Privacy-Preserving Financial Fraud Detection: Challenges and Solutions with Generative Models, Lifetime-Aware Detection, and Federated Boosting

Dae-Young Park
FSI and KAIST*
Gyeonggi-do and Daejeon, Republic of Korea
mainthread@fsec.or.kr

Jaeyoung Cheong†
TossBank Corp
Seoul, Republic of Korea
david.cheong@toss.im

Jingwan Ha‡
KBank Corp
Seoul, Republic of Korea
jingwan@kbanknow.com

Minsoo Park‡
KBank Corp
Seoul, Republic of Korea
minsoo.pk@kbanknow.com

Ohyeon Kwon‡
KakaoBank Corp
Gyeonggi-do, Republic of Korea
fyve.kwon@kakaobank.com

Saebitna Oh‡
KakaoBank Corp
Gyeonggi-do, Republic of Korea
kiara.oh@kakaobank.com

Youngjun Kwak‡
KakaoBank Corp
Gyeonggi-do, Republic of Korea
vivaan.yjkwak@kakaobank.com

In-Young Ko
KAIST
Daejeon, Republic of Korea
iko@kaist.ac.kr

Kyeongkyu Kim
FSI
Gyeonggi-do, Republic of Korea
kkkyu@fsec.or.kr

Abstract

While privacy regulations prohibit direct data sharing among institutions, improving fraud detection performance requires collaboration across banks. To mitigate this limitation, we have conducted a real-world case study on privacy-preserving financial fraud detection (FFD) in the South Korean banking sector. During the research, we have identified four major challenges in practice: (C1) the degradation of tabular generative models under extreme class imbalance and sparsity, (C2) the lack of utility–privacy joint evaluation methodology, (C3) the inability of detection models to capture irregular active lifetime of fraudulent activity, and (C4) the absence of robust federated gradient boosting under dynamic participation. In this work, we introduce two novel approaches: (i) Graph-theoretical Generative Models (GGMs), which leverage graph theories to generate high-utility synthetic tabular data; and (ii) Active Lifetime-Aware Fraud Transaction (ALAFT), which adjusts fraud scores by defining and modeling active lifetime of fraudulent patterns. Across two private banking datasets and a public benchmark, GGMs consistently outperform seven baselines, while ALAFT outperforms significant gains over six representative detectors, reducing false positives during high-risk periods. Finally, we outline our ongoing work, fraud scenario-aware and similarity-based FedXGB-Bagging with *KakaoBank*, *TossBank*, and *KBank* to enable secure collaboration and support nationwide anti-fraud efforts.

*Financial Security Institute (FSI) and Korea Advanced Institute of Science and Technology (KAIST)
†**These authors contributed equally to this work. Author order follows alphabetical order.**

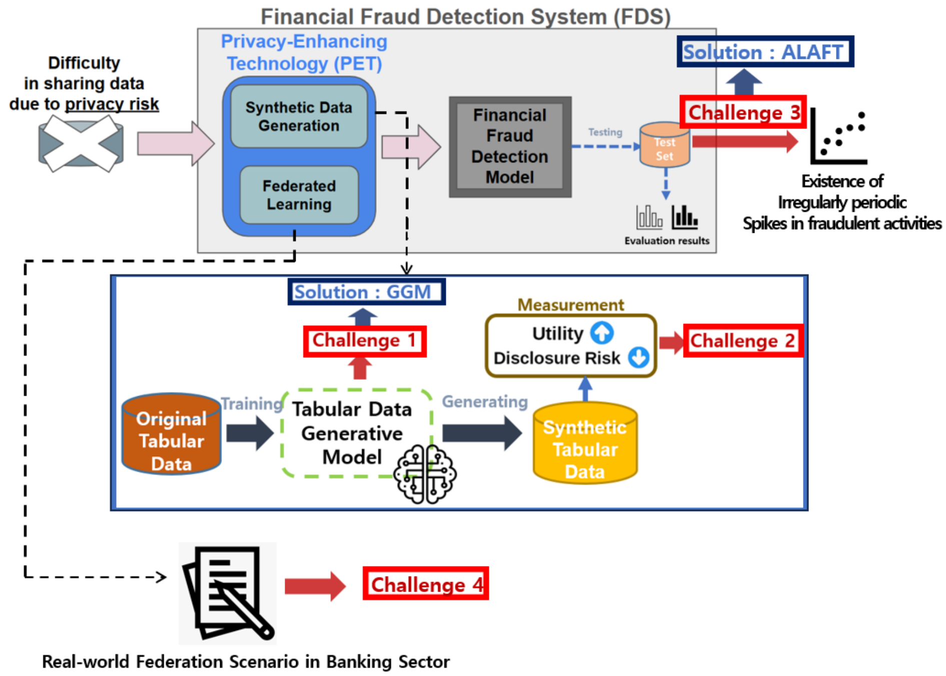


Figure 1: Overview of the challenges and the proposed solutions

1 Introduction

Financial fraud detection systems (FDS) face growing demands for higher detection accuracy while ensuring strict compliance with privacy regulations such as the Personal Information Protection Act (PIPA) [1, 2]. In South Korea's financial industry, collaboration between financial institutions is essential for improving fraud detection model performance, yet direct data sharing is infeasible due to privacy concerns. To overcome this limitation, our research has explored two primary privacy-enhancing technologies (PETs): synthetic data generation and federated learning (FL).

However, through years of practical engagement, we have identified four key challenges that lower the effectiveness of PET-based financial FDS in real-world deployments. These challenges emerge from financial fraud detection (FFD) data characteristics, evaluation methodology of tabular generative models for financial transactions, irregularly temporal active lifetime of fraudulent activities, and technical constraints in federation environments. In this paper, we present these challenges and solutions by combining empirical findings and validations from our previously published works at ACM CIKM 2024, IEEE BigComp 2024, ACM KDD 2025, and accepted one at ACM CIKM 2025. Figure 1 illustrates the four challenges (C1–4) and the proposed solutions (corresponding C1 and C3). In progress works, we are developing solutions for C1 and C3. We are also conducting proactive research with South Korea's three internet banks—KakaoBank, TossBank, and KBank—and will share results from our real-world case study to advance financial fraud detection systems and support nationwide anti-fraud efforts.

2 Four challenges in a real-world case study

Challenge 1: Difficulty in improving detection performance with tabular generative model for synthetic data due to inherent characteristics of FFD datasets. FFD dataset exhibits extreme class imbalance, high sparsity, and non-normal attribute distributions. When the intensity of these characteristics increases, the performance of existing tabular generative models deteriorates.

Challenge 2: Lack of practical methodology to evaluate both utility and disclosure risk of synthetic tabular data. While many studies focus on generation quality [3, 4, 5], few works provide quantitative methods to jointly assess utility and privacy disclosure risk of tabular synthetic data created by tabular generative models. This gap hinders deployment of the generated tabular synthetic data used in compliance-sensitive financial environments.

Challenge 3: Failure of existing fraud detection models to capture "active lifetime" periods of fraudulent activities. The fraudulent activities often exhibit concentrated bursts over irregularly temporal patterns driven by factors such as economic conditions [6, 7]. We observe the existing detection models underperform in detecting fraud transactions during these high-activity windows.

Challenge 4: Lack of practical FL algorithms for gradient boosting models and lack of scenario-aware FL studies. Currently, FL studies cover neural network-based models [8, 9, 10]. However, gradient boosting-based models are most commonly used in real-world FFD applications. In addition, when considering real-world financial scenarios, participating institutions face varying constraints: some join late, some drop out mid-training, and some contribute small datasets (quantity skew).

3 The proposed solutions

Graph-theoretical Generative Model (GGM) to address Challenge 1. We developed novel graph-theoretical generative models, named *SeparateGGM* and *SignedGGM*. GGM consists of several steps: (1) the features of the original data are augmented to enrich the data through a GNN model (i.e., GraphSAGE [11]); (2) separate directed K-NN graphs and signed directed K-NN graphs with positive and negative edges are created based on similarity between data instances and class relationships; (3) graph topological and connectivity analysis is conducted. Then, through the analysis, the separate and signed graphs are effectively selected based on criteria that align with the objective of synthetic data generation, maximizing the performance of the target detection model; (4) graph centrality indicators are calculated within the selected graphs to determine the influence score of each node (data instance), which is used as a weight for the objective function of our base generative model; (5) The base generative model (i.e., CTGAN[12]) is trained by the augmented data with the influence scores to generate tabular synthetic data.

Active Lifetime-Aware Approach to fraudulent Financial Transactions (ALAFT) to address Challenge 3. By considering active lifetimes of fraudulent activities on irregularly temporal patterns, ALAFT incorporates four ideas to enhance fraud detection performance during high-risk periods in banking transactions. (1) Consideration of active lifetime for specific customer, account, and transaction. We enrich transaction features by representing them as signed K-NN graphs at each level to capture temporal relationships within active lifetimes. (2) Consideration of confidence of being truly normal transaction among normal transactions in active lifetime. Within active lifetimes, we compute a "truly normal" confidence score for each normal transaction based on its similarity to known fraudulent and normal transactions. Transactions with high confidence are selectively sampled as normal. (3) Consideration of the fraud possibility of the remaining normal transactions in active lifetimes. Normal transactions not selected in the previous step may still contain undetected fraud. We assign each such transaction a fraud possibility value, derived from its temporal proximity to fraudulent transactions within the same active lifetime using multiple sliding window sizes. (4) Consideration of adjustment of the predicted fraud score based on the temporal distance of the nearest active lifetimes. Transactions closer to adjacent active lifetimes are more likely to be fraudulent. We calculate the temporal distance between each transaction and the nearest active lifetime (both before and after) and adjust predicted fraud scores accordingly.

4 Experiments

4.1 Experimental setup

Datasets. First, we utilize two private banking transaction datasets constructed under strict data privacy guidelines by an institute, Financial Security Institute (FSI) [3], one of South Korea organizations dedicated to enhancing cybersecurity and information protection within financial industry. We masked bank names to P and Q due to privacy policy [4]. These datasets include detailed information

[3]https://www.fsec.or.kr

[4]For detailed information, refer to A.2

of a view of customer, account, and transaction with the total 12 types of financial fraud scenarios [5]. Furthermore, second, to validate the generalized performance of our approach, we utilize a publicly available simulated banking dataset [13], which consists of 594,643 records over six months, including 16 merchant categories, demographic attributes. Among these, 1.21% are labeled fraudulent. This dataset holds only normal or fraud labels. All datasets were split 80:20.

GGM. We compare our proposed GGM with seven representative baselines covering diverse paradigms of tabular generative modeling: GC[14], CART[15], TVAE[16], TableGAN[17], CTGAN[12], DPHFlow[18], and TabDDPM[19]. These baselines are selected as they represent widely-used or state-of-the-art approaches for tabular synthetic data generation. For a fair comparison, the amount of generated synthetic data is fixed to match the size of the corresponding original dataset [20]. Fraud detection performance is evaluated using four metrics—Macro-F1, Weighted-F1, ROC-AUC, and PR-AUC—averaged over five popular detection models (Random Forest, LightGBM, MLP, LSTM+CNN, and TabNet). All experiments are repeated 100 times and average results are reported.

ALAFT. Six representative models were selected as base detectors: three machine learning models—SVM[21], RandomForest[22], XGBoost[23]—and three deep learning models—TabNet[24], SAINT[25], NODE[26]. We evaluate them in binary settings. Macro-F1 and ROC-AUC are the primary metrics for this setting, with false positive rate (FPR) additionally measured in the binary setting to account for operational burden from false positives. Performance is reported on the top-N% (1, 2, 5, 10%) transactions ranked by fraud score, reflecting practical constraints in manual inspection.

4.2 Empirical results

GGM. We presents the performance comparison of GGM (SeparateGGM and SignedGGM), and seven baseline tabular generative models. Despite the overall lower scores, SeparateGGM and SignedGGM consistently outperform all baselines in Macro-F1, ROC-AUC, and PR-AUC across the three datasets. Notably, SignedGGM achieves the highest ROC-AUC and PR-AUC in P-bank and Simulation, while SeparateGGM achieves marginally higher Macro-F1 in Q-bank.

We compares original base models and their ALAFT-enhanced versions on P-bank, Q-bank, and Simulation datasets for top 1% and 10% fraud scores. Across all base models and datasets, ALAFT consistently improves detection, with the best Macro-F1 and ROC-AUC highlighted in bold. The highest relative gains are marked with underlines, showing notable improvements especially for SVM and XGBoost.

5 Discussion and future direction

We identify four challenges and propose two solutions for privacy-preserving financial fraud detection. As next steps, we will (1) conduct ablation studies to quantify the contribution of individual components in GGM and ALAFT, (2) design practical solutions for both Challenge 2 (a need for effective utility–privacy evaluation) and Challenge 4 (a need for robust FL for gradient boosting).

We are developing fraud scenario-aware and similarity-based FedXGBBagging, enabling South Korea's internet banks (KakaoBank, TossBank, and KBank) to collaboratively train gradient boosting models without sharing raw data. By selecting and aggregating only similar decision trees across institutions, this approach aims to address heterogeneous data volumes, dynamic participation, and fraud pattern variability. We will continue to work closely with the three internet banks to advance state-of-the-art fraud detection technologies and contribute to the nationwide effort to combat financial fraud in South Korea.

Acknowledgments and Disclosure of Funding

This work was partly supported by the Institute of Information Communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2025-02218761, 75%) and IITP-

[5] A subset of these datasets was used in the South Korea's Data x AI Competition: `https://www.fsec.or.kr/bbs/detail?menuNo=66&bbsNo=11502`

ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT)(IITP-2025-RS-2020-II201795, 25%).

References

[1] Zidi Qin, Yang Liu, Qing He, and Xiang Ao. Explainable graph-based fraud detection via neural meta-graph search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4414–4418, 2022.

[2] Dawei Cheng, Sheng Xiang, Chencheng Shang, Yiyi Zhang, Fangzhou Yang, and Liqing Zhang. Spatio-temporal attention-based neural network for credit card fraud detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 362–369, 2020.

[3] Vamsi K Potluru, Daniel Borrajo, Andrea Coletta, Niccolo Dalmasso, Yousef El-Laham, Elizabeth Fons, Mohsen Ghassemi, Sriram Gopalakrishnan, Vikesh Gosai, Eleonora Kreav, et al. Synthetic data applications in finance. Technical report, 2023.

[4] Zhiqiang Wan, Yazhou Zhang, and Haibo He. Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 IEEE symposium series on computational intelligence (SSCI)*, pages 1–7. IEEE, 2017.

[5] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8, 2020.

[6] Hugo Van Driel. Financial fraud, scandals, and regulation: A conceptual framework and literature review. *Business History*, 2019.

[7] Waleed Hilal, S Andrew Gadsden, and John Yawney. Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems With applications*, 193:116429, 2022.

[8] Taek-Ho Lee, Suhyeon Kim, Junghye Lee, and Chi-Hyuck Jun. Harmosate: Harmonized embedding-based self-attentive encoder to improve accuracy of privacy-preserving federated predictive analysis. *Information Sciences*, 662:120265, 2024.

[9] Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho D Choi, and Sung-Ju Lee. Fedtherapist: Mental health monitoring with user-generated linguistic expressions on smartphones via federated learning. In *EMNLP*, 2023.

[10] Taek-Ho Lee, Suhyeon Kim, Junghye Lee, and Chi-Hyuck Jun. Word2vec-based efficient privacy-preserving shared representation learning for federated recommendation system in a cross-device setting. *Information Sciences*, 651:119728, 2023.

[11] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[12] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[13] Edgar Alonso Lopez-Rojas and Stefan Axelsson. Social simulation of commercial and financial behaviour for fraud detection research. In *Social Simulation Conference*, 2014.

[14] Huahua Wang, Farideh Fazayeli, Soumyadeep Chatterjee, and Arindam Banerjee. Gaussian copula precision estimation with missing values. In *Artificial Intelligence and Statistics*, pages 978–986. PMLR, 2014.

[15] William A Young, Gary R Weckman, Vijaya Hari, Harry S Whiting, and Andrew P Snow. Using artificial neural networks to enhance cart. *Neural Computing and Applications*, 21:1477–1489, 2012.

[16] Fadi Hamad, Shinpei Nakamura-Sakai, Saheed Obitayo, and Vamsi Potluru. A supervised generative optimization approach for tabular data. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 10–18, 2023.

[17] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10), 2018.

[18] Jaewoo Lee, Minjung Kim, Yonghyun Jeong, and Youngmin Ro. Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7345–7353, 2022.

[19] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 17564–17579, 2023.

[20] Dae-Young Park and In-Young Ko. An empirical study of utility and disclosure risk for tabular data synthesis models: In-depth analysis and interesting findings. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 67–74. IEEE, 2024.

[21] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

[22] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[23] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[24] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[25] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[26] Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312*, 2019.

[27] Dae-Young Park and In-Young Ko. Urban event detection from spatio-temporal iot sensor data using graph-based machine learning. In *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 234–241. IEEE, 2022.

[28] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10 (1):1–9, 2019.

[29] Claire Little, Mark Elliot, Richard Allmendinger, and Sahel Shariati Samani. Generative adversarial networks for synthetic data generation: a comparative study. *arXiv preprint arXiv:2112.01925*, 2021.

[30] Dae-Young Park. Graph-theoretical approach to enhance accuracy of financial fraud detection using synthetic tabular data generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5467–5470, 2024.

[31] Qinbin Li, Zeyi Wen, and Bingsheng He. Practical federated gradient boosting decision trees. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4642–4649, 2020.

[32] Chenyang Ma, Xinchi Qiu, Daniel Beutel, and Nicholas Lane. Gradient-less federated gradient boosting tree with learnable learning rates. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pages 56–63, 2023.

[33] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[34] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

A Supplementary Material

A.1 Implementation detail

GNN hyperparameters including the number of hidden features, number of folds for k-fold cross-validation, number of training epochs, and learning rate are set to 32, 5, 50, and 0.01, respectively [6]. The ReLU activator and Adam optimizer are used. To augment the raw features, the number of extracted features is set to 32 (through empirical search, we selected the value among 16, 32, 64, and 128). We also use Python-igraph 0.11.4 for graph construction and analysis. To address categorical attributes, we use a popular method, the one-hot encoder. δ for graph indicators is set to 0.0001 in the same way as [27]. For the base model (i.e., CTGAN), pac size, batch size, and the number of epochs are set to 11, 256, and 500, respectively, after empirically searched. DGL 2.0.0 library is used to implement the GNN models. The number of hidden features, number of K-folds, epochs, and learning rate are set to 32, 5, 50, and 0.01, respectively. To augment the raw features, the number of extracted features is set to 32 [7].

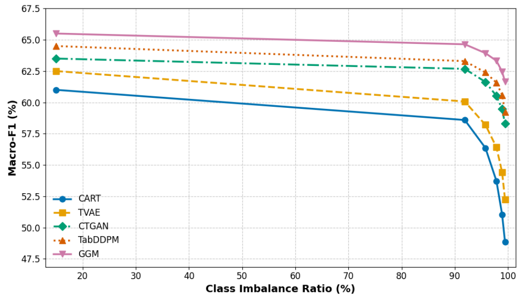
To implement SVM and RandomForest, Scikit-learn 1.5.2 library is used. Specifically, for SVM, the kernel function is set to a Radial Basis Function (RBF) kernel trick and a regularization parameter is set to 1.0. For RandomForest, the number of estimators, min samples split, a function to measure the quality of split points, and min samples leaf are set to 100, 2, Gini index, and 1, respectively. For XGBoost, XGBoost 2.1.3 library is used, with max depth, number of estimators, learning rate, child weights, and subsampling ratio set to 7, 100, 0.12, 1, and 1.0, respectively [8].

To implement TabNet model, pytorch-tabnet 2.1.3 library is used. The number of decision prediction layer width, attention embedding layer width, number of decision steps, maximum number of epochs, gamma, and batch size are set to 8, 8, 3, 200, 1.3, and 512, respectively. SAINT and NODE are implemented based on public repositories [9] [10]. For SAINT, the number of transformer layers, attention heads in each transformer layer, dropout ratio, epoch, batch size, an optimizer, and learning rate are set to 6, 8, 0.1, 100, 512, AdamW, and 0.01, respectively. For NODE, the number of total trees, tree depth, epoch, batch size, learning rate, and an optimizer are set to is 2048, 8, 100, 512, 0.001, and Adam. These values are selected after empirically searched. All experiments were conducted in Python 3.10.12 and Ubuntu 22.04.3 running on an Intel(R) Xeon(R) CPU @ 2.00GHz and A100 (CUDA version 12.2) with 51GB RAM.

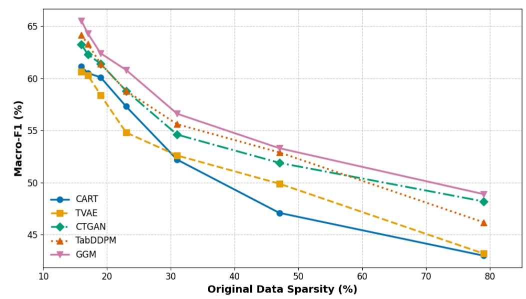
A.2 Detailed information of private banking datasets

We summarizes two private banking datasets (P-bank and Q-bank), covering customer, account, and transaction views over five years, with millions of records and low overall fraud rates. Label review was conducted based on two domain experts from the FSI, resulting in the removal of approximately 5% of the datasets that are deemed mislabeled, respectively. The label rates for fraud types range from 0.07% to 0.20% for P-bank and 0.09% to 0.24% for Q-bank.

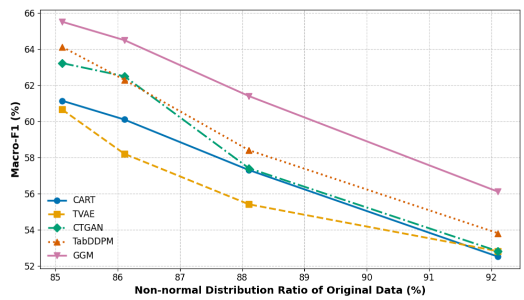
[6] We use the DGL 2.0.0 library to implement the GNN models.
[7] Through empirical search, we selected the value among 16, 32, and 64
[8] We usually follow default parameters which the library presents
[9] https://github.com/somepago/saint
[10] https://github.com/Qwicen/node



(a) Class imbalance



(b) Data sparsity



(c) Non-normal distribution

Figure 2: Investigation of challenge 1 stemming from the intensity of the three characteristics in the original FFD data and the changes in financial fraud detection performance of a target FFD model trained across the different generative models. The performance of existing tabular generative models declines as the intensity of the three FFD characteristics increases.

A.3 Detailed description of challenges

Challenge 1 As shown in Figure 2, we altered the class imbalance ratio of the original data, which is calculated as 1 minus the ratio of the minority class to the total data points, to more realistically simulate the challenges presented in real-world FFD dataset. This experimental setup allowed us to directly observe the effects of increasing class imbalance on the performance of FFD models trained using the generated tabular synthetic data. We undertook a more granular analysis by systematically removing 50% of the instances from the minority class in five successive stages, recalculating the new class imbalance ratio and model performance at each step. We similarly conducted experiments in terms of the change of sparsity ratio (increasing it exponentially by injecting zero value in some random cells) and attribute ratio with a non-normal distribution ratio (randomly choosing some continuous attributes following a normal distribution, which is determined using the Shapiro-Wilk test, and then transforming them to follow a multi-modal distribution). For example, in Figure 1-(b), the F1 score, which began at 77.3, dropped to 30.0, illustrates the impact of sparsity in the original data on model performance. For this preliminary study, we adopt CART, TVAE, CTGAN, TabDDPM, and GGM (i.e., SignedGGM) and use macro-F1. Thus, this investigation shows a pronounced decline in the performance of existing FFD models as the intensity of the three characteristics of the original data increases. We utilized CTGAN, a renowned tabular generative model for tabular synthetic data. In addition, we adopted average F1 and ROC-AUC values of two target detection models. We also used an average values of LightGBM and TabNet as performance indices because they are popular FFD models. To ensure the robustness of our preliminary study, this experimental process was repeated hundreds of times, and the average values of these iterations are used.

Challenge 2 While existing studies on tabular synthetic data focus on improving the utility of generated datasets, there is a notable lack of systematic approaches that jointly evaluate utility and disclosure risk in terms of tabular generative models. The existing works of the related methodologies often cover a single dataset or limited set of metrics, which fails to capture the trade-offs between the two aspects.

From our previous study, we observed that:

- Trade-off relationship: Higher utility often comes at the cost of increased disclosure risk, especially when the volume of synthetic data increases [11].
- Dataset dependency: Utility and disclosure risk vary significantly with data characteristics such as the proportion of continuous attributes, sparsity, and dimensionality.
- Metric limitations: Widely-used disclosure metrics such as Targeted Correct Attribution Probability (TCAP) fail to identify certain high-risk outlier records, leaving potential vulnerabilities unmeasured.
- Model-specific tendencies: Statistical and machine learning models may yield higher utility but also higher disclosure risk in some settings, while deep learning models often show lower disclosure risk but reduced utility on certain metrics.

These findings highlight the absence of a unified evaluation framework that can balance utility and privacy, adapt to diverse tabular datasets, and address metric blind spots (e.g., unmeasured outliers). Without such a framework, generating tabular synthetic data in compliance-sensitive finance domains including finance risks either underestimates disclosure threats or lowers too much data utility.

Challenge 3 The hypothesis that fraud scenario periodically exhibit a duration of specific minutes is grounded in evidence from financial economic studies [6, 7]. According to these studies, the trends are driven by a combination of factors such as economic conditions and technological advancements. For example, during periods of economic instability and the rapid adoption of new technologies in financial applications, consumers may be more vulnerable to financial fraud due to the desire to sympathize with fraud to earn money and fraudsters thus often align their fraudulent activities with these periods, targeting specific vulnerabilities that arise in such contexts [6].

Furthermore, we explicitly investigate the existence of the high-risk periods by using a financial fraud transactions (FFT) statistics [12]. Then, we observe there is a duration of specific minutes in which the intensity and frequency is exceptionally large, and it appears periodically. Two attributes of the statistics consists of frequency and intensity ratios across 30-minute intervals for 8 months, from March to September 2019. In addition, since the intensity (<0.15) and frequency (<0.001) have too different ranges of values, we normalize these indicators to visualize together. Therefore, we empirically observe high-risk periods (i.e., high fraud intensity and frequency) tends to recur periodically in financial transactions and the periods are not evenly distributed but appear repeatedly at intermittent intervals. Based on this rationale, we define the irregular temporal patterns involving fraudulent activities, named *the active lifetime*.

These active lifetimes are identified based on time slot (TS_i), wherein both the ratio of fraud intensity and frequency of fraudulent transactions exceeds predefined thresholds [13]. $I_{\text{threshold}}$ and $F_{\text{threshold}}$ are set to 1.2 and 4.0 based on the investigation in a previous work. Then, we define a set of active lifetime AL for a given set of time slots TS_n as follows: $AL = \{TS_j \mid I_{\text{fraud}}(TS_j) > I_{\text{threshold}}$ and $F_{\text{fraud}}(TS_j) > F_{\text{threshold}}\}$, where $I_{\text{fraud}}(TS_j)$ denotes the fraud intensity ratio at TS_j, representing the proportion of fraud amount to total transaction amount: $I_{\text{fraud}}(TS_j) = \frac{\text{fraud amount}_{TS_j}}{\text{total amount}_{TS_j}}$. $F_{\text{fraud}}(TS_j)$ denotes the fraud frequency ratio at TS_j, representing the ratio of fraud transactions to total transactions: $F_{\text{fraud}}(TS_j) = \frac{\text{fraud count}(TS_j)}{\text{total count}(TS_j)}$.

We shows that existing fraud detection models (XGBoost and TabNet) perform worse during active lifetimes compared to non-active periods, with higher false positive rates (FPR) and lower true positive rates (TPR) on The FFT statistics [30]. High FPR indicates the detection model falsely predicts legitimate transactions as fraud, causing unnecessary disruptions for customers. Low TPR indicates the detection model fails to predict many actual fraudulent transactions as fraud, greatly reducing the reliability of financial institutions.

[11]We adopt three utility metrics: pair-wise correlation, statistics, and pMSE-based score [20] which are most famous for measuring synthetic data utility. We also adopt two disclosure risk index: GU [28] and TCAP [29].

[12]https://www.fsec.or.kr/

[13]Based on the investigation in one of our previous works, this time slot is set to 30-minute intervals.

Challenge 4 Despite the widespread adoption of gradient boosting models such as XGBoost in real-world financial fraud detection, existing federated learning (FL) research predominantly focuses on neural network-based approaches and rarely addresses banking-specific operational constraints.

To account for heterogeneity of data volume across institutions, we apply a Dirichlet distribution with concentration parameter α to divide the datasets among institutions [14]. We also simulate two scenarios in which a client either randomly drops out during rounds 5–15 or joins during rounds 10–20. Performance decrease ratio is calculated as follows: Performance Decrease (%) $= \left(\frac{\text{Baseline} - \text{After Drop/Join}}{\text{Baseline}}\right) \times 100$. Bold and underlined values indicate the largest and second largest drops in performance, respectively. Our empirical analysis on four representative federated gradient boosting methods—SimFedXGB [31], FedXGBllr [32], FedXGBBagging [33], and FedXGBCyclic [34]—revealed two limitations:

- Vulnerability to data quantity skew: All methods suffered notable F1-score degradation under skewed data volume distributions across institutions (i.e., α = 0.1 Dirichlet split). This instability is likely due to unreliable gradient/Hessian statistics in clients (institutes) with small datasets, leading to suboptimal tree construction.
- Instability under participation dynamics: When simulating realistic scenarios where a bank drops out mid-training or a new bank joins, certain models (SimFedXGB, FedXGBCyclic) experienced severe performance drops (up to 19.33%), while others (FedXGBBagging, FedXGBllr) were more robust due to their ensemble aggregation mechanisms.

These findings underscore the need for fraud scenario-aware federated gradient boosting frameworks that can (1) adapt to heterogeneous data volumes and (2) maintain stability under dynamic client participation.

[14] In this experiment, we simulate collaboration among 5 financial institutions. We considered the number of major banks in real-world finance industry in South Korea.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present a real-world case study on privacy-preserving financial fraud detection in the South Korean financial industry. We identify four challenge and propose two solutions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have not yet developed a solution for C2 and C4 and described it as a current study in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not contribute a theoretical proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The source code is available through the previous works we mentioned in introduction. After the acceptance, they will be cited and a link will be given to use the source code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See above.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Experimental setting is explained to the extent possible in the page limitations. After acceptance, we will cite our previous works to help understand the results in detail.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We show statistical significance was confirmed through paired t-tests between our method and baselines for each metric and dataset (100 independent runs with p-values ≤ 0.0002).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe our computation resources in supplementary material A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed and confirmed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We explain positive societal impacts such as improving financial fraud detection accuracy and enabling secure collaboration among banks to support nationwide anti-fraud efforts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We do not use licenses for existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We used ChatGPT to only proofread sentences for grammar clarity. We then manually reviewed and verified these refinements.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.