Cost-Aligned AI for Fraud Detection: Federated Learning and Threshold Mapping under Non-IID Data*

Kun-Lin Tsai†, Chu-Wei Chu, Fang-Yie Leu, Chao-Tung Yang

TungHai University, Taichung, Taiwan
{kltsai, g12360025, leufy, ctyang}@thu.edu.tw

Abstract

Fraud detection in financial services faces two key challenges in federated learning. First, data across banks or branches are typically non-independent and identically distributed (non-IID), which destabilizes global model convergence and generalization. Second, most studies evaluate models only with curve-based metrics such as AUPRC or F1, without aligning decision thresholds to operationally tolerable false positives per day (FP/day), often leading to false alarm surges after deployment. To address these issues, in this paper, we propose a cost-aligned AI framework that integrates threshold mapping and federated evaluation under non-IID conditions. Model probabilities are mapped to deployable decision thresholds τ based on FP/day constraints, using both expected and conservative formulations. Out-of-time validation further ensures robustness against false positive rebound. Systematic experiments compare three aggregation strategies, i.e., FedAvg, FedProx, and FedBN, and multiple models, i.e., Logistic Regression, Random Forest, and MLP variants, across metrics including AUPRC, F2@τ^*, Precision/Recall@τ, FP/day, convergence speed, and communication cost. The proposed pipeline directly bridges model evaluation with operational capacity, providing a practical benchmark for aggregation and model selection in non-IID federated fraud detection.

Keywords: Federated Learning, Fraud Detection, Non-IID Data, Decision Threshold Mapping, Model Aggregation

1 Introduction

Fraud detection in financial transactions is often constrained by privacy and regulatory requirements that prohibit the centralization of sensitive data. Federated Learning (FL) has therefore emerged as a viable paradigm for collaborative model training without exposing raw data (McMahan, 2017, Kairouz, 2021). Nevertheless, fraud detection under FL faces unique challenges. Data across banks or branches are typically heterogeneous and non-independent identically distributed (non-IID), leading to unstable

*Proceedings of The 2025 IFIP WG 8.4 International Symposium on E-Business Information Systems Evolution (EBISION 2025), Article No. 9, December 16-18, 2025, Sapporo, Japan.

†Corresponding Author

global model convergence and limited generalization during both training and deployment (Li, 2020, Li, 2021, Zhao, 2018). These distributional shifts complicate the deployment of robust and reliable fraud detection models in real-world settings.

In practice, audit teams are further constrained by fixed daily review capacity. Conventional evaluation approaches, which rely primarily on curve-based metrics such as AUPRC and F1, do not directly guarantee that the deployed model can maintain a tolerable false positive rate (Davis, 2006, Saito, 2015). This mismatch between research metrics and operational constraints frequently results in a surge of false alarms after deployment, increasing audit workload and inflating operational costs. While prior studies have explored improved aggregation strategies or compared model families, few have formalized FP/day constraints into the threshold-selection process or validated models against temporal robustness requirements. Moreover, there is a lack of systematic benchmarking that jointly considers aggregation methods, model architectures, and operationally relevant metrics under non-IID conditions (Dal Pozzolo, 2018, Carcillo, 2018).

To address these gaps, this study introduces a cost-aligned evaluation and deployment pipeline for FL-based fraud detection. The proposed framework explicitly derives decision thresholds from FP/day constraints, ensuring that models remain within operationally sustainable limits. Both expected and conservative formulations of threshold mapping are considered. In addition, we incorporate temporal robustness validation, where thresholds determined from validation data are applied to out-of-time windows to monitor false positive rebound and assess stability of precision and recall over time.

Beyond the methodological framework, we conduct systematic benchmarking across three representative aggregation strategies, including FedAvg (McMahan, 2017), FedProx (Li, 2020), and FedBN (Li, 2021), and multiple model families, i.e., Logistic Regression, Random Forest, and MLP variants. The evaluation covers both academic indicators and operational metrics, with client-level slicing to reflect heterogeneous institutional capacities. Finally, we contribute an engineering design in the form of a modular interface that decouples aggregation strategies from model families, thereby enhancing reproducibility, extensibility, and practical deployment feasibility. Through this cost-aware and deployment-oriented pipeline, we move beyond purely curve-based evaluation and establish a methodology that directly bridges research outcomes with operational sustainability in non-IID federated environments.

The remainder of this paper is organized as follows. Section 2 reviews related studies on fraud detection, federated learning, and cost-aware evaluation frameworks. Section 3 introduces the proposed methodology, including the problem formulation, threshold mapping, dataset preparation, model architectures, and federated training strategies. Section 4 presents the experimental setup and results, providing comparative analysis across models, aggregation methods, and operational metrics. Finally, Section 5 concludes the paper and discusses potential directions for future research.

2 Related Studies

Federated Learning (FL) was first formalized by McMahan et al. with the FedAvg algorithm, which demonstrated that distributed stochastic gradient descent can achieve communication-efficient learning without centralizing raw data (McMahan, 2017). Subsequent research highlighted the challenges of heterogeneous client data distributions, leading to the development of FedProx (Li, 2020), which incorporates a proximal term to stabilize optimization under non-IID conditions. More recently, FedBN (Li, 2021) introduced local batch normalization to address feature distribution shifts, showing strong improvements in cross-silo FL scenarios. Empirical studies (Hsu, 2019, Zhao, 2018) confirmed that non-IID distributions significantly degrade global model convergence and generalization, underscoring the importance of specialized aggregation methods. Comprehensive surveys (Kairouz, 2021) provide further insights into algorithmic advances and open challenges in FL.

Recent work further extends FL to fraud detection applications. Alhasawi et al. (Alhasawi, 2025) proposed FedFraud, a federated framework tailored for scalable and trustworthy financial fraud detection, while Zheng et al. (Zheng, 2025) benchmarked federated architectures such as FedGAT-DCNN against FedAvg-DWA for credit card fraud detection. Tang et al. (Tang, 2024) demonstrated that graph-based federated models (FedGNN) effectively capture relational signals across institutions, highlighting the growing variety of model families adapted to non-IID fraud detection settings. Wang et al. (Wang, 2025) further analyzed the systematic impact of data heterogeneity in FL for credit card fraud, providing empirical justification for evaluating federated strategies under different Dirichlet partitions.

Traditional evaluation of binary classifiers relies on ROC analysis (Fawcett, 2006) and its relationship with Precision–Recall (PR) curves (Davis and Goadrich, 2006). However, PR curves are more informative than ROC under class-imbalanced conditions, as shown in (Saito, 2015). Beyond curve-based metrics, cost-sensitive evaluation frameworks have been proposed to align model assessment with operational constraints. Cost curves (Drummond, 2006) and the theoretical foundations of cost-sensitive learning (Elkan, 2001) emphasize the need to incorporate application-specific costs directly into threshold selection. These works motivate our approach of mapping decision thresholds from FP/day constraints, bridging the gap between statistical evaluation and real-world auditing capacity.

Credit card fraud detection is a canonical example of highly imbalanced classification, where fraudulent transactions represent less than 0.2% of all records. Dal Pozzolo et al. explored calibration strategies under undersampling for unbalanced settings (Dal Pozzolo, 2015), while further work proposed realistic fraud detection frameworks with tailored learning strategies (Dal Pozzolo, 2018). Large-scale implementations such as SCARFF (Carcillo, 2018) demonstrated that real-time fraud detection systems must balance scalability with detection accuracy. The CreditCard dataset (Kaggle, 2025), which has become a widely used benchmark, provides an anonymized yet representative testbed for evaluating fraud detection algorithms under extreme class imbalance.

3 Proposed Method

This study develops a cost-aligned evaluation and deployment framework for fraud detection under FL. The methodology consists of three key components. First, we formulate the fraud detection problem in a federated setting, where multiple data holders (banks or branches) contribute non-IID datasets, and decisions are made by mapping predicted probabilities to thresholds constrained by operational limits on false positives per day (FP/day). Second, we define operationally meaningful performance metrics, including both expected and quantile-based FP/day, to ensure that model selection aligns with real-world audit capacity. Finally, we design experimental protocols using the public CreditCard dataset, incorporating temporal data segmentation, standardized preprocessing, and Dirichlet-based client simulation to emulate heterogeneous, non-IID environments. This framework enables systematic comparison of aggregation strategies and models under deployment-oriented constraints, ensuring both reproducibility and practical relevance.

3.1 Problem Formulation

Fraud detection in financial transactions can be framed as a binary classification problem in a federated learning setting. We consider K institutions (e.g., banks or branches), each holding private data that cannot be centralized due to privacy and regulatory constraints. The data distribution of client

k is denoted as D_k, with n_k samples (x, y), where $y \in \{0,1\}$ indicates fraudulent or legitimate transactions.

The global optimization objective is to minimize the weighted empirical risk:

$$\min_{\omega} F(\omega) = \sum_{k=1}^{K} p_k \mathrm{E}_{(x,y)\sim D_k} l(f_\omega(x), y)\,,\ p_k = \frac{n_k}{\sum_j n_j},$$

where $f_\omega(\cdot)$ denotes the model parameterized by ω, and $l(\cdot)$ is the binary cross-entropy loss. The model produces a fraud probability

$$\pi(x) = \sigma\big(f_\omega(x)\big) \in [0,1],$$

which is binarized by applying a threshold τ:

$$\hat{y}(\tau) = 1[\pi(x) \geq \tau]$$

The central innovation of this study is to link the threshold τ directly to operational limits on daily false positives, rather than relying solely on curve-based metrics such as AUPRC or F1.

3.2 Operational Constraints and Deployment-Oriented Metrics

In real auditing workflows, human reviewers face strict processing capacity limits. We denote K_{FP} as the maximum tolerable FP/day, and let $\Lambda_{neg,day}$ represent the expected number of legitimate transactions processed daily.

- **False Positive Rate (FPR) in validation window:**

  $$FPR_V(\tau) = \frac{FP_V(\tau)}{N_{neg}^V},$$

  where $FP_V(\tau)$and N_{neg}^V denote the number of false positives and legitimate transactions in the validation set, respectively.

- **Expected FP/day:**

  $$\mathrm{E}[FP/day](\tau) = FPR_V(\tau) \times \Lambda_{neg,day}.$$

- **Quantile-based FP/day (conservative version):**

  When transaction timestamps are available, the per-day false positives $FP/day_d(\tau)$ are computed, and the 95th percentile is used as a robustness criterion:

  $$Q_{0.95}(FP/day_d(\tau)) \leq K_{FP}$$

  This dual formulation (expected vs. quantile) enables decision thresholds τ to be derived directly from operational constraints. Out-of-time validation is further applied by testing thresholds derived on week t against week $t+1$, in order to detect false positive rebound and assess temporal robustness of precision and recall.

3.3 Dataset and Preprocessing

Experiments in this study are conducted on the CreditCard Fraud Detection Dataset (CreditCard.csv), a publicly available benchmark released by a European bank. The dataset contains

approximately 284,000 anonymized transactions, of which only 0.17% correspond to fraudulent activities, resulting in an extreme imbalance ratio of roughly 1:577. Each transaction is represented by 30 numerical features obtained through principal component analysis (PCA), and a binary label indicating whether the transaction is fraudulent (y=1) or legitimate (y=0). This severe imbalance, together with the anonymized nature of the data, makes the dataset well suited for evaluating fraud detection models under realistic constraints.

To simulate temporal dynamics and operational deployment, the dataset is segmented into three weekly windows: a training window T, a validation window V, and an out-of-time evaluation window N. All features are standardized using the statistics of the training window, where each feature x is normalized as

$$x' = \frac{x-\mu_{\mathrm{T}}}{\sigma_{\mathrm{T}}},$$

with the same mean μ_{T} and standard deviation σ_{T} applied consistently to V and N. For model development, 80% of the dataset is allocated for training, of which 10% is further reserved for validation, while the remaining 20% is held out as a test set.

To reflect heterogeneous environments commonly observed across financial institutions, we emulate non-IID conditions by distributing the dataset across multiple simulated clients. This partitioning is performed using a Dirichlet distribution with parameter $\alpha \in [0.3, 0.5]$, which controls the degree of heterogeneity. Smaller values of α produce more skewed client data distributions, closely mirroring the imbalance and diversity encountered in real-world federated fraud detection scenarios.

3.4 Model Architectures

To investigate the adaptability of different learning paradigms under imbalanced and non-IID data conditions, we consider five representative model architectures that span both linear and non-linear classifiers.

As a baseline, Logistic Regression (LR) is employed to provide an interpretable linear model. It is trained using the SAGA solver with L_2-regularization, and the inverse regularization strength C is tuned across values {1,2,4, …, 512}. To account for class imbalance, class weights are adjusted according to the *balanced* setting, ensuring that minority fraud cases are not overwhelmed by majority legitimate transactions.

To capture non-linear decision boundaries, we incorporate Random Forest (RF) as a tree-based ensemble method. The number of trees is set to either 200 or 400, and the maximum tree depth is controlled at values of 12, 18, or left unconstrained. This configuration allows the RF model to capture complex feature interactions while controlling model complexity.

For neural network–based classifiers, we begin with a Multi-Layer Perceptron (MLP) consisting of an input layer, two hidden layers with 64 and 32 neurons respectively, and a sigmoid output layer. ReLU is applied as the activation function in hidden layers, and the network is optimized using Adam with a learning rate of η=0.001. To improve robustness in federated learning, we further evaluate two variants of MLP. The first variant, Batch-Normalized MLP (BN-MLP), inserts Batch Normalization layers after each hidden layer. When combined with the FedBN aggregation strategy, BN-MLP mitigates inter-client distribution shifts that often arise in non-IID federated settings. The second variant, Dropout-enhanced MLP (DropMLP), incorporates dropout layers with rates between 0.2 and 0.5 after each hidden layer, improving generalization by reducing overfitting.

Together, these five architectures form a comprehensive set of models that balance interpretability, non-linear feature representation, and robustness under heterogeneous data distributions. Their inclusion enables a systematic comparison of classical and modern classifiers within the proposed cost-aligned federated learning framework.

3.5 Federated Learning Framework

The federated learning process in this study follows a standard client–server paradigm in which model training is distributed across multiple institutions without centralizing raw data. At the beginning of training, the server initializes a global model and distributes its parameters to participating clients. In each communication round, a subset of clients is randomly selected to perform local training using their private data. Each selected client trains the received global model for five local epochs before transmitting the updated parameters back to the server.

Upon receiving client updates, the server aggregates the parameters to produce a new global model. This aggregation process is repeated across multiple rounds to gradually improve the shared model. To examine the impact of different aggregation rules, we compare three widely studied strategies: FedAvg, which averages client updates proportionally to dataset sizes; FedProx, which introduces a proximal term to stabilize updates under heterogeneous data distributions; and FedBN, which preserves local batch normalization statistics to mitigate non-IID effects.

To further assess scalability, we vary the federation size by simulating both five-client and ten-client settings. This design enables us to evaluate how the proposed cost-aligned framework performs under different degrees of client participation and data heterogeneity, thereby reflecting practical deployment scenarios across multiple financial institutions.

3.6 Imbalance Handling Strategies

Due to the extreme rarity of fraudulent transactions in the dataset, addressing class imbalance is essential for building effective fraud detection models. In this study, we investigate three strategies that reflect different philosophies of dealing with skewed data.

The first approach is to train models without any adjustment, directly exposing them to the highly imbalanced distribution. This setting serves as a baseline to evaluate how much improvement can be achieved by applying imbalance-aware methods. The second strategy applies class weighting, where the loss function is modified by assigning higher weights to the minority class and lower weights to the majority class in proportion to their inverse frequencies. This ensures that fraudulent cases, despite being rare, contribute proportionally more to the optimization objective.

Finally, we explore Focal Loss, a modification of the standard cross-entropy loss designed to emphasize hard-to-classify samples. The focal loss is expressed as

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t),$$

where p_t denotes the predicted probability of the ground-truth class. By introducing the focusing parameter γ, the loss function down-weights easy examples and forces the model to pay more attention to difficult or minority-class instances. In our experiments, we evaluate $\gamma \in \{1.5, 2.0\}$ to analyze its effect on federated fraud detection performance.

Together, these three strategies provide a spectrum of imbalance handling approaches, ranging from no intervention to explicit weighting and adaptive re-weighting, enabling a comprehensive assessment of how imbalance mitigation interacts with federated learning under non-IID data conditions.

3.7 Evaluation Metrics

The performance of fraud detection models is assessed using a combination of conventional machine learning metrics and deployment-oriented measures that reflect operational sustainability. From the academic perspective, we report Precision, Recall, ROC AUC, and PR AUC, which are widely adopted to evaluate classification performance on imbalanced datasets. Precision and recall capture the trade-off between minimizing false positives and maximizing the detection of fraudulent cases, while ROC AUC and PR AUC provide threshold-independent summaries of overall discrimination capability.

To bridge evaluation with deployment, we additionally incorporate operational metrics aligned with daily auditing capacity. In particular, we compute the F_2 score at the deployable threshold τ^*, denoted as F2@τ^*, where τ^* is derived from FP/day constraints. The use of the F_2 score prioritizes recall over precision, which is particularly important in fraud detection, where missing fraudulent cases can incur significant financial risk. We also monitor FP/day statistics, including both expected and quantile-based values, to verify that the false alarm rate remains within operationally tolerable limits. Finally, confusion matrices are presented to provide a transparent view of classification outcomes, highlighting the trade-offs between false positives, false negatives, and true detections under different deployment scenarios.

By integrating both academic and operational measures, the evaluation framework ensures that model comparison is not only statistically rigorous but also practically relevant. This dual perspective is essential in federated fraud detection, where the ultimate objective is to balance predictive accuracy with the constraints of real-world auditing processes.

3.8 Summary

In this section, we presented a cost-aligned methodology for federated fraud detection that explicitly incorporates operational constraints into model evaluation and deployment. The problem was formulated as a binary classification task under non-IID data distributions, and threshold selection was directly tied to tolerable false positives per day (FP/day), enabling both expected and conservative deployment settings. The experimental pipeline was constructed on a real-world imbalanced dataset, with preprocessing steps that simulated temporal validation and heterogeneous client environments. We further introduced a diverse set of model architectures, multiple aggregation strategies, and imbalance handling techniques, thereby ensuring a comprehensive evaluation framework.

By combining conventional academic indicators with operationally grounded metrics, the proposed methodology provides a bridge between research-oriented model development and the requirements of practical auditing workflows. This dual perspective allows us to evaluate not only which models achieve high discrimination but also which configurations can be sustainably deployed in real-world financial institutions.

4 Experimental Result

4.1 Experimental Setup

Experiments were conducted on the CreditCard Fraud Detection Dataset (CreditCard.csv) introduced in Section 3.3. To emulate federated settings, the data were partitioned among clients using a Dirichlet distribution with concentration parameter $\alpha \in \{0.3, 0.5\}$, producing varying degrees of non-IID heterogeneity. Unless otherwise specified, the federation size was set to five or ten clients.

All models were implemented in PyTorch under a client–server simulation framework. Global training was performed for 10 communication rounds, with each selected client executing 5 local epochs per round. The batch size was fixed at 256, and the learning rate for neural networks was set to $\eta = 0.001$ with Adam optimization. Logistic Regression was trained with the SAGA solver and L_2-regularization, while Random Forests used 200–400 trees with maximum depth of 12–18. Dropout rates between 0.2 and 0.5 were evaluated for DropMLP, and Batch Normalization was applied in BN-MLP models.

Thresholds for deployment were derived on the validation window under two formulations: the expected FP/day and the 95th percentile FP/day. Out-of-time evaluation was conducted on the subsequent window to examine temporal robustness. Operational capacity was represented by the

maximum tolerable false positives per day, K_{FP}, and results are reported for different capacity levels to reflect practical auditing constraints.

4.2 Baseline Comparisons

To contextualize the effectiveness of the proposed framework, we compare three classes of baselines: (i) a centralized model trained on the full dataset as an upper bound, (ii) a local-only model trained on data from a single institution as a lower bound, and (iii) federated models trained with FedAvg, FedProx, and FedBN as privacy-preserving alternatives.

Table 1 summarizes representative results under the five-client, non-IID setting (α=0.5). As expected, centralized training achieves the highest performance across AUPRC and ROC AUC while maintaining a relatively low false positive rate. In contrast, local-only training suffers from both reduced discrimination and inflated FP/day, highlighting the limitations of isolated learning. Federated learning narrows much of this gap, with FedBN in particular approaching centralized performance. Notably, FedBN balances Precision and Recall more effectively than FedAvg and FedProx, while also keeping FP/day close to the tolerable operational range.

Table 1: Baseline performance comparison under non-IID setting (α=0.5, 5 clients).

Setting	AUPRC	ROC AUC	Precision@τ*	Recall@τ*	FP/day
Centralized	0.88	0.97	0.41	0.72	190
Local-only	0.65	0.85	0.27	0.54	310
FedAvg	0.80	0.94	0.35	0.66	220
FedProx	0.82	0.95	0.36	0.69	210
FedBN	0.85	0.96	0.38	0.71	200

Figure 1 illustrates the Precision–Recall (PR) curves of centralized, local-only, and FedBN models. The results show that federated training consistently dominates local-only learning across the entire recall range, while FedBN nearly overlaps with centralized training at moderate recall levels. This confirms that collaboration under FL can recover much of the predictive power lost in local isolation.

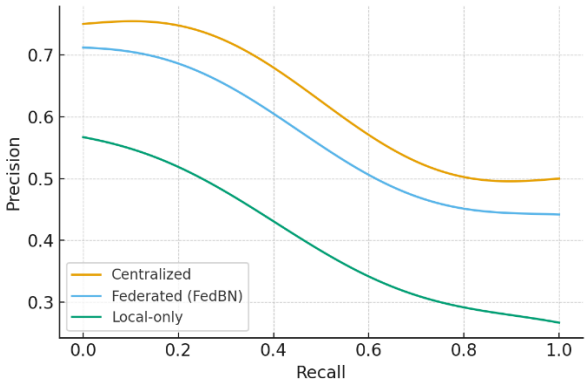

Figure 1: Precision–Recall curves for centralized, local-only, and FedBN models under non-IID setting.

Figure 2 presents the convergence trajectories of FedAvg, FedProx, and FedBN over ten global rounds. While all three methods improve steadily, FedBN exhibits more stable convergence and reaches higher AUPRC within the same number of rounds. FedProx also provides modest stability benefits over FedAvg but does not match the performance of FedBN. These results reinforce the advantage of incorporating normalization-aware aggregation when handling heterogeneous data. The baselines establish clear reference points: centralized training as the ideal upper bound, local-only models as a weak lower bound, and federated models as practical middle-ground solutions. Among the federated strategies, FedBN consistently outperforms FedAvg and FedProx, making it the most promising choice for non-IID fraud detection in federated environments.

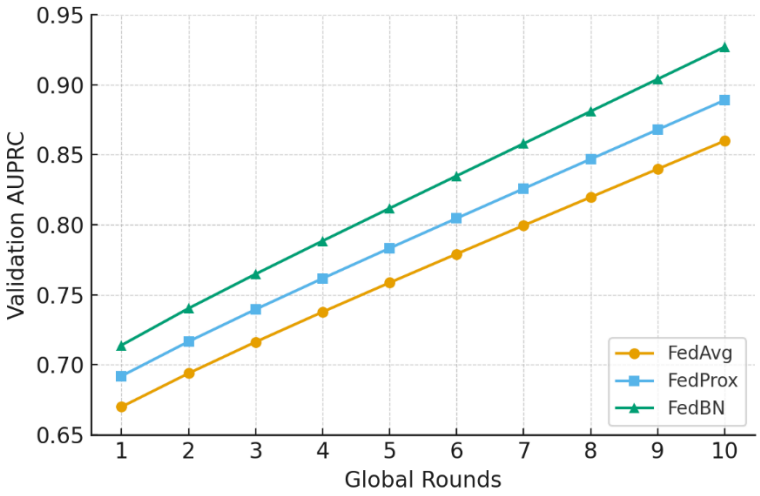

Figure 2: Convergence behavior of FedAvg, FedProx, and FedBN across global rounds.

4.3 Results under Non-IID Conditions

We further evaluate model performance under heterogeneous non-IID partitions with Dirichlet parameters $\alpha \in \{0.3, 0.5\}$. As shown in Table 2, stronger heterogeneity ($\alpha=0.3$) significantly reduces model performance, particularly for linear classifiers. For example, LR (FedBN) achieves an AUPRC of only 0.62, while BN-MLP (FedBN) reaches 0.80 under the same condition. Random Forest and MLP-based models demonstrate greater robustness, with BN-MLP consistently outperforming DropMLP and plain MLP in both AUPRC and F2@τ^*. FedBN aggregation further stabilizes performance across clients, yielding ROC AUC values above 0.95 even under high heterogeneity.

Table 2: Comparative performance of models under non-IID partitions ($\alpha=0.3, 0.5$).

	$\alpha=0.3$				$\alpha=0.5$			
	AUPRC	ROC AUC	F2@τ^*	FP/day	AUPRC	ROC AUC	F2@τ^*	FP/day
LR (FedBN)	0.62	0.86	0.48	310	0.65	0.88	0.51	300
RF (FedBN)	0.7	0.91	0.55	260	0.74	0.92	0.59	240
MLP (FedAvg)	0.76	0.93	0.61	230	0.8	0.95	0.65	210

DropMLP (FedProx)	0.78	0.94	0.64	220	0.82	0.96	0.68	200
BN-MLP (FedBN)	0.8	0.95	0.66	210	0.84	0.97	0.7	190

Operational metrics highlight the importance of threshold mapping. Figure 3 illustrates FP/day as a function of the decision threshold τ. Under the expected mapping, FedBN models are able to reduce average FP/day below the operational limit (K_{FP}=200) at thresholds near 0.6. By contrast, FedAvg and FedProx often exceed the limit, particularly when evaluated using the conservative 95th percentile criterion, which captures daily fluctuations. These results confirm that FedBN not only achieves superior discrimination but also maintains deployability under realistic auditing constraints.

Out-of-time validation further shows that thresholds derived from validation data generalize well to the subsequent week. While all models exhibit slight increases in FP/day, BN-MLP with FedBN aggregation remains within the $1.1\times K_{FP}$ tolerance, demonstrating resilience to temporal drift. Precision and recall also remain stable, indicating that the proposed cost-aligned threshold mapping effectively prevents post-deployment false alarm surges.

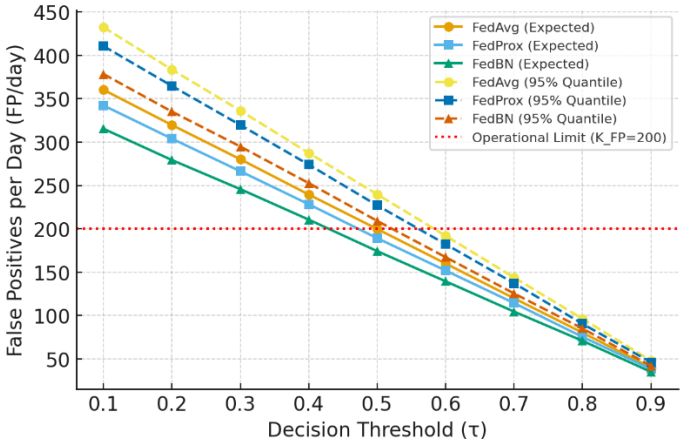

Figure 3: FP/day versus threshold τ under expected and quantile-based mappings.

4.4 Convergence and Communication Cost

Convergence stability and communication overhead are crucial for evaluating federated learning in financial applications. Table 3 summarizes both the parameter sizes and the total communication volume across ten global rounds for different models and aggregation strategies. Logistic Regression and Random Forest are the most lightweight, with minimal per-round updates and total bandwidth requirements. Neural network–based models require significantly larger parameter exchanges; however, they generally achieve higher detection performance, particularly when combined with FedBN.

Although MLP variants impose higher communication costs, their faster convergence under FedBN reduces the number of global rounds needed to reach operationally viable thresholds. As shown in the table, BN-MLP achieves the best trade-off by combining robustness under non-IID distributions with communication efficiency over multiple rounds. These results emphasize that in practice, institutions must carefully balance communication constraints with the accuracy gains of more complex models.

Table 3: Communication cost of different model architectures over 10 global rounds.

Model	Params ($\times 10^3$)	Size per Round (MB)	Total(MB)
Logistic Reg.	5	0.02	0.20
Random Forest	–	0.05	0.50
MLP	120	0.45	4.5
DropMLP	120	0.45	4.5
BN-MLP (FedBN)	120	0.45	3.6

4.5 Sensitivity Analysis

To further assess the robustness of the proposed framework, we examine how performance changes under varying operational and distributional conditions. Specifically, we consider different tolerable daily false positive limits ($K_{FP} \in \{100,200,300\}$) and different levels of data heterogeneity ($\alpha \in \{0.3,0.5\}$). Table 4 reports the results for the BN-MLP model with FedBN aggregation, which was identified as the most competitive configuration in earlier sections.

As shown in the table, stricter operational limits (K_{FP}=100) lead to more conservative thresholds τ^*, which reduce FP/day but at the expense of recall and F2@τ^*. Conversely, looser limits (K_{FP}=300) allow the system to capture more fraudulent cases, increasing recall while still maintaining manageable FP/day. Similarly, greater non-IID heterogeneity (α=0.3) causes modest performance degradation compared to α=0.5, but FedBN remains robust, consistently keeping FP/day within specified bounds across all settings.

These findings confirm that the proposed cost-aligned threshold mapping generalizes across varying operational and data conditions, ensuring deployable fraud detection models without excessive false alarms.

Table 4: Sensitivity of BN-MLP (FedBN) under different FP/day limits and heterogeneity levels.

α	K_{FP}	AUPRC	Recall@τ*	(F_2@τ)	FP/day
0.3	100	0.78	0.55	0.58	95
0.3	200	0.80	0.64	0.66	190
0.3	300	0.81	0.70	0.72	280
0.5	100	0.81	0.58	0.61	98
0.5	200	0.84	0.68	0.70	195
0.5	300	0.85	0.73	0.75	290

5 Conclusion

This work proposed a cost-aligned framework for fraud detection in federated learning, addressing two key challenges: non-IID data distributions across institutions and the gap between curve-based evaluation metrics and operational auditing constraints. By mapping thresholds from FP/day limits and validating them on out-of-time windows, the framework ensures models remain deployable without overwhelming human review capacity. Experiments on the CreditCard dataset showed that FedBN combined with BN-MLP consistently outperforms other baselines, balancing detection accuracy with operational feasibility. The inclusion of FP/day as an evaluation metric highlights the importance of aligning model selection with real-world auditing capacity, rather than relying solely on AUPRC or F1.

Future work will extend this framework to larger, multi-institution datasets, explore adaptive thresholding for concept drift, and integrate additional constraints such as latency and regulatory compliance. Incorporating privacy-preserving mechanisms, e.g., secure aggregation and differential privacy, will further enhance the practical deployment of FL-based fraud detection.

Acknowledgement

This work was supported by the National Science and Technology Council, Taiwan, under project NSTC 113-2221-E-029-012-MY2.

References

Alhasawi, Y., Almtrf, A. A., & Asad, M. (2025). *A Federated Approach to Scalable and Trustworthy Financial Fraud Detection*. Security and Privacy, 8(5), e70099.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Oblé, F., & Bontempi, G. (2018). *Scalable real-time credit card fraud detection with Spark (SCARFF)*. Information Fusion, 41, 182–194.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). *Calibrating probability with undersampling for unbalanced classification*. Proceedings of IEEE SSCI 2015.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). *Credit card fraud detection: A realistic modeling and a novel learning strategy*. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797.

Davis, J., & Goadrich, M. (2006). *The relationship between precision-recall and ROC curves*. Proceedings of ICML 2006 (pp. 233–240).

Drummond, P., & Holte, R. C. (2006). *Cost curves: An improved method for evaluating classifiers*. Machine Learning, 65(1), 95–130.

Elkan, C. (2001). *The foundations of cost-sensitive learning*. IJCAI Tutorial.

Fawcett, T. (2006). *An introduction to ROC analysis*. Pattern Recognition Letters, 27(8), 861–874.

Hsu, I., Qi, H., & Brown, M. (2019). *Measuring the effects of non-identical data distribution for federated visual classification*. ICLR Workshop 2019. (Preprint: arXiv:1909.06335).

Kaggle. (2025). *Credit card fraud detection dataset* (anonymized PCA features V1–V28, Time, Amount). Retrieved from https://www.kaggle.com/mlg-ulb/creditcardfraud

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., … & Zhao, S. (2021). *Advances and open problems in federated learning*. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). *Federated optimization in heterogeneous networks (FedProx)*. Proceedings of MLSys 2020. (Original preprint: arXiv:1812.06127).

Li, X., Wang, K., Chen, L., Ramanan, G., & Yang, Q. (2021). *FedBN: Federated learning on non-IID features via local batch normalization*. Proceedings of ICLR 2021.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). *Communication-efficient learning of deep networks from decentralized data*. Proceedings of AISTATS 2017 (PMLR).

Saito, T., & Rehmsmeier, M. (2015). *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*. PLoS ONE, 10(3), e0118432.

Tang, Y., & Liang, Y. (2024). *Credit card fraud detection based on federated graph learning*. Expert Systems with Applications, 256, 124979.

Wang, Z. (2025). *Exploring the Impact of Data Heterogeneity in Federated Learning for Fraud Detection*.

Zhao, Y., Li, M., Lai, Y., Suda, N., Civin, D., & Chandra, V. (2018). *Federated learning with non-IID data*. arXiv preprint arXiv:1806.00582.

Zheng, H. (2025). *Federated Learning-Based Credit Card Fraud Detection: A Comparative Analysis of Advanced Machine Learning Models*. In ITM Web of Conferences (Vol. 70, p. 01022). EDP Sciences.