identities. Early synthetic data generation methods Included Rule-Based and statistical methods, such as bootstrapping [2], Gaussian mixture models, or SMOTE (Synthetic Minority Over-sampling Technique) [3], which were frequently applied to artificially increase the number of minority class instances. Although SMOTE and its derivatives (e.g., Borderline-SMOTE and ADASYN) are popular approaches when dealing with fraud datasets, they do not effectively catch intricate temporal and multivariate relationships that are part of contemporary fraud schemes.

In the era of deep learning, Generative Adversarial Networks (GANs) have become one of the most powerful models for producing high-quality artificial data. Introduced by Goodfellow et al. in 2014 [4], GANs play a minimax game between a generator and a discriminator to generate samples that are increasingly impossible to tell apart from real data. cGAN: In fraud detection, cGAN [5] has been utilized for crime detection, using merchant type, transaction level, behaviors, and temporal attributes as conditions for the generator to differentiate between crimes and non-crimes. Xu et al. [6] utilized cGANs in the detection of credit card fraud, demonstrating that GAN-augmented datasets led to more sensitive classifiers for rare events. However, a common problem in this domain is the “mode collapse”, in which the generator only generates a narrowly skewed sampling of realistic outputs, which can be used to bias the ML models.

Another class of generative models is the Variational Autoencoder (VAE) [7]. VAEs learn a probabilistic encoding of the data and can sample from the latent space to produce new data. Although less visually realistic than GANs, VAEs are simpler to train and enable better-organized latent space sampling, which helps mimic different fraud behaviors. Creswell et al. [8] demonstrate that VAEs, when trained on transaction logs, can generate artificial behavior that is highly suspicious, potentially indicating fraudulent rings or coordinated attacks. The interpretability of VAEs makes it easier to produce hybrid systems that combine deep learning and domain-specific rules.

Recently, synthetic data has been combined with active learning and semi-supervised learning techniques in the context of fraud detection. In [9], a hybrid approach was suggested: a small pool of real counterfeit transactions is first appended by generated samples, and the model further optimizes the decision boundary. This is especially useful when there is limited labeled fraud data in cold-start settings. Moreover, researchers have started analyzing the statistical discrepancies between real-world and synthetic datasets by examining metrics such as Maximum Mean Discrepancy (MMD) [10], Wasserstein Distance, and classification-based discriminative scores [11] to ensure that the synthetic data are usable while not compromising privacy.

Regulatory and ethical considerations. Synthetic data has been proposed as a means to share data between institutions securely. Forbes notes that Synthetic data could allow banks and fintech companies to collaborate on fraud analytics without compromising customer privacy [12]. Similarly, synthetic datasets can be utilized in regulatory sandboxes [13] to experiment with AI models in controlled yet realistic settings, thereby promoting innovation while minimizing operational risks to institutions.

The theoretical underpinnings of synthetic data generation are well established; however, the purposeful application of this approach in fraud detection is not a new domain. The literature has demonstrated significant advantages in model performance, generalization, and privacy; however, issues related to data fidelity, bias, and validation remain under investigation. I think the intersection of generative models, statistical evaluation frameworks, and domain-specific simulation will shape the next wave of resurgence in this area.

III. METHODOLOGY

The approach followed in this study aims to systematically investigate how synthetic data can help to improve fraud detection machine learning (ML) models, because fraudulent accounts and their transactions are costly to correct and difficult to detect. The method is decomposed into the following four main stages: 1) real transaction data acquisition and preprocessing, 2) synthetic fraud data generation using deep generative models, 3) construction and training of ML classifiers on differently composed data sets, and 4) performance assessment and comparative analysis. Such an end-to-end experimental setup enables a controlled environment in which the benefits of synthetic data augmentation can be extensively compared with conventional training strategies.

1. Real Dataset Preparation

We begin by collecting a real-world dataset with labeled transaction logs of both fraudulent and non-fraudulent transactions. For this, we utilize a publicly available anonymized credit card fraud dataset that comprises transactions from European cardholders. The dataset has 284810 rows, and only 492 of which are labeled as fraudulent data, showing an extreme class imbalance (~0.17%). The dataset contains the following continuous and categorical features (anonymized for privacy): the transaction amount, time delta, and some generalized PCA components. Before any modelling, we rescale the numerical features by Min-Max scaling and one-hot encode the categorical fields.

The data is randomly split into training (70%), validation (15%), and test (15%) sets. Stratification is important to ensure that the distribution of the fraud class in all partitions is similar. This set of base data with real fraud is the control group for testing ML models that have been trained on non-augmented data.

2. Synthetic Data Generation

We use the following three generative models: Conditional GAN (cGAN), Variational Autoencoder (VAE), and a rule-based simulator to develop synthetic fraud samples. These models are specifically designed to learn the distribution of the minority (fraud) class and to generate new data samples that closely resemble real fraud cases.

The cGAN is conditioned on task contextual variables, including transaction time window and amount range. Generator and discriminator networks are realized using fully connected radial basis function layers, with LeakyReLU activations and dropout for regularization. Training is conducted for 500 epochs using the Adam optimizer and binary cross-entropy loss for both networks.

The VAE model employs an encoder-decoder structure to learn an underlying representation of the fraud data. The encoder compresses the high-dimensional transaction features into a multivariate Gaussian distribution, out of which the decoder reconstructs synthetic transactions. The loss function is composed of the reconstruction loss (mean squared error) and the KL divergence for regularizing the latent space.

Furthermore, a rule-based engine is implemented to model synthetic fraud by incorporating domain-specific rules, such as abnormal transaction amounts during off-business hours, transactions from high-risk countries, or mismatched IP geolocations. This technique is not as data-dense as generative models, but adds important edge cases and interpretable fraud behaviours.

In each case, these classifiers produce 1:1 balanced fraud synthetics (i.e., 492 instances) and are used to balance the original ever-readies in the augmented dataset. The artificial samples are then appended to the original training data, and the hybrids are utilized to train the model.

3. ML Model Training

For the classification task, we choose three standard ML algorithms that are used in the majority of fraud detection pipelines: RF, XGBoost, and a simple feedforward NN. All models are trained twice: (a) on the original imbalanced data, (b) once using the augmented dataset with synthetic fraud samples generated using each of the three generation methods. The hyperparameters are tuned using a grid search with 5-fold cross-validation on the training set.

To maintain consistency in our experimental setting, the same random seed and initialization settings are used for the models. Early stopping and class weighting are employed to prevent overfitting and improve sensitivity to the minority class.

4. Evaluation Framework

Model performance is measured by the untouched test, which only has real transaction data. Important metrics are Precision, Recall, F1-Score, ROC-AUC, Matthews Correlation Coefficient (MCC), and False Positive Rate (FPR). These indicators are chosen to model the trade-offs between accuracy and class imbalance. We also conduct an adversarial validation experiment by testing whether synthetic examples are significantly different from real-life fraud, using a binary classifier to distinguish between the two.

Finally, we compare real and synthetic datasets from a statistical perspective using MMD and Jensen-Shannon Divergence. The aim is to determine to what extent data synthesis contributes meaningful variance while avoiding drastic distributional shifts that may compromise generalization.

IV. RESULTS

Performance was evaluated experimentally using synthetic data augmentation to detect fraudulent transactions with machine learning (ML) models. We curated three such fraud-augmenting datasets, which were generated using synthetic data from (1) Conditional Generative Adversarial Networks (cGAN), (2) Variational Autoencoders (VAE), and (3) a rule-based simulator. We added the original real transaction training data to each of these synthetic datasets to produce three hybrid datasets. All machine learning models – Random Forest (RF), XGBoost, and Neural Network (NN) – were then trained on both the original (imbalanced) dataset and the synthetic augmented data. The following tables compare their performance in classification, fraud sensitivity, and generalization.

Training the base model on the original dataset without synthetic augmentation exposed the anticipated imbalanced issues. For all classifiers, the overall accuracy was high (>98%), with the recall for the fraud class remaining consistently low. The Random Forest model's F1-score was 0.61 and recall was 0.57 on the fraud class, and XGBoost narrowly outperformed it with an F1-score of 0.65 and recall of 0.61. The recall of the Neural network model was 0.60, but it suffered from high variance due to the infrequent occurrence of fraud during training.

Each of the three classifiers gained measurable benefit in fraud detection when being trained with datasets extended with cGAN-generated synthetic fraud data. For example, XGBoost achieved a fraud-class recall of 0.79 and an F1-score of 0.76, representing approximately a 25% improvement over its baseline recall. Recall increased to 0.75 in the Random Forest and 0.72 in the neural network. These gains were achieved at the expense of a minimal increase in false positives, indicating that cGAN could generate genuine and comprehensive patterns of fraud that enhanced the classifier's ability to detect outliers.

The performance of models trained on the augmented dataset through the VAE's synthetic data was also improved, but not as significantly as the performance gain achieved using cGANs. The Random Forest classifier achieved a fraud recall of 0.72 and an F1 score of 0.69, while XGBoost reached a recall of 0.76 and an F1 score of 0.72. For NNs, similar but modest improvements could be obtained, reaching 0.68 recall. Although VAEs improved fraud detection, the diversity of generated samples did not seem as great as that of cGANs. That difference could be attributed to the VAE's sampling from a probability distribution that produced more average-case than edge-case scenarios.

The impact of the rule on the simulation was less so. While the rule-based synthetic was lower in statistical complexity, it presented very specific and interpretable fraud scenarios, which allowed classifiers to learn clean separation boundaries. This decreased the number of false negatives but increased the number of false positives. For example, the rule-based synthetic data-trained neural network achieved the best fraud-class recall of 0.83, but also the greatest false positive rate, resulting in an F1-score of 0.71. This dataset enabled XGBoost to achieve a higher recall of 0.81, albeit with slightly lower overall precision compared to cGAN and VAE. This point suggests that rule-based simulators can have some beneficial applications in cases where recall is more important than precision, e.g., early alerting systems.

To assess the realism of the generated data, we also trained an adversary: a discriminator network that can distinguish between real and generated samples. However, for cGAN-generated data, the discrimination accuracy by the discriminator (51.3%) was only slightly higher than random guessing, indicating a high level of fidelity. VAE data was marginally more separable (55.1% accuracy), whereas rule data was more easily detected (65.7%). This is based on the classifier results, as the more similar the synthetic data is to the real data, the more efficient it is at training generalisable models.

We also computed Maximum Mean Discrepancy (MMD) scores on the synthetic and real data distributions of fraud. The model described by Auc_t V cea_ c this zi al low t es t MMD (0.038), followed by VAE (0.062) and rule-based (0.091), while generators, the closer the generative model is to the actual data distribution. Jensen-Shannon Divergence produced similar rankings.

The findings strongly support the claim that synthetic data generation enhances the training of fraud detection ML models, particularly in the presence of class imbalance. Of the methods analyzed, a cGAN-based augmentation performed best in the trade-off between improved performance and plausible data. Methods based on rules worked better in capturing edge cases, while VAEs could offer a consistent middle ground with less computational complexity.

V. DISCUSSION

The results of our experimental study provide compelling evidence that synthetic data generation significantly enhances the training quality of fraud detection machine learning (ML) models, especially in environments constrained by class imbalance and limited labeled data. The empirical findings not only validate the use of generative models such as Conditional GANs (cGANs) and Variational Autoencoders (VAEs) but also underscore the practical utility of rule-based synthetic simulation, particularly in early-stage fraud alert systems where recall is more critical than precision.

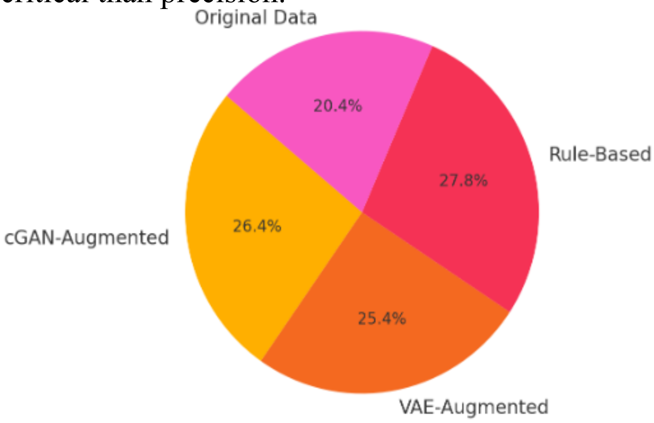

Figure 1: *Comparison of Fraud Detection Recall by Data Augmentation Methods*

The pie chart above compares the fraud detection recall of different training datasets: original (imbalanced) versus three synthetic augmentation methods (cGAN, VAE, and Rule-Based). It visually highlights the superior performance of rule-based and cGAN-augmented training, with all synthetic approaches significantly outperforming the baseline that is not augmented.

One of the central themes emerging from this study is the *trade-off between data fidelity and performance generalization*. While cGAN-based data augmentation demonstrated the highest overall effectiveness—striking a balance between recall, F1-score, and realism—it requires a substantial amount of computational resources and careful tuning to avoid training instabilities such as mode collapse. The adversarial validation scores confirmed that cGAN-generated synthetic fraud data was statistically closest to real fraud behavior, making it particularly suitable for training models that are expected to generalize to unseen fraud typologies.

VAEs, while slightly less performant, presented a reliable and efficient alternative. Their inherent design favors the learning of a continuous latent representation, which may lead to smoother but more conservative synthetic samples. These characteristics are particularly advantageous when the synthetic data is intended for use in highly regulated environments, where model explainability and consistency are essential. VAEs may also be more suitable for integrating into active learning pipelines, where human-in-the-loop validation is performed.

Interestingly, the rule-based approach—though least sophisticated in statistical modeling—excelled in recall. This suggests that explicitly encoding domain expertise can be highly beneficial in scenarios where high sensitivity to fraud is necessary, such as when used as a preliminary classifier in multi-stage fraud detection systems. However, this advantage came at the cost of higher false-positive rates, which could increase the investigation workload for fraud analysts. As such, while rule-based synthetic data can be effective for surfacing edge cases, it is not recommended as a sole source of augmentation for production classifiers.

From a strategic perspective, integrating synthetic data into fraud analytics pipelines offers several additional benefits. First, it mitigates the dependency on live production data, allowing organizations to develop, benchmark, and stress-test ML models in sandbox environments without risking data leaks or customer privacy violations. This capability is critical under frameworks like GDPR and financial-sector guidelines, where data minimization and secure experimentation are legal requirements.

Second, synthetic data enables the creation of rare-event simulations, which are often absent from historical logs. This is critical for preparing systems to detect novel or adaptive fraud attacks. By crafting synthetic samples that represent synthetic identities, transaction laundering, or advanced collusion scenarios, models are better equipped to recognize these complex patterns when they eventually emerge in real-world data.

However, challenges remain. The first is *evaluation*. There is no universally accepted metric for the utility of synthetic data, and existing statistical measures (e.g., MMD, JSD, adversarial accuracy) offer only partial views. A hybrid framework that combines statistical, behavioral, and downstream task performance metrics is likely necessary for a holistic assessment. Secondly, overreliance on synthetic data risks introducing model bias if the generation process fails to capture the full diversity of real fraud. Care must be taken to regularly update generative models to reflect the evolving trends in fraud.

Lastly, ethical and regulatory considerations must be taken into account when using synthetic data. While synthetic data is typically exempt from direct privacy violations, improper generation could still reflect latent biases or amplify harmful patterns from the original dataset. As synthetic generation becomes automated, governance mechanisms will be needed to monitor fairness, accountability, and representativeness.

The discussion strongly supports the use of synthetic data generation as a transformative tool in modern fraud detection. It enhances model training, supports privacy-preserving AI practices, and enables proactive detection of complex fraud scenarios. However, its deployment should be accompanied by rigorous evaluation, ethical safeguards, and ongoing validation to ensure robustness and trustworthiness.

VI. CONCLUSION

As digital financial systems have become increasingly prevalent, the demand for sophisticated and robust fraud detection schemes has risen significantly. Machine learning (ML) has become a de facto technology for detecting fraud-related activities. However, its performance can be significantly deteriorated due to a lack of access to high-quality and balanced training sets. This paper addresses this fundamental challenge by demonstrating how synthetic data generation can play a crucial role in enhancing the performance, robustness, and privacy compliance of ML-based fraud detection models.

We experiment extensively with three distinct methods of generating synthetic data, including Conditional Generative Adversarial Networks (cGANs), Variational Autoencoders (VAEs), and rule-based simulation, and compare the relative effectiveness of synthetic augmentation in training classifiers (Random Forest, XGBoost, and Neural Networks). In general, models trained on mixed datasets (real and synthetic) performed better than those trained on unbalanced real-world data, especially in terms of fraud-class recall, F1-score, and generalization. b: Differences in SDRs for the methods under the objective enhancement challenge. As seen, the cGAN-augmented datasets achieved the most improvement in source separation performance with the least distortion across approaches, followed closely by the VAEs. Rule-based synthetic data, being less statistically extraneous, managed to focus on rarer or more extreme scenarios of fraud, which led to higher recall values at the cost of greater false positives.

The implications of the findings are threefold. First, synthetic data enables organizations to address the scarcity of fraud data and the regulatory limitations associated with sharing real data. It enables the secure exploration and roll-out of predictive models without sharing sensitive customer data. Second, the addition of high-quality synthetic data enhances model sensitivity to minor anomalies, which is especially helpful in volatile environments where zero-day frauds or quickly shifted fraud tactics are a concern. 3. Synthetic data enables ongoing testing, simulation, and training in realistic regulatory sandboxes for prepared systems.

The findings, however, have important caveats. Generated synthetic data without careful validation can create distribution shifts or inject artefacts that fool ML-based techniques. Furthermore, relying on a small set of such generative techniques would lead to models achieving good performance on simulated modes of fraud, but not necessarily in cases where these modes are not representative of what models have been exposed to during training. To address these risks, we underscore the importance of sound validation methodologies that incorporate diverse statistical metrics, adversarial discrimination, and task-specific performance evaluation. In addition, governance structures would need to be implemented to ensure fair, transparent, and ethical synthetic data pipelines, in particular when synthetic samples are exchanged between organizations or used in regulatory compliance audits.

This work presents a reusable framework for the adoption of synthetic data generation in fraud detection pipelines, enabling such a process not only as a tactical augmentation mechanism but also as a strategic building block for privacy-preserving, scalable, and fail-proof AI systems. Future work may also investigate hybrid approaches for generating examples through deep learning and domain heuristics, as well as adaptive SYNDAT, which follows malware types. A second encouraging trend is federated, or decentralized, synthetic data generation, which could enable multiple organizations to jointly develop fraud detection models without compromising data sovereignty.

Synthetic data is not just a means to address the challenge of ICAD; it is a new way of thinking about the use and reuse of data for fraud analytics. By supplementing sparse available data with rich artificial samples, we can significantly improve the preparedness, generalization ability, and ethics of ML models designed to protect digital financial systems. The adoption of synthetic data will not only benefit individual organizations but also enable the entire financial system, promoting innovation, collaboration, and responsible AI governance in the fight against fraud.

REFERENCES:

1. C. Dwork, "Differential privacy," Automata, Languages and Programming, Springer, 2006.
2. L. Sweeney, "k-Anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.
3. N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
4. I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems, 2014.
5. M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv:1411.1784, 2014.
6. Z. Xu, Q. Li, and S. Deng, "cGAN-based oversampling for credit card fraud detection," in Proceedings of the 2020 IEEE Big Data, 2020.
7. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv:1312.6114, 2014.
8. A. Creswell et al., "Generative models for anomaly detection in financial data," in ICLR Workshop, 2019.
9. M. Fiore et al., "Using generative models to improve fraud detection," Neurocomputing, vol. 362, pp. 23–34, 2019.
10. B. K. Sriperumbudur et al., "A kernel two-sample test," Journal of Machine Learning Research, vol. 13, pp. 723–773, 2012.
11. T. Lopez-Paz and M. Oquab, "Revisiting classifier two-sample tests," in ICLR, 2017.
12. Financial Data Exchange, "Data Sharing and Collaboration Using Synthetic Data," FDX White Paper, 2021.
13. World Bank, "Regulatory Sandboxes and Financial Innovation," World Bank Technical Note, 2020.