Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 6 (2025) 1366 - 1374 e-ISSN: 2580-0760

Robust Aggregation Strategies in Federated Learning for Credit Risk Assessment

Sulthonika Mahfudz Al Mujahidin[1], Michael Angello Qadosy Riyadi[2], Adinda Mariasti Dewi[3], Mustafa Kamal[4*]
[1,2,3,4]Department of Information Technology (Surabaya City Campus), Telkom University, Surabaya, Indonesia

[1]sulthonikamahfudzam@student.telkomuniversity.ac.id, [2]michaelangello@student.telkomuniversity.ac.id,
[3]adindamariasti@student.telkomuniversity.ac.id, [4*]mustafakamal@telkomuniversity.ac.id

Abstract

Financial institutions face challenges in credit risk assessment due to fragmented data and strict privacy regulations, which hinder predictive modeling and increase financial risks. Federated Learning (FL) enables privacy-preserving collaborative modeling without sharing raw data. This study evaluates five FL aggregation methods—Federated Averaging (FedAvg), Weighted Average, Median Aggregation, Federated Proximal (FedProx), and Stochastic Controlled Averaging (SCAFFOLD)—using logistic regression on the Credit Approval dataset (690 records, five clients) with non-IID label and feature distributions. Local models were trained and aggregated over 50 rounds. Median Aggregation outperformed the other methods, achieving an F1-score of 97.85% and a recall of 80.6% (vs. 72.3% for others), demonstrating robustness against data skewness. However, global model performance (85.22% for FedAvg, Weighted Average, FedProx, SCAFFOLD; 85.80% for Median) remained static across rounds, indicating limited convergence due to rapid local model convergence and non-IID challenges. The high communication cost of 50 rounds highlights a trade-off between accuracy and efficiency, necessitating optimized strategies like adaptive regularization or client sampling. This study advances theoretical understanding of FL under heterogeneity and provides practical guidance for secure, regulation-compliant credit risk modeling in financial institutions. Future work should explore larger datasets, multi-round convergence, and privacy mechanisms like differential privacy to mitigate risks such as model inversion attacks while ensuring compliance

Keywords: credit risk; federated learning; financial technology; non-IID data; privacy-preserving machine learning

How to Cite: S. M. A. Mujahidin, M. A.Q. Riyadi, A. M. Dewi, and M. Kamal, "Robust Aggregation Strategies in Federated Learning for Credit Risk Assessment", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 6, pp. 1366 - 1374, Dec. 2025.
Permalink/DOI: https://doi.org/10.29207/resti.v9i6.7133

Received: August 30, 2025
Accepted: November 16, 2025
Available Online: December 7, 2025

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

In the financial industry, credit risk assessment is critical for portfolio stability, yet fragmented data across banks, financing companies, and credit bureaus complicates comprehensive modeling, increasing the risk of inaccurate decisions and non-performing loans [1] - [5]. Strict privacy regulations, such as Indonesia's Law No. 27/2022 on Personal Data Protection and OJK rules, prohibit raw data sharing without explicit consent and robust safeguards, posing technical and legal barriers to cross-institutional collaboration [6] - [7]. These restrictions necessitate costly security infrastructure and multi-layered audits, limiting direct data integration and prompting exploration of privacy-preserving solutions like Federated Learning (FL) [13] - [16].

Federated Learning enables financial institutions to collaboratively train credit risk models without sharing sensitive data, using a horizontal scheme where local models are trained and only encrypted parameters are aggregated [17], [18]. Studies show FL improves generalization over siloed models, which often suffer from overfitting and poor performance in diverse customer segmentation [19]. However, non-IID data—characterized by label skew (e.g., class imbalances) and feature skew (e.g., varying risk profiles)—challenges FL's effectiveness, as global model convergence is disrupted by parameter bias [20] - [27]. This is particularly critical in Indonesia's financial sector, where institutions serve diverse markets with distinct risk characteristics [28].

Standard Federated Averaging (FedAvg) performs well with homogeneous data but degrades under non-IID

conditions due to weight divergence and gradient deviation, as evidenced by reduced accuracy in heterogeneous settings [29], [30]. Robust aggregation methods, such as Median Aggregation to mitigate outliers, Federated Proximal (FedProx) to reduce client drift via proximal regularization, and Stochastic Controlled Averaging (SCAFFOLD) to correct client-specific variations, aim to address these challenges [24], [31] - [35]. These methods show promise in improving F1-scores and recall under label skew, though their application in real financial datasets remains underexplored [31], [33], [36], [37].

This study evaluates five FL aggregation strategies—FedAvg, Weighted Average, Median Aggregation, FedProx, and SCAFFOLD—using the Credit Approval dataset in a five-client setup with balanced sample sizes but significant label and feature skew. Employing logistic regression, the study conducts a single model update over 50 rounds to assess initial efficiency and robustness. Weighted Average serves as a baseline alongside FedAvg, while FedProx and SCAFFOLD introduce advanced mechanisms to tackle heterogeneity. The evaluation addresses the critical gap in applying robust aggregation to real financial data, offering insights into secure, regulation-compliant credit risk modeling.

The study's contributions are twofold: theoretically, it advances understanding of FL performance under non-IID conditions in financial contexts; practically, it guides institutions in deploying collaborative models while mitigating risks like data inference attacks [17, 28]. By comparing aggregation strategies, this research highlights their impact on model convergence and performance, addressing the unique challenges of Indonesia's heterogeneous financial ecosystem and paving the way for scalable, privacy-preserving solutions.

2. Methods

This research simulates a federated learning (FL) environment among financial institutions, where each client independently trains a local model on non-IID data subsets while preserving privacy by avoiding direct raw data exchange. The workflow begins with data collection, followed by preprocessing steps such as handling missing values, one-hot encoding categorical features, and standardizing numerical features. The processed dataset is then divided among multiple clients, where each client trains a local model using logistic regression. Afterward, the locally trained models are aggregated at the global level using various strategies, including FedAvg, Weighted Average, median, FedProx, and SCAFFOLD, as illustrated in Figure 1. This approach not only ensures confidentiality of sensitive financial data but also facilitates collaborative learning that enhances the robustness and generalization of the global model.

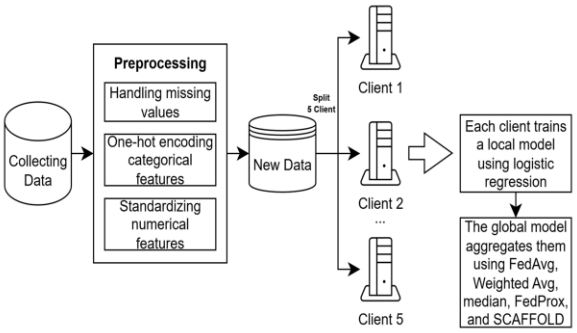


Figure 1. Research Flow Diagram

2.1 Dataset

The Credit Approval dataset from the UCI Machine Learning Repository comprises 690 entries, divided into five non-IID subsets, each representing a financial institution. It includes 15 predictor attributes (6 numerical, 9 categorical) and a binary target variable (credit approval/rejection), with a class imbalance of 44.5% positive and 55.5% negative, and 4.85% missing values. Feature details are provided in Table 1.

Table 1. Characteristics of the Credit Approval Dataset

Variable Name	Role	Type	Missing Values
A16	Target	Categorical	No
A15	Feature	Continuous	No
A14	Feature	Continuous	Yes (13)
A13	Feature	Categorical	No
A12	Feature	Categorical	No
A11	Feature	Continuous	No
A10	Feature	Categorical	No
A9	Feature	Categorical	No
A8	Feature	Continuous	No
A7	Feature	Categorical	Yes (9)
A6	Feature	Categorical	Yes (9)
A5	Feature	Categorical	Yes (6)
A4	Feature	Categorical	Yes (6)
A3	Feature	Continuous	No
A2	Feature	Continuous	Yes (12)
A1	Feature	Categorical	Yes (12)

2.2 Preprocessing

To prepare the dataset for logistic regression in the FL framework, a preprocessing pipeline addressed the 4.85% missing values and non-IID characteristics [25]. Numerical features (A2, A3, A8, A11, A14, A15) were imputed using the mean, and categorical features (A1, A4, A5, A6, A7, A9, A10, A12, A13) using the mode, following financial data best practices [38]. Categorical features were one-hot encoded to ensure model compatibility without ordinal assumptions [39]. Numerical features were standardized to zero mean and unit variance to eliminate scale-based bias and accelerate convergence [40]. This ensured consistency across clients and enhanced model robustness.

2.3. Experimental Setup

The experiment was conducted with five clients, each representing a financial institution, where the dataset was evenly split into 138 samples per client but exhibited varying class distributions. For instance, Client 2 contained 86.23% positive classes, while Client 3 consisted of 94.93% negative classes, leading to a pronounced label imbalance, as illustrated in Figure 2. This configuration mirrors practical cross-institutional challenges, emphasizing the heterogeneity that is commonly encountered in real-world scenarios. Moreover, such diversity among clients provides a realistic testbed to evaluate the effectiveness and robustness of federated learning aggregation strategies under non-IID conditions, ensuring that the developed global model can adapt to diverse and imbalanced data sources.

Each client trained a logistic regression model locally, followed by 50 rounds of central aggregation using five methods: Federated Averaging (FedAvg), Weighted Average, Median Aggregation, Federated Proximal (FedProx), and Stochastic Controlled Averaging (SCAFFOLD). Statistical analysis revealed significant feature skew, particularly in A11 (CV = 0.9586) and A15 (CV = 1.0353), impacting global learning, as detailed in Table 2.

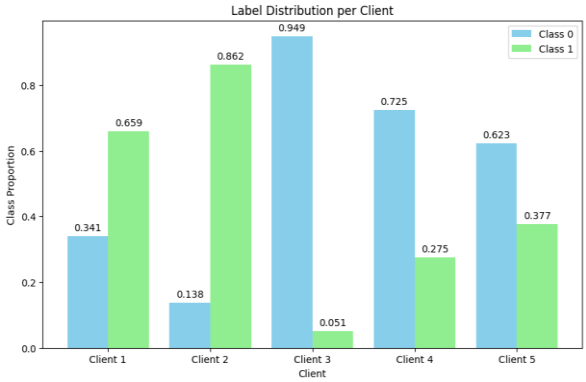


Figure 2. Binary Label Distribution per Client

Table 2. Variance, Numerical Feature Coefficient of Variation, and Number of Samples per Client

Feature	Variance Client 1	Variance Client 2	Variance Client 3	Variance Client 4	Variance Client 5	Coefficient of Variation	Number of Samples
A2	0.8909	1.1978	0.8634	1.1367	0.8265	0.1556	138
A3	1.1628	1.1007	0.8416	0.8889	0.8448	0.1409	138
A8	1.4832	1.6902	0.3208	0.5363	0.4973	0.6236	138
A11	2.6022	0.9541	0.0616	0.4906	0.5101	0.9586	138
A14	1.0273	0.6731	1.7669	0.7146	0.7408	0.4171	138
A15	1.7022	0.4666	2.6883	0.0409	0.1007	1.0353	138

2.4 Model

Logistic regression was selected for binary classification due to its computational efficiency and interpretability in financial contexts [19]. No hyperparameter optimization was performed to ensure experimental consistency, focusing evaluation on aggregation strategies rather than model tuning [40].

This positions logistic regression as a classification and sensitivity probe for assessing aggregation methods under data heterogeneity.

2.5 Federated Learning Aggregation Strategy

To form a global model in federated learning, this study implements five parameter aggregation methods to evaluate sensitivity to non-IID data distribution: Federated Averaging (FedAvg), Weighted Average, Median Aggregation, Federated Proximal (FedAvg), and Stochastic Controlled Averaging (SCAFFOLD). All five are presented in mathematical equations to clarify the mechanism of combining local models on each client.

FedAvg is a basic method that calculates the arithmetic mean of local model parameters trained by each client. Each client has an equal contribution, regardless of the size of the data it has, as formulated in Equation 1.

$$w_g = \frac{1}{N}\sum_{i=1}^{N} w_i, \quad b_g = \frac{1}{N}\sum_{i=1}^{N} b_i \tag{1}$$

w_g and b_g denote the global model's weight vector and intercept, while w_i and b_i represent those from client i, with N = 5. This equal weighting approach is efficient when data distribution is uniform across clients, but it faces substantial challenges in heterogeneous settings. In cases where the dataset exhibits extreme class imbalance, such as 86.23 percent positive class in Client 2 and 94.93 percent negative class in Client 3 (see Figure 2), outlier parameters can dominate the aggregation process and compromise the representativeness of the global model.

In addition to the equal contribution approach, aggregation can also be done by considering the proportion of data owned by each client. This method is known as Weighted Average, and is formulated as Equation 2.

$$w_g = \sum_{i=1}^{N}\left(\frac{n_i}{n}\right) w_i, \quad b_g = \sum_{i=1}^{N}\left(\frac{n_i}{n}\right) b_i \tag{2}$$

w_g and b_g represent the global model's weights and intercept, while w_i dan b_i and n_i refer to the local parameters and sample count of the i-th client, with n denoting the total number of samples across all clients. In the current configuration, each client receives 138 samples out of 690, making the relative weights identical and thus rendering the aggregation outcome practically equivalent to FedAvg.

However, this approach remains relevant as it reflects real-world conditions in the financial sector, where data

volumes across institutions are often imbalanced [23]. The evaluation aims to assess how weighting schemes influence model aggregation under non-IID conditions, especially in the presence of feature and label skew. A summary of numerical feature variability across clients is provided in Table 2.

While both FedAvg and weighted averaging perform well under uniform conditions, they remain vulnerable to extreme client deviations that often arise in federated financial systems. These limitations motivate the adoption of Median Aggregation, which consolidates model parameters by computing the element-wise median across clients using Equation 3.

$$w_{g=}\text{median}(w_1, w_2, \dots, w_N),\ b_g = \text{median}(b_1, b_2, \dots, b_N) \quad (3)$$

Here, w_g and b_g denote the global model's parameters, while w_i and b_i represent those from client i. The median is computed element-wise over all five clients.

This method effectively filters out outliers, making it suitable for non-IID settings like finance, where parameter skew is common. As shown in Table 2, features such as A11 and A15 display high variation, and extreme label imbalance (see Figure 2) further increases aggregation bias risk. Median Aggregation is thus not merely an alternative but a mitigation strategy to ensure stability and fairness in multi-institution collaboration.

To further address the challenges of non-IID data, Federated Proximal (FedProx) introduces a proximal term to the local training objective, encouraging local models to stay close to the global model while optimizing on client data [34]. This method modifies the local update by adding a regularization term, formulated as Equation 4

$$\min_{w_i}\left[F_i(w_i) + \frac{\mu}{2}|w_i - w_g|^2\right],\quad w_g = \sum_{i=1}^{N}\left(\frac{n_i}{n}\right)w_i,\quad b_g = \sum_{i=1}^{N}\left(\frac{n_i}{n}\right)b_i \quad (4)$$

Here, $F_i(w_i)$ is the local loss function for client i, μ is the proximal coefficient, w_i and w_g are the local and global weight vectors, and n_i and n denote the client and total sample counts, respectively. The global parameters w_g and b_g are computed similarly to Weighted Average. FedProx mitigates client drift in heterogeneous settings, such as the extreme label imbalances shown in Figure 2, by balancing local optimization with global consistency, making it suitable for financial datasets with diverse risk profiles.

Stochastic Controlled Averaging (SCAFFOLD) enhances robustness by using control variates to correct for client-specific variations during local training [35]. It adjusts local updates to account for the difference between local and global gradients, with the global model aggregated as Equation 5.

$$w_g = \sum_{i=1}^{N}\left(\frac{n_i}{n}\right)\left(w_i + c_i - c_g\right),\quad b_g = \sum_{i=1}^{N}\left(\frac{n_i}{n}\right)\left(b_i + \bar{c_i} - \bar{c_g}\right) \quad (5)$$

c_i and c_g are the client and server control variates, respectively, and $\bar{c_i}$ and $\bar{c_g}$ are their respective means for the intercept. The control variates c_i are updated locally to capture client-specific drift, as observed in the feature variability in Table 2. SCAFFOLD's mechanism reduces the impact of non-IID distributions, such as the class imbalances in Figure 2, ensuring better convergence in cross-institutional financial collaborations.

2.6 Federated Learning Algorithm

The FL process involves 50 rounds of aggregation, as shown in Figure 3.

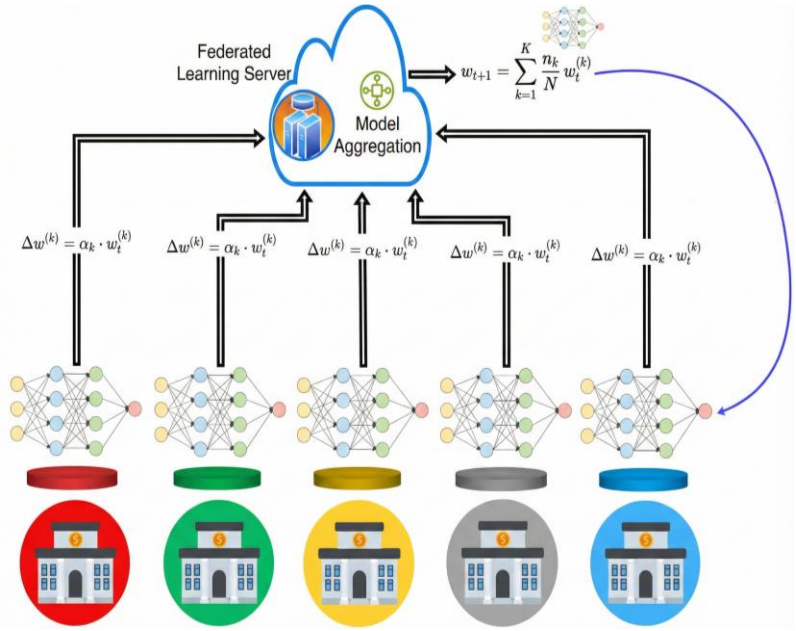


Figure 3. Federated Learning Architecture in the Context of Multi-Financial Institutions.

The server initializes a global model, distributes it to clients, and aggregates local parameters without accessing raw data. Each client trains locally, evaluates performance, and sends parameters and metrics to the server. The algorithm, detailed in Algorithm 1, ensures privacy through parameter-only exchanges.

Algorithm 1. Federated Learning with Communication Rounds
Input: A set of clients N, each with private local datasets that are not shared with the central server, number of communication rounds R = 50.

Procedure: Initialize a global model structure on the central server.

For each round r = 1 to R:

Distribute the current global model to all clients.
Each client independently performs the following steps:
Initialize the local model with the received global model parameters.
Train the local model using its private dataset.
Evaluate the trained local model on its own data and store the resulting performance metrics.
Transmit the trained local model parameters (e.g., weights, intercepts) to the central server.

Send the corresponding evaluation metrics to the server.

The central server then executes the following actions:

Aggregate all received local parameters into a single global model using a predefined method (e.g., FedAvg, Weighted Average, Median, FedProx, or SCAFFOLD).
Maintain strict privacy by never accessing or storing raw client data.
The newly aggregated global model is prepared for the next round or final evaluation.
After R rounds, the central server broadcasts the final global model to all clients.
Each client evaluates this global model on their respective local data and sends the performance metrics to the central server.

Finally, the server collects all local and global evaluation metrics from the final round for comparative analysis, ensuring that no raw data is exchanged throughout the entire process.

Output: An aggregated global model trained from local updates over R rounds, and performance metrics from each client based on both local and global models.

This implementation enables systematic evaluation of the five aggregation strategies within the heterogeneous Credit Approval dataset context, providing empirical foundation for comparative analysis of FL effectiveness under non-IID conditions.

2.7 Evaluation and Statistical Test

Performance is assessed using accuracy, precision, recall, and F1-score. In credit approval, high recall minimizes undetected risky loans, while high precision reduces false rejections [41], [42]. The F1-score balances these, critical for non-IID data [43]. Accuracy serves as a baseline but is secondary due to label imbalances [44]. Global models from FedAvg, Weighted Average, Median Aggregation, FedProx, and SCAFFOLD are evaluated, with F1-score prioritized. Statistical significance is tested using the Shapiro-Wilk normality test, followed by ANOVA (normal data) or Kruskal-Wallis (non-normal data), ensuring robust comparison of aggregation strategies.

3. Results and Discussions

This section evaluates the federated learning (FL) implementation for financial systems, focusing on local and global model performance, the impact of data heterogeneity on aggregation methods, challenges in cross-iteration updates, and trade-offs between accuracy and communication efficiency. Statistical analysis and practical implications in financial contexts are discussed, alongside privacy and security considerations.

3.1 Local Model Performance Evaluation

Local model evaluation demonstrates the FL system's ability to train distributed models effectively before aggregation. Table 3 summarizes performance across five clients, with accuracies ranging from 84.06% to 97.83%. Client 2 achieves the highest F1-score (98.72%) due to its balanced data distribution (86.23% positive, Figure 2), highlighting the efficacy of local training in favorable conditions. However, Client 3's performance degrades significantly (recall 42.86%, F1-score 60.00%) due to extreme label skew (94.93% negative), underscoring the impact of non-IID data on local generalization. This variability emphasizes the need for robust aggregation strategies to address heterogeneity in financial datasets, where feature (e.g., A11, A15, Table 2) and label skews are common.

Table 3. Local Model Performance per Client

Client	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Client 1	84.06	85.57	91.21	88.30
Client 2	97.83	100.00	97.48	98.72
Client 3	97.10	100.00	42.86	60.00
Client 4	96.38	92.31	94.74	93.51
Client 5	95.65	97.92	90.38	94.0

3.2 Evaluation of Global Models by Aggregation Method

Five aggregation methods—Federated Averaging (FedAvg), Weighted Average, Median Aggregation, Federated Proximal (FedProx), and Stochastic Controlled Averaging (SCAFFOLD)—were evaluated for their sensitivity to non-IID financial data, as shown in Table 4. FedAvg and Weighted Average yield identical performance (average accuracy 85.22%, F1-score 72.85%) due to uniform dataset sizes (138 samples per client), rendering weighting ineffective. Both struggle with Client 3's extreme label skew (94.93% negative), resulting in a low F1-score

(30.77%). Outlier parameters from imbalanced clients dominate the arithmetic mean, compromising global model representativeness.

Median Aggregation mitigates outliers by selecting element-wise medians, achieving a higher F1-score (97.85%) for Client 2's balanced data but lower accuracy (63.04%) for Client 1 due to feature skew (Table 2). This robustness comes at the cost of reduced sensitivity to extreme heterogeneity. FedProx, designed to reduce client drift via a proximal term, performs identically to FedAvg (accuracy 85.22%, F1-score 72.85%), suggesting limited impact from the regularization parameter (\mu) in this logistic regression setup. Similarly, SCAFFOLD's control variates fail to address Client 3's skew, mirroring FedAvg's performance. These results highlight that non-IID data, particularly label and feature skews, poses a significant challenge across all methods, with Median Aggregation offering partial robustness.

Table 4. Global Model Performance per Client and Aggregation Method

Method	Client	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
FedAvg	Client 1	73.19	75.47	87.91	81.22
FedAvg	Client 2	83.33	100.00	80.67	89.30
FedAvg	Client 3	93.48	33.33	28.57	30.77
FedAvg	Client 4	86.23	69.39	89.47	78.16
FedAvg	Client 5	89.86	97.50	75.00	84.78
Weighted Average	Client 1	73.19	75.47	87.91	81.22
Weighted Average	Client 2	83.33	100.00	80.67	89.30
Weighted Average	Client 3	93.48	33.33	28.57	30.77
Weighted Average	Client 4	86.23	69.39	89.47	78.16
Weighted Average	Client 5	89.86	97.50	75.00	84.78
Median Aggregation	Client 1	63.04	65.15	94.51	77.13
Median Aggregation	Client 2	96.38	100.00	95.80	97.85
Median Aggregation	Client 3	93.48	33.33	28.57	30.77
Median Aggregation	Client 4	82.61	61.67	97.37	75.51
Median Aggregation	Client 5	93.48	95.74	86.54	90.91
FedProx	Client 1	73.19	75.47	87.91	81.22
FedProx	Client 2	83.33	100.00	80.67	89.30
FedProx	Client 3	93.48	33.33	28.57	30.77
FedProx	Client 4	86.23	69.39	89.47	78.16
FedProx	Client 5	89.86	97.50	75.00	84.78
SCAFFOLD	Client 1	73.19	75.47	87.91	81.22
SCAFFOLD	Client 2	83.33	100.00	80.67	89.30
SCAFFOLD	Client 3	93.48	33.33	28.57	30.77
SCAFFOLD	Client 4	86.23	69.39	89.47	78.16
SCAFFOLD	Client 5	89.86	97.50	75.00	84.78

3.3 Challenges in Cross-Iteration Updates

Figure 4 illustrates the average accuracy of global models over 50 communication rounds, with FedAvg, Weighted Average, FedProx, and SCAFFOLD at 85.22% and Median Aggregation at 85.80%. The lack of performance improvement across rounds indicates rapid local model convergence to client-specific optima, limiting global model refinement. This stagnation, driven by non-IID data (e.g., Client 3's 94.93% negative class), suggests that cross-iteration updates fail to adapt to heterogeneous distributions. Local models dominate with client-specific patterns, hindering global convergence and highlighting the need for adaptive strategies, such as dynamic (\mu) tuning in FedProx or optimized control variates in SCAFFOLD.

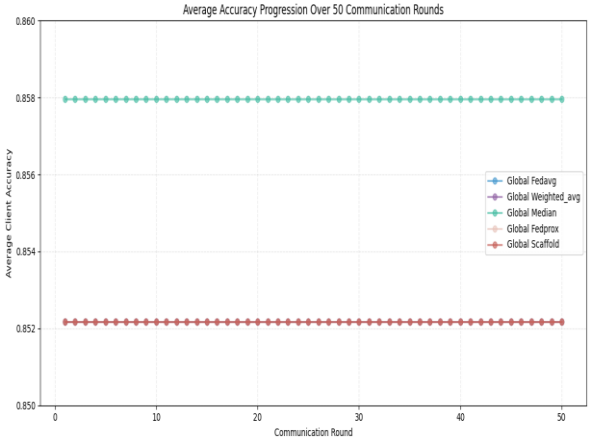


Figure 4. Comparison of aggregation methods across 50 communication rounds

Figure 4 visualizes the average accuracy of global models over 50 rounds, showing constant values: FedAvg, Weighted Average, FedProx, and SCAFFOLD at 85.22%, and Median Aggregation at 85.80%. The lack of variation reflects the challenge of cross-iteration updates in non-IID settings, where local model dominance and rapid convergence prevent global model refinement.

3.4 Trade-Off Between Accuracy and Communication Efficiency

The high communication cost of 50 rounds without performance gains underscores a critical trade-off. While FL preserves privacy by avoiding data sharing, the lack of global model improvement in heterogeneous settings increases computational inefficiency. For instance, Client 3's low F1-score (30.77%) across all methods reflects the challenge of integrating diverse financial patterns.

Strategies like selective client participation or fewer rounds could reduce communication overhead while maintaining accuracy, particularly in large-scale financial consortia where thousands of institutions collaborate.

3.5 Local vs. Global Performance Comparison

Local models outperform global models across all metrics (average accuracy 94.20%, precision 95.16%) compared to global averages of 85.22% (FedAvg, Weighted Average, FedProx, SCAFFOLD) and 85.80% (Median Aggregation). This gap, visualized in Figure 5, validates FL's strength in local specialization, particularly for Client 2's balanced data (F1-score 98.72%). Median Aggregation's higher recall (80.56%) versus FedAvg's 72.32% reflects its robustness to outliers, but its lower precision (71.22%) indicates sensitivity to extreme heterogeneity. The consistent global performance across rounds further suggests that local convergence limits global adaptability, thereby necessitating more advanced aggregation techniques for improved outcomes.

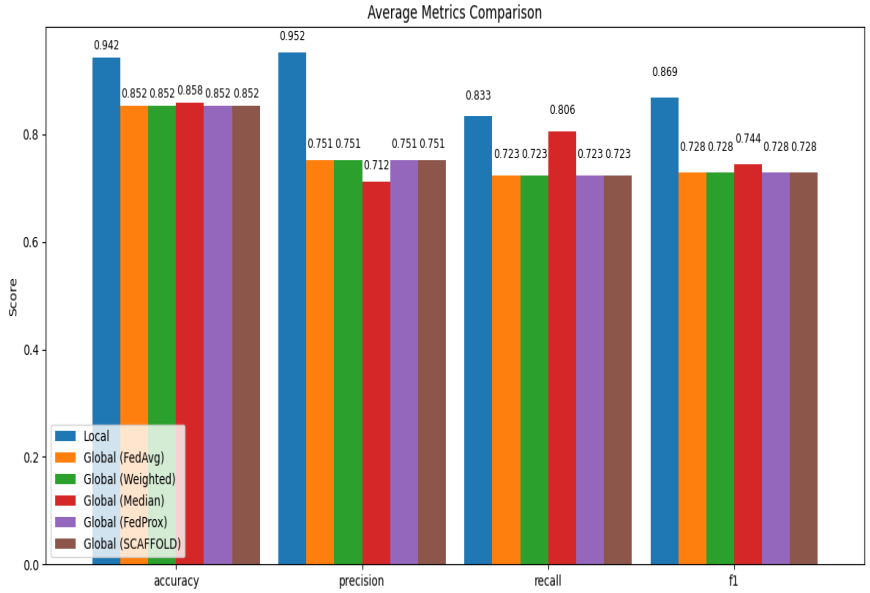


Figure 5. Average Performance of Local and Global Models

3.6 Statistical Analysis

The Shapiro-Wilk test (Table 5) confirms heterogeneity in FL performance. Local models exhibit non-normal distributions (accuracy p-value 0.0071, F1-score 0.0418), reflecting variable performance due to client-specific data characteristics.

Table 5. Shapiro-Wilk Normality Test Results

Group	Metric	W	p-value	Normality
Local	Accuracy	0.6876	0.0071	Not normal
Local	F1-Score	0.7662	0.0418	Not normal
Global FedAvg	Accuracy	0.9488	0.7287	Normal
Global FedAvg	F1-Score	0.7188	0.0149	Not normal
Global Weighted Average	Accuracy	0.9488	0.7287	Normal
Global Weighted Average	F1-Score	0.7188	0.0149	Not normal
Global Median Aggregation	Accuracy	0.8102	0.0978	Normal
Global Median Aggregation	F1-Score	0.8531	0.2046	Normal
Global FedProx	Accuracy	0.9488	0.7287	Normal
Global FedProx	F1-Score	0.7188	0.0149	Not normal
Global SCAFFOLD	Accuracy	0.9488	0.7287	Normal
Global SCAFFOLD	F1-Score	0.7188	0.0149	Not normal

Global models show mixed results: FedAvg, Weighted Average, FedProx, and SCAFFOLD have normal accuracy (p-value 0.7287) but non-normal F1-scores (p-value 0.0149), indicating stable accuracy but variable precision-recall balance due to non-IID data. Median Aggregation achieves normality in both metrics (p-value 0.0978, 0.2046), confirming its robustness but lower precision. These findings underscore the challenge of achieving consistent global performance in heterogeneous financial systems.

3.7 Practical Contributions

The FL implementation offers practical value for financial institutions, such as banks collaborating on credit risk models without sharing sensitive customer data. For example, a consortium of regional banks could use Median Aggregation to mitigate the impact of imbalanced loan default datasets, ensuring robust global models for fraud detection.

Simulations with larger datasets (e.g., millions of transactions across global banks) could further validate scalability, addressing real-world challenges like regulatory compliance and cross-border data heterogeneity.

3.8 Privacy and Security Considerations

FL's privacy-preserving nature is critical for financial applications, but risks remain. Adversaries could exploit model updates to infer client data (e.g., via model inversion attacks), particularly in imbalanced datasets like Client 3's. Differential privacy could mitigate this by adding noise to updates, though it may reduce accuracy.

Secure aggregation protocols, such as homomorphic encryption, could further protect parameter exchanges but increase computational costs. Future work should

explore these trade-offs in regulated financial environments to ensure compliance with standards like GDPR or CCPA [45].

3.9 Limitations and Future Work

The study's limitations include the small dataset (690 instances), which limits scalability insights for large-scale financial systems. The 50-round evaluation may not capture long-term convergence dynamics, and unexplored heterogeneity types (e.g., quantity skew, temporal drift) restrict generalizability.

Future research should investigate multi-round convergence, larger datasets, and combined heterogeneity scenarios. Integrating differential privacy and secure aggregation will be critical to address privacy risks in real-world financial FL deployments.

4. Conclusions

This study evaluated five federated learning aggregation strategies—Federated Averaging (FedAvg), Weighted Average, Median Aggregation, Federated Proximal (FedProx), and Stochastic Controlled Averaging (SCAFFOLD)—using the Credit Approval dataset to assess credit risk under non-IID conditions, revealing that local models (e.g., Client 2's F1-score of 98.72%) outperformed global models, with Median Aggregation achieving the highest global F1-score (97.85%) and recall (80.6% vs. 72.3% for FedAvg and Weighted Average). This highlights Median Aggregation's robustness in handling data heterogeneity, offering practical value for privacy-preserving credit risk modeling in financial institutions while advancing theoretical understanding of FL in skewed data environments; however, limitations such as the small dataset (690 instances), 50 aggregation rounds, and lack of advanced privacy mechanisms suggest future research into larger-scale datasets, multi-round convergence, and differential privacy integration to enhance regulatory compliance and model scalability.

Acknowledgements

The authors would like to thank the Directorate of Research and Community Service (PPM) of Telkom University that has supported this research.

References

[1] Tolulope Esther Edunjobi and Opeyemi Abayomi Odejide, "Theoretical frameworks in AI for credit risk assessment: Towards banking efficiency and accuracy," *Int. J. Sci. Res. Updates*, vol. 7, no. 1, pp. 092–102, Mar. 2024, doi: 10.53430/ijsru.2024.7.1.0030.

[2] I. Gasmi, S. Neji, N. Mansouri, and M. Soui, "Bank credit risk prediction using machine learning model," *Neural Comput & Applic*, vol. 37, no. 16, pp. 10333–10350, June 2025, doi: 10.1007/s00521-025-11044-5.

[3] S. Shi, R. Tse, W. Luo, S. D'Addona, and G. Pau, "Machine learning-driven credit risk: a systemic review," *Neural Comput & Applic*, vol. 34, no. 17, pp. 14327–14339, Sept. 2022, doi: 10.1007/s00521-022-07472-2.

[4] V. S. Naresh, M. Thamarai, and S. Reddi, "Privacy-Preserving Bank Loan Approval with Federated Learning: A Secure and Collaborative Multi-Institutional Prediction Framework," *Comput Econ*, May 2025, doi: 10.1007/s10614-025-10978-1.

[5] J. A. Kumar and S. R. Babu, "Enhancing Credit Scoring with Alternative Data and Machine Learning for Financial Inclusion," 2025.

[6] C. Xie and S. Hu, "Open banking: an early review," *JIDE*, vol. 4, no. 2, pp. 73–82, July 2024, doi: 10.1108/JIDE-03-2024-0009.

[7] T. L. Nguyen *et al.*, "Blockchain-Empowered Trustworthy Data Sharing: Fundamentals, Applications, and Challenges," *ACM Comput. Surv.*, vol. 57, no. 8, pp. 1–36, Aug. 2025, doi: 10.1145/3718082.

[8] M. M. Hossain, M. Mamun, A. Munir, M. M. Rahman, and S. H. Chowdhury, "A Secure Bank Loan Prediction System by Bridging Differential Privacy and Explainable Machine Learning," *Electronics*, vol. 14, no. 8, p. 1691, Apr. 2025, doi: 10.3390/electronics14081691.

[9] K. Lazaros, D. E. Koumadorakis, A. G. Vrahatis, and S. Kotsiantis, "Federated Learning: Navigating the Landscape of Collaborative Intelligence," *Electronics*, vol. 13, no. 23, p. 4744, Nov. 2024, doi: 10.3390/electronics13234744.

[10] C. Ma, H. Zhao, K. Zhang, L. Zhang, and H. Huang, "A federated supply chain finance risk control method based on personalized differential privacy," *Egyptian Informatics Journal*, vol. 31, p. 100704, Sept. 2025, doi: 10.1016/j.eij.2025.100704.

[11] A. Rohendi and D. B. Kharisma, "Personal data protection in fintech: A case study from Indonesia," *J. Infras. Policy. Dev.*, vol. 8, no. 7, p. 4158, July 2024, doi: 10.24294/jipd.v8i7.4158.

[12] B. Prasetyo, I. G. A. K. R. Handayani, and A. Sulistiyono, "Data Protection Laws in Indonesia: Navigating Privacy in the Digital Age," *Side: Scientific Development Journal*, vol. 2, no. 1, pp. 9–16, Feb. 2025, doi: 10.59613/petfxv64.

[13] A. B. Wiedyasari and Wardah Yuspin, "Perlindungan Konsumen Data Pribadi Nasabah Bank Syariah Indonesia ditinjau dari POJK Nomor 6/POJK.07/2022," *ULREV*, vol. 8, no. 1, Apr. 2024, doi: 10.29303/ulrev.v8i1.331.

[14] R. A. Kasri, B. S. Indrastomo, N. D. Hendranastiti, and M. B. Prasetyo, "Digital payment and banking stability in emerging economy with dual banking system," *Heliyon*, vol. 8, no. 11, p. e11198, Nov. 2022, doi: 10.1016/j.heliyon.2022.e11198.

[15] M.-S. Jameaba, "Digitization Revolution, FinTech Disruption, and Financial stability: Using the Case of Indonesian Banking Ecosystem to highlight wide-ranging digitization opportunities and major challenges," July 16, 2020, *Social Science Research Network, Rochester, NY*: 3529924. doi: 10.2139/ssrn.3529924.

[16] A. Perdana, S. Arifin, and N. Quadrianto, "Algorithmic trust and regulation: Governance, ethics, legal, and social implications blueprint for Indonesia's central banking," *Technology in Society*, vol. 81, p. 102838, June 2025, doi: 10.1016/j.techsoc.2025.102838.

[17] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," Feb. 13, 2019, *arXiv*: arXiv:1902.04885. doi: 10.48550/arXiv.1902.04885.

[18] C. Wu, "Data privacy: From transparency to fairness," *Technology in Society*, vol. 76, p. 102457, Mar. 2024, doi: 10.1016/j.techsoc.2024.102457.

[19] Z. Wang, J. Xiao, L. Wang, and J. Yao, "A novel federated learning approach with knowledge transfer for credit scoring," *Decision Support Systems*, vol. 177, p. 114084, Feb. 2024, doi: 10.1016/j.dss.2023.114084.

[20] S. Pingulkar and D. Pawade, "Federated Learning Architectures for Credit Risk Assessment: A Comparative Analysis of Vertical, Horizontal, and Transfer Learning Approaches," in *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, Pune, India: IEEE, Oct. 2024, pp. 1–7. doi: 10.1109/ICBDS61829.2024.10837430.

[21] S. K. Aljunaid, S. J. Almheiri, H. Dawood, and M. A. Khan, "Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection," *JRFM*, vol. 18, no. 4, p. 179, Mar. 2025, doi: 10.3390/jrfm18040179.

[22] S. Mohammed *et al.*, "The effects of data quality on machine learning performance on tabular data," *Information Systems*, vol. 132, p. 102549, July 2025, doi: 10.1016/j.is.2025.102549.

[23] Z. Lu, H. Pan, Y. Dai, X. Si, and Y. Zhang, "Federated Learning With Non-IID Data: A Survey," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19188–19209, June 2024, doi: 10.1109/JIOT.2024.3376548.

[24] R. Zhang, Y. Zhang, Y. Zhao, B. Jia, and W. Lian, "FedCVG: a two-stage robust federated learning optimization algorithm," *Sci Rep*, vol. 15, no. 1, p. 18357, May 2025, doi: 10.1038/s41598-025-02722-4.

[25] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020, doi: 10.1109/MSP.2020.2975749.

[26] T. Yu, E. Bagdasaryan, and V. Shmatikov, "Salvaging Federated Learning by Local Adaptation," Mar. 03, 2022, *arXiv*: arXiv:2002.04758. doi: 10.48550/arXiv.2002.04758.

[27] Y. Chang, X. Shi, X. Zhao, Z. Chen, and D. Ma, "Dual prompt personalized federated learning in foundation models," *Sci Rep*, vol. 15, no. 1, p. 28026, July 2025, doi: 10.1038/s41598-025-11864-4.

[28] J. Meng, J. Chen, and D. Zhao, "Federated learning and information sharing between competitors with different training effectiveness," *Journal of Economy and Technology*, vol. 3, pp. 1–9, Nov. 2025, doi: 10.1016/j.ject.2024.12.003.

[29] S. Abdulrahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5476–5497, Apr. 2021, doi: 10.1109/JIOT.2020.3030072.

[30] S. Wang *et al.*, "Adaptive Federated Learning in Resource Constrained Edge Computing Systems," *IEEE J. Select. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, June 2019, doi: 10.1109/JSAC.2019.2904348.

[31] Z. Jovanovic, Z. Hou, K. Biswas, and V. Muthukkumarasamy, "Robust integration of blockchain and explainable federated learning for automated credit scoring," *Computer Networks*, vol. 243, p. 110303, Apr. 2024, doi: 10.1016/j.comnet.2024.110303.

[32] M. García-Márquez, N. Rodríguez-Barroso, M. V. Luzón, and F. Herrera, "Improving (α , f) -Byzantine resilience in federated learning via layerwise aggregation and cosine distance," *Knowledge-Based Systems*, vol. 326, p. 114004, Sept. 2025, doi: 10.1016/j.knosys.2025.114004.

[33] T. Wang, Z. Zheng, and F. Lin, "Federated learning framework based on trimmed mean aggregation rules," *Expert Systems with Applications*, vol. 270, p. 126354, Apr. 2025, doi: 10.1016/j.eswa.2024.126354.

[34] J. Cui, Y. Li, Q. Zhang, Z. He, and S. Zhao, "A Federated Learning Framework Using FedProx Algorithm for Privacy-Preserving Palmprint Recognition," in *Biometric Recognition*, S. Yu, W. Jia, X. Shu, X. Yuan, J. Gui, J. Tang, C. Shan, and Q. Liu, Eds., Singapore: Springer Nature Singapore, 2025, pp. 187–196.

[35] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., in Proceedings of Machine Learning Research, vol. 119. PMLR, July 2020, pp. 5132–5143. [Online]. Available: https://proceedings.mlr.press/v119/karimireddy20a.html

[36] Q. Li, B. He, and D. Song, "Model-Contrastive Federated Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, June 2021, pp. 10708–10717. doi: 10.1109/CVPR46437.2021.01057.

[37] W. Wu, L. He, W. Lin, R. Mao, C. Maple, and S. Jarvis, "SAFA: A Semi-Asynchronous Protocol for Fast Federated Learning With Low Overhead," *IEEE Trans. Comput.*, vol. 70, no. 5, pp. 655–668, May 2021, doi: 10.1109/TC.2020.2994391.

[38] I. Kholod *et al.*, "Open-Source Federated Learning Frameworks for IoT: A Comparative Review and Analysis," *Sensors*, vol. 21, no. 1, p. 167, Jan. 2021, doi: 10.3390/s21010167.

[39] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Comput Stat*, vol. 37, no. 5, pp. 2671–2692, Nov. 2022, doi: 10.1007/s00180-022-01207-6.

[40] M. V. Luzón *et al.*, "A Tutorial on Federated Learning from Theory to Practice: Foundations, Software Frameworks, Exemplary Use Cases, and Selected Trends," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 4, pp. 824–850, Apr. 2024, doi: 10.1109/JAS.2024.124215.

[41] Z. Liu, Z. Zhang, H. Yang, G. Wang, and Z. Xu, "An innovative model fusion algorithm to improve the recall rate of peer-to-peer lending default customers," *Intelligent Systems with Applications*, vol. 20, p. 200272, Nov. 2023, doi: 10.1016/j.iswa.2023.200272.

[42] T. M. Alam *et al.*, "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets," *IEEE Access*, vol. 8, pp. 201173–201198, 2020, doi: 10.1109/ACCESS.2020.3033784.

[43] C. N. Nwafor, O. Nwafor, and S. Brahma, "Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach," *Sci Rep*, vol. 14, no. 1, p. 25174, Oct. 2024, doi: 10.1038/s41598-024-75026-8.

[44] G. Mo, G. Zhang, C. Tan, W. Zhang, and Y. Guo, "Reassessment of Corporate Credit Risk Identification: Novel Discoveries from Integrated Machine Learning Models," *Comput Econ*, Dec. 2024, doi: 10.1007/s10614-024-10801-3.

[45] M. Taufiq and A. S. Kenyo, "The Legal Protection of Personal Data in the Digital Era: A Comparative Study of Indonesian Law and the GDPR," *International Journal of Business, Law, and Education*, vol. 6, no. 2, pp. 1260–1268, Aug. 2025, doi: 10.56442/ijble.v6i2.1178.