

Synthetic Data Generation and Federated Learning as Innovative Solutions for Data Privacy in Finance

Elif Özcan^{1,2}, Ruşen Akkuş Halepmollası^{1,2} and Yusuf Yaslan¹

¹*Faculty of Computer and Informatics Engineering, Istanbul Technical University, Turkey*

²*TÜBİTAK Informatics and Information Security Research Center, Kocaeli, Turkey*

{elif.ozcan, rusen.halepmollasi}@tubitak.gov.tr; {ozcane22, halepmollasi, yyaslan}@itu.edu.tr

Keywords: Finance, Synthetic Data, Federated Learning, Artificial Intelligence.

Abstract: Financial services generate vast, complex and diverse datasets, yet data privacy issues pose significant challenges for secure usage and collaborative analysis. Synthetic data generation can offer an innovative solution while preserving privacy without exposing sensitive information. Also, federated learning enables collaborative model training across clients while maintaining data privacy. In this study, we used Default Credit Card dataset and employed diffusion based synthetic data generation to evaluate its impact on centralized and federated learning approaches. To this end, we offer comprehensive benchmarking of synthetic, real, and hybrid datasets by employing four machine learning classifiers both centrally and federated. Our findings demonstrate that synthetic data effectively improves results, especially when combined with real data. We also conduct client specific experiments in federated learning when addressing highly imbalanced or incomplete class distributions. Moreover, we evaluate FedF1 aggregation method, which aims to improve global model performance by optimizing F1-score. To the best of our knowledge, this is the first study to integrate synthetic data generation and federated learning on a financial dataset to provide valuable insights for secure and collaborative learning.


1 INTRODUCTION


Artificial Intelligence (AI) has been a transformative and innovative force in the financial sector, including banking, insurance, trading, risk management, and modern FinTech services (Cao, 2022). AI applications, particularly Machine Learning (ML) and Deep Learning (DL), are crucial for modeling the complex linear and nonlinear behaviors of financial variables to address problems beyond the scope of traditional models (Ahmed et al., 2022). Meanwhile, ML models trained on vast amounts of financial data can achieve higher scores in terms of evaluation metrics and enable more robust and efficient data driven decisions. Financial services generate vast, complex and diverse dataset; however, the sensitive and personally identifiable features of financial data create significant challenges and limitations for its usage and sharing (Assefa et al., 2020).


One promising solution to handle data privacy and security issues is synthetic data generation, which

mirrors the statistical properties and patterns of real data without sharing sensitive data (Lu et al., 2023). It aims to protect the privacy of customers due to laws such as General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019) and Health Insurance Portability and Accountability Act (HIPAA) (Cohen and Mello, 2018). Also, sharing realistic synthetic data between institutions and within research community allows training ML models on privacy-compliant datasets and enables the development of effective solutions to technical challenges. Moreover, synthetic data can address the lack of historical data for certain events to provide counterfactual data for testing strategies, as well as can handle class imbalances in datasets to improve the performance of ML models, particularly in cases like fraud detection (Assefa et al., 2020).

Federated Learning (FL) emerges as another innovative approach to address privacy concerns in financial data analysis by enabling multiple institutions to collaboratively train ML models without the need to share sensitive or raw data (Yang et al., 2019). Data remains securely within the institutions' premises, and only model updates are shared. It ensures that

^a <https://orcid.org/0009-0002-3423-131X>

^b <https://orcid.org/0000-0002-9941-2712>

^c <https://orcid.org/0000-0001-8038-948X>

privacy is maintained, as the underlying financial data never leaves the organization, thus complying with regulatory constraints such as GDPR (Truong et al., 2021). FL also facilitates data collaboration across institutions and allows them to leverage diverse datasets to improve model results without violating privacy policies (Mothukuri et al., 2021).

Considering the aforementioned data privacy and security issues, in this study, we employed two innovative approaches: (i) Synthetic data generation and (ii) FL. We leveraged a diffusion model for synthetic data generation and explored its impact on both centralized and FL approaches across several ML algorithms, including Logistic Regression (LR), Support Vector Classifier (SVC), Stochastic Gradient Descent Classifier (SGDC), and Multi Layer Perceptron (MLP). We benchmark and evaluate the real, synthetic and hybrid data in centralized and FL approaches under six distinct experimental scenarios. Additionally, we conducted two case studies to analyze specific challenges. Case Study 1 focused on the impact of FL and synthetic data at client level and Case Study 2 focused on addressing highly imbalanced and incomplete class distributions. In this context, our contributions are as follows:

- We present comprehensive benchmarking of synthetic, real and hybrid data in both centralized and FL environments. To the best of our knowledge, this is the first study to comprehensively investigate the integration of synthetic data generation and FL approach using Default Credit Card dataset to address critical issues in data privacy, accessibility, and class imbalance.
- We introduce a client-level analysis in FL to investigate whether it improves model outputs, particularly in scenarios with imbalanced or incomplete class distributions.
- We evaluate a novel FedF1 aggregation method (Aktaş et al., 2024) to optimize global model performance in FL to explore its ability to handle heterogeneity and imbalance clients.

Our contributions provide a robust framework for integrating synthetic data generation and FL approach in financial applications to address data privacy, security and accessibility issues.

Structure of the Paper. Section 2 summarizes previous works on synthetic data generation and FL approaches in finance. In Section 3, we present the methodology. Section 4 describes the dataset and experimental setup of case studies. In Section 5, the results of the study are reported and discussed. Section 6 concludes the paper and offers future work.

2 LITERATURE REVIEW

In this section, we review existing literature on synthetic data generation and FL in the financial domain. This review is organized into two subsections: the role of synthetic data in financial applications and advancements in federated learning for finance.

2.1 Synthetic Data Generation in Finance

Synthetic data generation plays a crucial role in financial applications by addressing various issues such as data scarcity, class imbalance, and privacy constraints. Khaled et al. (Khaled et al., 2024) explored the use of synthetic data to improve ML models for credit card fraud detection. Authors employed the SMOTE technique to address the severe class imbalance in financial datasets, where fraudulent transactions are significantly underrepresented. By training ML models on the generated synthetic data, they observed notable improvements in accuracy and recall, particularly in detecting minority class detection. This research underscores the potential of synthetic data to mitigate data imbalance challenges and improve the performance of fraud detection models in the financial sector.

Building on the promise of synthetic data, Jolicoeur-Martineau et al. (Jolicoeur-Martineau et al., 2023) proposed a novel framework that integrates score-based diffusion models with conditional flow matching for tabular data generation and imputation by using XGBoost. That approach is specifically designed to handle mixed-type tabular data, including both categorical and numerical features, a common challenge in tabular data modeling. Through extensive experimentation on 27 datasets from diverse domains, the method demonstrated superior performance in data generation tasks compared to state-of-the-art DL-based generative models while maintaining competitive results in data imputation scenarios. Additionally, key advantage of the proposed approach is its efficiency, as it can leverage parallel CPU training and bypass the need for computationally expensive GPUs. This work highlights the potential of combining advanced generative modeling techniques with traditional ML algorithms to effectively address tabular data challenges.

Furthermore, Sattarov et al. (Sattarov et al., 2023) introduced FinDiff, a novel diffusion-based model specifically designed to generate synthetic tabular data in the financial domain. The model addresses the challenges associated with mixed-type data, such as the coexistence of numerical and categorical fea-

tures. FinDiff was rigorously evaluated on three real-world financial datasets and focused on regulatory tasks including economic scenario modeling, stress testing, and fraud detection—key applications in finance where data availability and privacy are critical concerns. Their experimental results demonstrated that FinDiff can preserve the statistical properties of the original dataset and generate high-fidelity synthetic data while ensuring utility and privacy. Thus, the model offers a robust solution to data-sharing challenges in the financial industry. Moreover, authors highlighted the versatility of FinDiff in supporting downstream ML tasks and showing competitive performance compared to traditional methods. FinDiff not only enhances data accessibility but also aligns with the regulatory requirements of the financial sector. Therefore, it can be a valuable tool for FL applications.

2.2 Federated Learning in Finance

FL has received significant attention from researchers and practitioners as it enables collaborative model training without sharing sensitive data (Ülver et al., 2023)(Zhang et al., 2021)(Yurtoğlu et al., 2024). Wang et al. (Wang et al., 2024) proposed Federated Knowledge Transfer (FedKT), a FL approach developed for credit scoring while preserving data privacy. The approach enables collaboration among financial institutions without sharing raw data to address privacy concerns in credit scoring. A key challenge in FL is the heterogeneity of data distributions across participants, which can hinder the learning capacity of the global model. For this purpose, FedKT combines fine-tuning and knowledge distillation techniques to effectively extract general knowledge from the global model's early layers and task-specific knowledge from its outputs. Experimental evaluations on four distinct credit datasets demonstrated that FedKT outperforms existing FL algorithms in terms of predictive performance and robustness. Its ability to balance privacy preservation with high model performance makes it particularly valuable in the financial sector, where data sensitivity and regulatory compliance are critical.

In addition to privacy concerns, data imbalance poses a significant challenge in FL environments. Zhang et al. (Zhang et al., 2024) explored the challenges posed by data imbalance in FL for credit risk forecasting, a critical task in financial decision-making. They analyzed the performance of three ML models—Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), and eXtreme Gradient Boosting (XGBoost)—across multiple datasets with vary-

ing client numbers and data distribution patterns. They achieved an average performance improvement of 17.92% and their findings revealed that FL models significantly outperformed local models for non-dominant clients with smaller, highly imbalanced datasets. However, for dominant clients with larger datasets, FL models offered no clear advantage over local models, thus, authors highlighted potential disincentives for their participation. The study emphasized the need for strategies to mitigate the effects of data imbalance and ensure equitable benefits for all participants in FL environments.

Trust and interpretability are also critical for the adoption of FL in finance. Awosika et al. (Awosika et al., 2023) introduced a novel approach that combines FL and eXplainable Artificial Intelligence (XAI) to enhance financial fraud detection systems. FL enables multiple financial institutions to collaboratively train a shared fraud detection model without exchanging sensitive customer data, thereby upholding data privacy and confidentiality. The integration of XAI ensures that the model's predictions are interpretable by human experts and fosters transparency and trust in the system. Authors conducted experiments on realistic transaction datasets and demonstrated that the FL-based fraud detection system consistently achieved high performance metrics. They underscored FL's potential as an effective and privacy-preserving tool in combating financial fraud.

3 METHODOLOGY

In this section, we provide our methodology that involves synthetic data generation, FL approach and ML algorithms.

3.1 Synthetic Data Generation

In this study, we follow synthetic data production procedure presented in FinDiff: Diffusion Models for Financial Tabular Data Production (Sattarov et al., 2023). For this purpose, we used Gaussian Diffusion Models to generate synthetic data customized to mixed-type tabular datasets, which are common in financial applications. The methodology is intended to overcome the issues of working with heterogeneous data that contains both numerical and category variables.

Gaussian Diffusion Models operate by gradually transforming data distributions through a two-step process. In the forward diffusion phase, Gaussian noise is incrementally added to the original data,

effectively smoothing its complex structure into a noise-dominated state. Thus, the learning of high-dimensional relationships within the data is facilitated. In the reverse diffusion phase, noise is systematically removed and reconstructed synthetic samples that approximate the original data distribution. A learned score function guides the reverse process to ensure that the generated data aligns closely with the original dataset’s statistical and structural properties (Ho et al., 2020)(Sohl-Dickstein et al., 2015).

In line with the methodology presented in (Sattarov et al., 2023), we also evaluated the quality and utility of the generated synthetic data using several metrics, including fidelity, utility, synthesis, and privacy. *Fidelity* measures how well the synthetic data replicated the statistical properties of the original data, both at the column level (individual features) and row level (holistic data structures). *Utility* evaluates the ability of synthetic data to support downstream ML tasks such as fraud detection and credit scoring. *Synthesis* ensures that the generated data maintained structural alignment with the original dataset. *Privacy* assesses resistance to privacy attacks such as membership inference.

3.2 Federated Learning

FL is a decentralized ML approach designed for training models collaboratively across multiple clients while preserving data privacy. Unlike traditional centralized approaches, where data is collected and processed on a central server, FL ensures that data remains on the clients. Only model updates, such as gradients and weights, are shared with a central server, in which the model parameters are aggregated to create a global model. As shown in the Figure 1, each client trains the model locally on its own dataset to create a global model without sharing the data and ensure that sensitive data never leaves the clients’ environment. After local training, each client sends only the model parameters to the server that aggregates the weights from all clients to update the global model. The updated global model is shared with all clients, and the process is repeated for several iterations until the model converges.

In this study, we employed two aggregation methods. The first method, namely FedAvg, computes a weighted average of model updates from participating clients based on their dataset sizes to ensure proportional contribution to the global model (McMahan et al., 2017). The second method, namely FedF1, aggregates model updates by assigning weights based on the clients’ F1-scores (Aktaş et al., 2024). FedF1 prioritizes contributions from clients with higher F1-

scores to reflect more stable and accurate local models. In other words, by using F1-scores as a weighting factor, it aims to improve the overall performance and reliability of the global model, particularly in scenarios with imbalanced datasets or heterogeneous client performance.

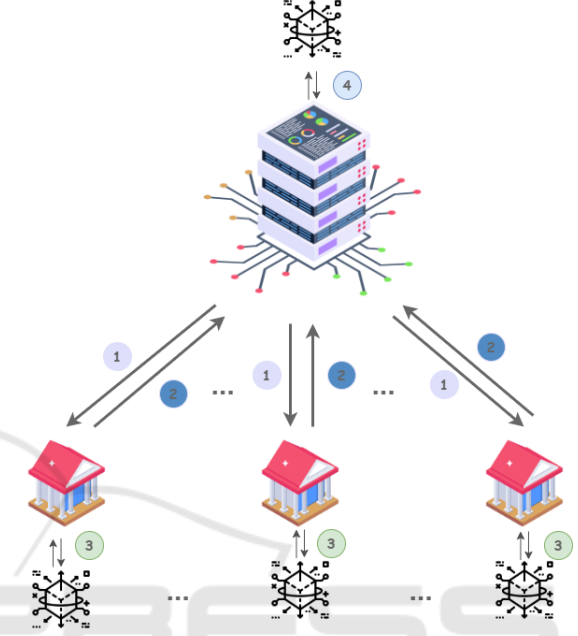


Figure 1: Synthetic Data Generation.

3.3 Machine Learning Algorithms

When comparing the both synthetic–real data and central–decentral approaches, we examined the feasibility of using four classifiers, namely LR, SVC, SGDC and MLP. The employed classifiers are, in short, described below:

Logistic Regression. (Hosmer et al., 2013) models the probability of a binary class label using the sigmoid function by transforming a linear combination of input features into a probability score. It is a parametric and discriminative method and focuses on the direct mapping between features (independent variables) and class labels (dependent variables). In this study, we conducted hyperparameter tuning using the training and validation datasets, resulting in the best parameters: $C = 0.01$, $max_iter = 5000$, $penalty = l1$, $solver = saga$, and $class_weight = balanced$.

Support Vector Classifier. (Cortes and Vapnik, 1995) constructs a hyperplane that separates classes in the feature space with maximum margin. In this study, we employed a *linear kernel* to model linearly

separable data. Hyperparameter tuning focused on improving the model's handling of class imbalances and convergence properties, with the optimal parameters identified as $C = 0.1$, $max_iter = 1000$, and $class_weight = balanced$.

Stochastic Gradient Descent Classifier. (Bottou, 2010) is a linear classifier that leverages Stochastic Gradient Descent for optimization. It iteratively updates the model parameters by computing the gradient of the loss function with respect to a single training example to make it highly efficient for large-scale and sparse datasets. In this study, we fine-tuned the hyperparameters, and determined the optimal configuration as $alpha = 0.1$ (regularization term), $max_iter = 5000$ (maximum number of iterations), $penalty = elasticnet$ (combination of L1 and L2 regularization), and $l1_ratio = 0.5$ (balance between L1 and L2 regularization).

Multi-Layer Perceptron. (Goodfellow et al., 2016) is a feedforward neural network that captures non-linear relationships between input features and target labels using multiple layers of neurons. Training is performed using backpropagation to optimize the weights of the network. The hyperparameter tuning process determined the best configuration as $activation = relu$, $alpha = 0.001$, $hidden_layer_sizes = (50,)$, $learning_rate = adaptive$, $max_iter = 200$, and $solver = adam$.

4 EXPERIMENTAL SETUP

In this section, we present the details of the dataset we utilized and the experiments we conducted, including the training configuration, evaluation scenarios, case studies, and evaluation metrics. The overall flow of the experimental setup is shown in Figure 2.

4.1 Dataset

In this study, we used Default of Credit Card Clients (DCCC) dataset (Yeh and Lien, 2009), obtained from the UCI Machine Learning Repository. DCCC dataset includes 30,000 records of credit card clients in Taiwan and includes both categorical and numerical features. It provides a comprehensive set of attributes, including demographic information, payment history, bill statements, and a default payment indicator. It is complete, with no missing values, and the target variable is binary. The dataset has a class imbalance, with class ratio of 3:1.

4.2 Training Configuration

We randomly split 90% of the data for training and 10% for testing. Random splitting of the dataset can lead to significant variations in the target variable ratios, which may impact model performance, especially since our dataset is imbalanced with a small number of samples in the default class. To address this issue, we ensured that the data split was performed with stratification. To obtain more reliable results, we repeated the experiments 10 times, each with a different random seed to shuffle the order of the samples, and calculated the average performance scores across all runs.

We utilized the training data to generate synthetic data from the training set. For this purpose, we employed the diffusion model, as detailed in Section 3. Moreover, we partitioned the training data equally into five subsets, representing five distinct clients for FL setup. We built federated models using the real train set distributed across those clients and evaluated models' performance using the real test set. When exploring the potential of synthetic data in FL, we independently applied the same diffusion model to each client to generate client-specific synthetic data. Please note that we generated synthetic data for each client using only their local data, as clients in real-world FL scenarios cannot access to each other's data. Overall, to evaluate the impact of data augmentation on model performance, we trained federated models on three data types: real data, synthetic data and hybrid data which is a combination of real and synthetic data. Furthermore, to compare the FL approach with the centralized approach, we used the same train-test configurations when building ML models centrally. Similar to the FL approach, we trained centralized models on three data types: real data, synthetic data and hybrid data.

Table 1 provides a detailed summary of the data distribution across centralized and federated setups to highlight the class distribution in each subset. In all scenarios, we used only the test set split from the real data for evaluation. In other words, model testing across all configurations was performed on the real test set. Thus, we ensured consistency and comparability across centralized and federated setups, as well as across models trained on real, synthetic, and combined datasets.

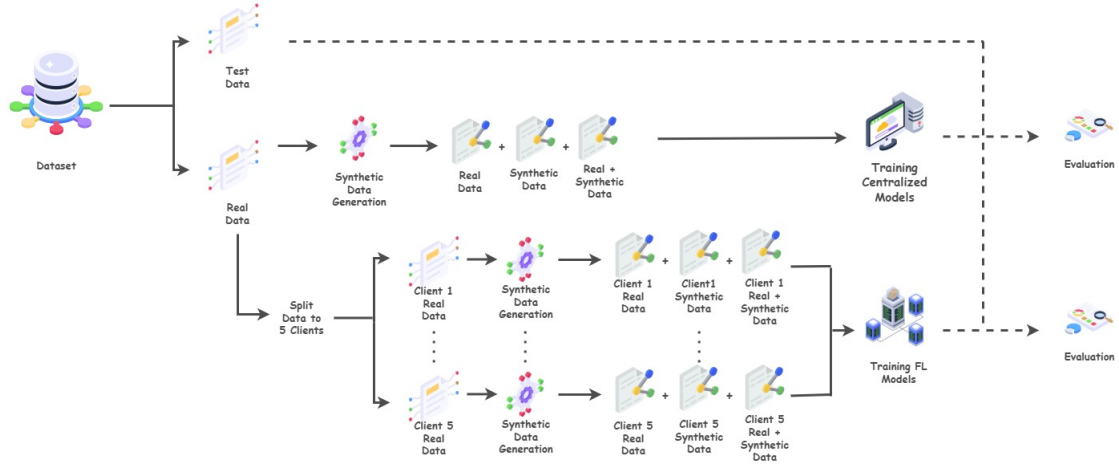


Figure 2: Illustration of proposed experimental setup.

Table 1: Data distribution across clients and test set.

	# Samples Class 0	# Samples Class 1	Class 1 Proportion (%)
Central Data	21023	5977	22.1
Client 1	4229	1171	21.6
Client 2	4209	1191	22.0
Client 3	4218	1182	21.8
Client 4	4187	1213	22.4
Client 5	4180	1220	22.5
Test Data	2341	659	21.9

4.3 Benchmarking: Evaluating Synthetic Data in Centralized and FL Approaches

To investigate the impact of synthetic data on FL, we conducted a comprehensive analysis using four algorithms: Logistic Regression (LR), Support Vector Classifier (SVC), Stochastic Gradient Descent Classifier (SGDClassifier) and Multi-Layer Perceptron (MLP). Our goal is to evaluate the performance of both centralized and federated models on varying data types and provide insights into how synthetic data influences learning outcomes. Also, we designed the experiments around six distinct scenarios:

- **Central+Real Data.** In this scenario, centralized models are trained solely on the real data. The objective is to assess the baseline performance of centralized models without the influence of synthetic data.
- **Central+Synthetic Data.** In this scenario, centralized models are trained using only synthetic data generated from the training set. This allows us to evaluate the impact of synthetic data in a centralized learning environment by comparing performance with the real data scenario.

- **Central+Hybrid Data.** This scenario involves centralized models trained on a combination of real and synthetic data. The goal is to assess the effectiveness of data augmentation.
- **FL+Real Data.** In this scenario, federated models are trained using only real data distributed across clients. This scenario provides a baseline for FL performance with real data.
- **FL+Synthetic Data.** In this scenario, federated models are trained using synthetic data generated for each client. The purpose is to explore the potential of synthetic data in a FL environment and evaluate how it influences model performance compared to real data.
- **FL+Hybrid Data.** This scenario involves federated models trained on a combination of real and synthetic data. By incorporating both types of data, the setup evaluates the impact of data augmentation on FL performance, similar to the hybrid data scenario in centralized models.

4.4 Case Study 1: Evaluating FL Performance with Synthetic Data

To further explore the benefits of FL at the client level, we conducted an additional case study focusing on three selected clients (Client 1, Client 3, and Client 5). We evaluated their performance under various configurations to gain a deeper understanding of the impact of FL and the use of synthetic data at the client level. We utilized SVC and MLP based on the benchmarking results of which MLP achieved the highest accuracy scores and SVC reached the best F1-scores.

For each client, we performed experiments using locally trained centralized models on three different data types: real data, synthetic data, and hy-

brid data. We trained the models on the individual client's data to represent a baseline for local training without federated collaboration. In FL setup, we conducted each model training over 10 rounds using the FedAvg aggregation method. We evaluated two distinct model types at the end of the training process to assess the impact of global collaboration and client-specific fine-tuning:

- *Global Federated Model:* This model represents the aggregated global model produced by the server after the 10th communication round. It reflects the combined knowledge learned from all participating clients.
- *Client-Adapted Federated Model:* This model is derived from the global federated model after the 10th round but is further fine-tuned locally on each client's own data.

With the above model types, we aim to explore, in detail, the performance improvements that FL can bring to local clients when data centralization is not feasible due to privacy concerns, regulatory restrictions, or operational constraints.

4.5 Case Study 2: Addressing Imbalanced Class Distribution

In this case study, we investigated the potential of FL to address a scenario where clients have highly imbalanced or incomplete class distributions. To simulate such a scenario, we created three clients: one client with no samples labeled as class 1 and two clients with balanced class distributions. The setup reflects real-world situations, such as a bank branch with no recorded fraud cases, while other branches have sufficient data for both classes. The data distribution among the clients and the test set is summarized in Table 2.

Table 2: Data distribution across clients and test set for Case Study 2.

	Class 0	Class 1
Client 1	5976	0
Client 2	2988	2988
Client 3	2988	2988
Test Data	659	659

When the first client trains a local model independently, it has no knowledge of class 1 due to the absence of positive samples on its dataset. Therefore, its local model can be incapable of predicting the positive class. On the other hand, by participating in FL, that client can leverage knowledge aggregated from other clients and gain access to information about

class 1 without sharing its raw data, thus preserving data privacy.

Additionally, we explored the limitations of synthetic data generation in this context. Even if synthetic data were generated for the first client, the absence of positive samples would prevent the generation of meaningful data for class 1. Thus, it highlights a critical scenario where FL provides a unique advantage over local models and synthetic data augmentation.

For aggregation in this study, we employed the FedF1 method, designed to optimize the federated model for imbalanced data scenarios. By focusing on F1-score optimization during aggregation, FedF1 ensures that the global model performs effectively across clients with differing class distributions.

5 ANALYSIS OF THE RESULTS

In this study, we evaluate the performance of synthetic data in centralized and FL approaches using four ML models, including LR, SVC, SGDC, and MLP. In Case Study 1, we investigate FL performance with synthetic data. With Case Study 2, we address imbalanced class distribution using FL approach. We evaluated the synthetic data generated for centralized training to assess its similarity to the real data. We also compared feature distributions and inter-feature relationships between the real and synthetic datasets.

Figure 3 shows the probability distributions of selected features (e.g., Gender, Pay0, Age, and Limitbal) and illustrate a close match between synthetic dataset and real dataset. Furthermore, Figure 4 presents the column pair trends and correlations and also demonstrate strong consistency across both datasets. The evaluation using the SDV framework provided additional metrics to quantify the quality of the synthetic data as follows:

- *Column Shapes Score:* 92.27%
- *Column Pair Trends Score:* 77.29%
- *Overall Quality Score:* 84.78%

In the subsections, we discuss, in detail, the effect of the different data types, FL and centralized approaches, selected ML algorithms, and imbalanced class distribution.

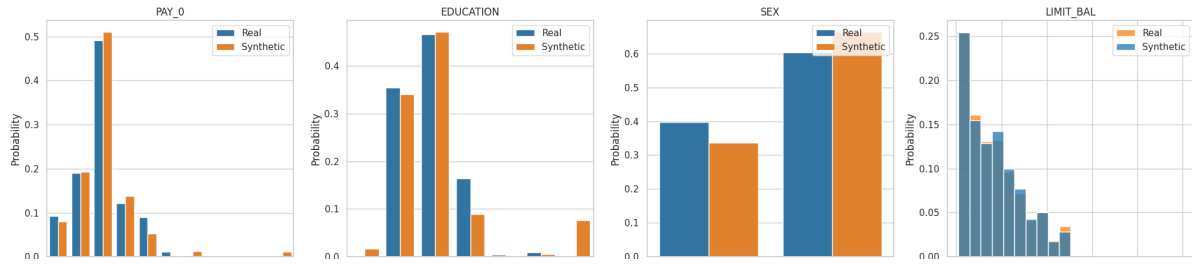


Figure 3: Probability distributions of selected features for real and synthetic data.

Table 3: Comparative analysis of different algorithms for Central and FL models.

	SGD				SVC				LR				MLP			
	F1	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1	Acc	Recall	Prec
Central - Real Data	0.4926	0.6320	0.55	0.56	0.5977	0.6540	0.64	0.60	0.5699	0.6913	0.57	0.56	0.4922	0.7713	0.52	0.57
Central - Synthetic Data	0.4852	0.5804	0.55	0.55	0.6386	0.7190	0.66	0.63	0.5797	0.7003	0.58	0.57	0.4548	0.7723	0.50	0.53
Central - Hybrid Data	0.4939	0.6080	0.55	0.58	0.6335	0.7057	0.66	0.62	0.5772	0.6960	0.58	0.57	0.4667	0.7795	0.51	0.62
FL - Real Data	0.4573	0.5589	0.53	0.54	0.6037	0.6637	0.65	0.60	0.5695	0.6943	0.57	0.56	0.3944	0.5684	0.48	0.49
FL - Synthetic Data	0.3829	0.4561	0.52	0.48	0.6674	0.7980	0.65	0.70	0.5532	0.6913	0.55	0.55	0.3954	0.5055	0.49	0.49
FL - Hybrid Data	0.4009	0.4601	0.53	0.54	0.6447	0.7370	0.66	0.63	0.5568	0.6960	0.56	0.55	0.3042	0.3446	0.48	0.51

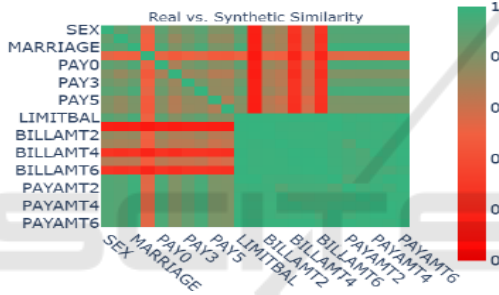


Figure 4: Column pair trends for real and synthetic data.

5.1 Benchmarking: Evaluating Synthetic Data in Centralized and FL Approaches

In this subsection, we compare the performance of centralized and FL approaches using real data, synthetic data and hybrid data.

The results presented highlight an insight into the effectiveness of synthetic data in ML models, both in centralized and FL frameworks. One of the primary goals of this study was to evaluate whether models trained on synthetic data could achieve performance comparable to, or even exceed, those trained on real data. From Table 3, we observe that while different algorithms show varying levels of performance, the use of synthetic data produces results that are comparable to—and occasionally better than—those obtained using real data. Figure 5 visually illustrates this observation, showcasing the performance distribution of centralized and FL models across different data configurations. The violin plots clearly demonstrate that synthetic data yields performance distribu-

tions comparable to those of real data, further highlighting its effectiveness in both centralized and FL settings. For instance, in the centralized models, the SVC algorithm achieved an F1-score of 0.6386 with synthetic data, outperforming the corresponding F1-score of 0.5977 when trained on real data. A similar trend is observed in FL models, where the synthetic data-based SVC achieved an F1-score of 0.6674, exceeding the F1-score of 0.6037 obtained with real data.

To delve deeper into the comparative performance of the datasets, Table 4 provides a focused analysis of F1-scores across all experimental setups. This table shows that in both centralized and FL models, synthetic data consistently performs competitively. Although hybrid datasets often improve model performance compared to real data (e.g., the F1-score of 0.4939 for the SGD method in centralized settings or the F1-score of 0.6447 for the SVC algorithm in FL), the gains over synthetic data are typically insignificant. This indicates that using synthetic data alone is usually enough to produce competitive results, even though hybrid data may be useful in certain situations. In many instances, hybrid data perform similarly to synthetic data, suggesting that the additional complexity of integrating real data may not always be necessary.

The results underline that the quality of model predictions is unaffected by the use of synthetic data. In contrast, synthetic data's competitive performance shows that it can potentially ease privacy issues, since using the synthetic data provides similar results without compromising the privacy of sensitive data in both centralized and FL environments. The success of synthetic data in achieving comparable or superior per-

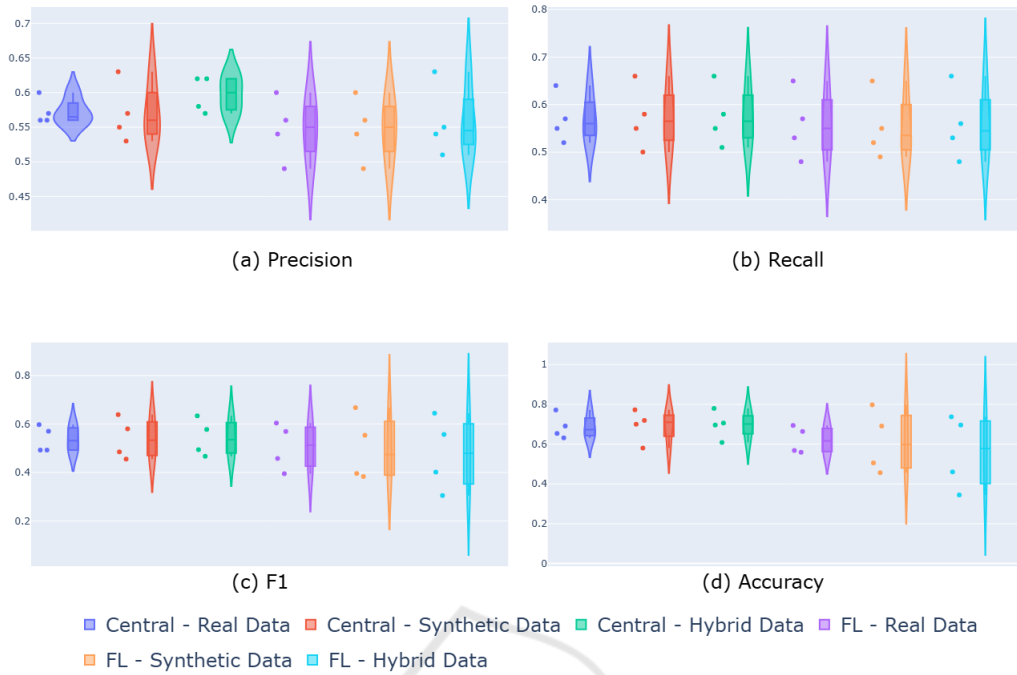


Figure 5: Performance distribution of centralized and federated models across different data configurations.

Table 4: F1-Scores Across Centralized and Federated Learning Models Using Real, Synthetic, and Hybrid Data

	Centralized Models			FL Models		
	Real Data	Synthetic Data	Hybrid Data	Real Data	Synthetic Data	Hybrid Data
SGD	0.4926	0.4852	0.4939	0.4573	0.3829	0.4009
SVC	0.5977	0.6386	0.6335	0.6037	0.6674	0.6447
LR	0.5699	0.5797	0.5772	0.5695	0.5532	0.5568
MLP	0.4922	0.4548	0.4667	0.3944	0.3954	0.3042

formance reaffirms its value as an alternative, especially in domains where real data is limited, sensitive, or inaccessible.

5.2 Case Study 1: Evaluating FL Performance with Synthetic Data

In this subsection, we compare the performance of centralized and FL approaches across different clients for FL global models, as well as fine-tuned FL models.

The results presented in Table 5 and Table 6 provide insights into the client-level impact of FL models compared to centralized models. From the results, it is evident that FL models generally achieve performance that is comparable to, or even exceeds, that of centralized models across most clients. For instance, in Table 5, the FL global model for SVC achieved an F1-score of 0.6649 for Client 2, outperforming the centralized model's F1-score of 0.5603. Similarly, in Table 6, for the MLP model, the FL model achieved an F1-score of 0.4673 for Client 2, which is compara-

ble to the centralized model's F1-score of 0.4673. The results show the effectiveness of FL in maintaining or improving performance at the client level, ensuring that models trained in a decentralized manner are capable of matching the results of centralized training.

5.3 Case Study 2: Addressing Imbalanced Class Distributions

In this case study we focused on the limitations of synthetic data in addressing class imbalance, particularly when certain classes are entirely absent from a client's local dataset. In such scenarios, synthetic data generation alone fails to resolve the issue, as it relies solely on the local data distribution and cannot create representations for missing classes. FL, however, overcomes this limitation by aggregating knowledge from multiple clients, enabling the global model to learn from distributed datasets where the missing class is present.

Table 7 highlights the performance of centralized models and FL strategies, FedAvg and FedF1. The FL

Table 5: Client-Specific Evaluation of SVC in Centralized and Federated Learning Models.

	Client 1				Client 2				Client 3			
	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
Central - Real Data	0.6016	0.6428	0.5939	0.6510	0.5868	0.6261	0.5603	0.6020	0.6263	0.6723	0.6275	0.6893
FL - Real Data	0.6016	0.6428	0.5939	0.6510	0.5868	0.6261	0.5603	0.6020	0.6263	0.6723	0.6275	0.6893
FL Global - Real Data	0.6078	0.6498	0.6037	0.6637	0.6078	0.6498	0.6037	0.6637	0.6078	0.6499	0.6037	0.6637
Central - Synthetic Data	0.6792	0.6431	0.6558	0.7867	0.5948	0.6188	0.5988	0.6867	0.7117	0.6099	0.6270	0.8003
FL - Synthetic Data	0.6792	0.6436	0.6576	0.7867	0.5948	0.6188	0.5988	0.6867	0.7117	0.6099	0.6270	0.8003
FL Global - Synthetic Data	0.6996	0.6514	0.6673	0.7980	0.6996	0.6514	0.6673	0.7980	0.6996	0.6514	0.6673	0.7980
Central - Hybrid Data	0.6215	0.6546	0.6277	0.7060	0.5761	0.6092	0.5579	0.6097	0.6895	0.6594	0.6710	0.7917
FL - Hybrid Data	0.6215	0.6546	0.6277	0.7060	0.5761	0.6092	0.5579	0.6097	0.6895	0.6594	0.6709	0.7917
FL Global - Hybrid Data	0.6373	0.6581	0.6447	0.7370	0.6373	0.6581	0.6447	0.7370	0.6373	0.6581	0.6447	0.7370

Table 6: Client-Specific Evaluation of MLP in Centralized and Federated Learning Models.

	Client 1				Client 2				Client 3			
	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
Central - Real Data	0.5190	0.5035	0.4679	0.7645	0.4982	0.5036	0.4673	0.7574	0.5240	0.5107	0.4850	0.7549
FL - Real Data	0.5406	0.5286	0.5248	0.7030	0.5557	0.5378	0.5286	0.7299	0.5588	0.5617	0.5380	0.6671
FL Global - Real Data	0.4922	0.4833	0.3944	0.5684	0.4922	0.4833	0.3944	0.5684	0.4922	0.4833	0.3944	0.5684
Central - Synthetic Data	0.4546	0.4974	0.4461	0.7687	0.4894	0.5014	0.4595	0.7644	0.4510	0.4903	0.4478	0.7524
FL - Synthetic Data	0.4941	0.4988	0.4688	0.7433	0.5094	0.5130	0.4904	0.7067	0.4985	0.5031	0.4770	0.7266
FL Global - Synthetic Data	0.4871	0.4920	0.3954	0.5055	0.4871	0.4920	0.3954	0.5055	0.4871	0.4920	0.3954	0.5055
Central - Hybrid Data	0.5414	0.5044	0.4631	0.7689	0.4997	0.5035	0.4634	0.7643	0.4828	0.4972	0.4632	0.7515
FL - Hybrid Data	0.5607	0.5496	0.5271	0.6910	0.5415	0.5265	0.4974	0.6906	0.5580	0.5752	0.5482	0.6348
FL Global - Hybrid Data	0.5093	0.4869	0.3042	0.3446	0.5093	0.4869	0.3042	0.3446	0.5093	0.4869	0.3042	0.3446

Table 7: Case 2 - Comparison of Centralized Learning and Federated Learning with Different Strategies.

	F1			Precision			Recall			Accuracy		
	Real	Synthetic	Hybrid	Real	Synthetic	Hybrid	Real	Synthetic	Hybrid	Real	Synthetic	Hybrid
Central	0.3333	0.3333	0.3333	0.7500	0.7500	0.7500	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
FL Avg	0.3333	0.3333	0.3333	0.7500	0.7500	0.7500	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
FL Avg Global	0.3891	0.3894	0.3544	0.5399	0.5656	0.5913	0.5013	0.5015	0.5035	0.50013	0.5001	0.5035
FL FedF1	0.3333	0.3333	0.3333	0.7500	0.7500	0.7500	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
FL Global FedF1	0.4000	0.4003	0.4118	0.4952	0.4894	0.5998	0.4996	0.4799	0.4937	0.4996	0.4799	0.4937

Global FedF1 model achieves an F1-score of 0.4000 for real data, outperforming the centralized model (F1-score: 0.3333). This demonstrates FL's ability to leverage data from other clients to predict underrepresented classes effectively, a capability that synthetic data alone cannot provide.

Moreover, the FedF1 strategy proves more effective than FedAvg by prioritizing F1-scores during aggregation, thereby enhancing the global model's ability to handle imbalanced data. For example, when using hybrid data (real + synthetic), the FL Global FedF1 model achieves a F1-score of 0.4118, compared to 0.3544 for FedAvg. These results underscore the critical role of FL in scenarios where local data distributions are severely imbalanced.

These results demonstrate that FL, particularly with the FedF1 strategy, effectively addresses class imbalance in decentralized environments. Unlike synthetic data generation, which is constrained by local data distributions, FL aggregates distributed knowledge across clients, enabling robust model training even in the absence of certain classes. This highlights FL's potential as a practical approach for scenarios where class imbalance cannot be resolved through conventional means.

6 FINAL REMARKS

This study highlights the potential of combining synthetic data and FL to address critical challenges in ML, such as data privacy, class imbalance, and decentralized learning. Synthetic data emerged as a reliable alternative to real data, consistently achieving comparable performance across both centralized and FL settings. Its effectiveness underscores its applicability for applications where privacy or data accessibility constraints make the use of real data impractical. These results are particularly relevant for applications in financial institutions, where data privacy regulations prevent direct data sharing between entities, the combination of synthetic data and FL enables collaborative learning without compromising sensitive information, ensuring that institutions can achieve performance comparable to centralized models without exposing their data.

At the same time, FL demonstrated its capacity to enhance model robustness by leveraging distributed knowledge across clients. This capability proved particularly crucial in scenarios where synthetic data alone was insufficient, such as when certain classes were entirely absent from a client's local dataset. By

integrating data from other clients, FL effectively mitigated these limitations, enabling the global model to address class imbalances and improve prediction accuracy. Compared to traditional centralized learning approaches, this combination not only preserves data privacy but also enhances model robustness by leveraging distributed knowledge, making it particularly effective in scenarios with class imbalances or missing labels.

The findings suggest that synthetic data and FL are not only complementary but also mutually reinforcing. Synthetic data provides the foundation for privacy-preserving ML, while FL extends this foundation to handle more complex challenges inherent in decentralized environments. Together, these approaches form a robust framework for developing high-performing and privacy-conscious ML models suitable for real-world applications.

For future work, there are several potential directions to build upon our current findings. First, the impact of alternative synthetic data generation methods could be examined, focusing on how different techniques influence model performance in both centralized and FL frameworks. Furthermore, expanding the scope of the study to include diverse datasets from various domains would help validate the robustness and applicability of the proposed approach. Another promising avenue involves testing more advanced classification algorithms to explore their potential for improving both predictive accuracy and generalization across heterogeneous environments. These directions would collectively contribute to a deeper understanding of the interplay between synthetic data and FL in addressing real-world ML challenges.

REFERENCES

- Ahmed, S., Alshater, M. M., El Ammari, A., and Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, 61:101646.
- Aktaş, M., Akkuş Halepmollası, R., and Töreyn, B. U. (2024). Enhancing credit risk assessment with federated learning through a comparative study. In *8th EAI International Conference on Robotic Sensor Networks*.
- Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., and Veloso, M. (2020). Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8.
- Awosika, T. et al. (2023). Transparency and privacy: The role of explainable ai and federated learning in financial fraud detection. *Journal of Financial Technology and Ethics*, 8(1):15–30.
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.
- Cao, L. (2022). Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3):1–38.
- Cohen, I. G. and Mello, M. M. (2018). Hipaa and protecting health information in the 21st century. *Jama*, 320(3):231–232.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Hoofnagle, C. J., Van Der Sloot, B., and Borgesius, F. Z. (2019). The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.
- Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley.
- Jolicoeur-Martineau, A. et al. (2023). Generating and imputing tabular data via diffusion and flow based gradient boosted trees. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Khaled, A. et al. (2024). Synthetic data generation and impact analysis of machine learning models for enhanced credit card fraud detection. *Journal of Artificial Intelligence and Applications*, 12(3):45–60.
- Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., and Wei, W. (2023). Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.
- Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640.
- Sattarov, E. et al. (2023). Findiff: Diffusion models for financial tabular data generation. *Financial Data Science Journal*, 9(2):75–90.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.
- Truong, N., Sun, K., Wang, S., Guitton, F., and Guo, Y. (2021). Privacy preservation in federated learning: An insightful survey from the gdpr perspective. *Computers & Security*, 110:102402.
- Ülver, B., Yurtoğlu, R. A., Dervişoğlu, H., Halepmollası, R., and Haklıdır, M. (2023). Federated learning in predicting heart disease. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

- Wang, H. et al. (2024). A novel federated learning approach with knowledge transfer for credit scoring. *IEEE Transactions on Neural Networks and Learning Systems*.
- Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19.
- Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480.
- Yurtoğlu, R. A., Dervişoğlu, H., Ülver, B., Halepmollası, R., and Hakkıdır, M. (2024). A novel transformation through digital twin and federated learning integration: A case study on cardiovascular disease prediction. In *International Conference on Information and Communication Technologies for Ageing Well and e-Health*, pages 91–113. Springer.
- Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., and Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, 216:106775.
- Zhang, L. et al. (2024). The effects of data imbalance under a federated learning approach for credit risk forecasting. *International Journal of Data Mining and Analytics*, 16(4):230–245.

