Journal of
Cybersecurity and Privacy

MDPI

Article

A Trust-Aware Incentive Mechanism for Federated Learning with Heterogeneous Clients in Edge Computing

Jiantao Xu, Chen Zhang, Liu Jin and Chunhua Su *

Graduate School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Fukushima Prefecture, Japan; d8252108@u-aizu.ac.jp (J.X.); d8252109@u-aizu.ac.jp (C.Z.); d8242103@u-aizu.ac.jp (L.J.)

* Correspondence: chsu@u-aizu.ac.jp

Abstract

Federated learning enables privacy-preserving model training across distributed clients, yet real-world deployments face statistical, system, and behavioral heterogeneity, which degrades performance and increases vulnerability to adversarial clients. Existing incentive mechanisms often neglect participant credibility, leading to unfair rewards and reduced robustness. To address these issues, we propose a Trust-Aware Incentive Mechanism (TAIM), which evaluates client reliability through a multi-dimensional trust model incorporating participation frequency, gradient consistency, and contribution effectiveness. A trust-weighted reward allocation is formulated via a Stackelberg game, and a confidence-based soft filtering algorithm is introduced to mitigate the impact of unreliable updates. Experiments on FEMNIST, CIFAR-10, and Sent140 demonstrate that TAIM improves accuracy by up to 4.1%, reduces performance degradation under adaptive attacks by over 35%, and ensures fairer incentive distribution with a Gini coefficient below 0.3. TAIM offers a robust and equitable FL framework suitable for heterogeneous edge environments.

Keywords: federated learning; incentive mechanism; trust management; edge computing; heterogeneous clients; robust aggregation

check for updates

Academic Editor: Christos Bouras

Received: 17 May 2025
Revised: 23 June 2025
Accepted: 23 June 2025
Published: 25 June 2025

Citation: Xu, J.; Zhang, C.; Jin, L.; Su, C. A Trust-Aware Incentive Mechanism for Federated Learning with Heterogeneous Clients in Edge Computing. *J. Cybersecur. Priv.* **2025**, *5*, 37. https://doi.org/10.3390/jcp5030037

1. Introduction

With the advancement of edge intelligence and data privacy regulations such as GDPR and HIPAA, achieving cross-device collaborative learning while preserving data locality has become a key challenge in building intelligent systems [1,2]. Federated learning (FL), with its distributed modeling paradigm of "keeping data local", has attracted significant attention in recent years [3,4]. By performing local training on edge devices and uploading only model parameters, FL effectively reduces the risk of data leakage and improves deployability in sensitive scenarios such as smart terminals, healthcare, and financial risk control, as illustrated in Figure 1.

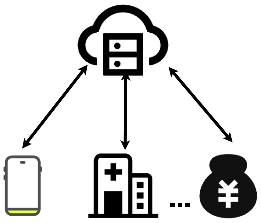

Figure 1. Typical application domains of federated learning in edge environments.

Nevertheless, the deployment of FL in real-world environments still faces critical challenges, particularly due to heterogeneity in statistical properties, system performance, and client behavior [5,6], as shown in Figure 2. First, data collected by different clients often come from distinct users, environments, or tasks, leading to highly non-independent and imbalanced distributions, which severely hinder global model convergence and generalization [7,8]. Second, due to the heterogeneous resource capacities of edge devices, issues such as communication delays, interrupted training, and intermittent connectivity commonly arise, widening performance gaps at the system level. More critically, client behaviors in open environments are uncontrollable and may include strategic participation, perturbed gradient uploads, or even collusive poisoning attacks, potentially destabilizing FL training or crashing the global model [9,10].

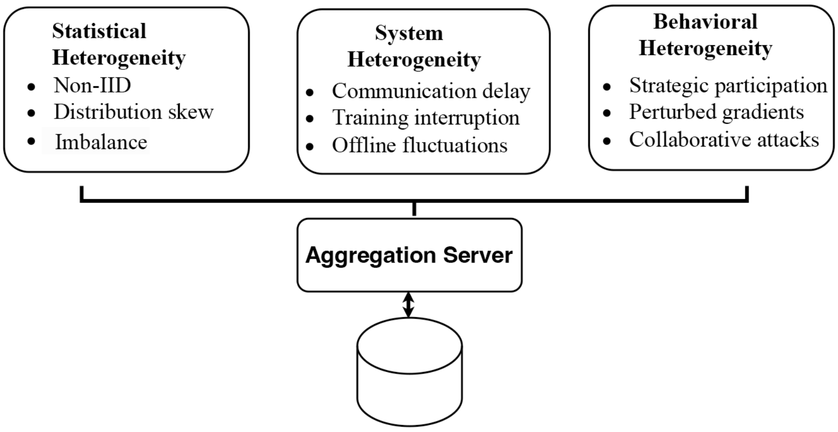

Figure 2. Three types of heterogeneity in federated learning: statistical, system, and behavioral.

To address these issues, extensive research has been conducted in robust aggregation, security mechanisms, trust management, and incentive design [11,12]. Robust aggregation algorithms such as the Krum [13] and the Trimmed Mean [14] effectively eliminate outlier updates and enhance system resilience, but most adopt rigid filtering strategies and lack tolerance for weak or high-variance clients. In terms of trust mechanisms, approaches such as FedTrust [15] and TrustFL [16] assess client credibility based on model similarity or behavioral patterns, yet are often decoupled from incentive schemes. Existing incentive mechanisms typically design reward allocation based on static indicators (e.g., data volume, training time), which struggle to detect opportunistic behaviors masked as honesty, and may result in incentive abuse under adversarial strategies [17,18].

It is particularly noteworthy that most existing works treat trust modeling, robust aggregation, and incentive mechanisms as separate submodules, lacking a systematic integrated modeling approach. Trust scores fail to feed back into resource allocation, and aggregation strategies do not respond adaptively to incentive feedback, leading to persistent internal incentive bias, behavior distortion, and training uncontrollability. Furthermore, due to the inherently dynamic nature of client behaviors, one-shot scoring or static thresholds fail to capture long-term trends and fluctuations, often resulting in misjudgments and misallocated rewards that compromise fairness and stability.

To tackle the above challenges, this paper proposes a unified Trust-Aware Incentive Mechanism(TAIM) that integrates client trust modeling, incentive feedback, and robust aggregation. Guided by the principle of “trust drives participation, incentive motivates resource investment, aggregation ensures robustness”, TAIM achieves deep coupling between strategic and optimization layers to enhance system robustness, fairness, and sustainable participation. The main contributions are as follows:

1. We design a dynamic trust modeling framework that integrates participation frequency, gradient consistency, and contribution effectiveness, capturing client behavioral trajectories and quantifying their stability and reliability;
2. A trust-driven incentive mechanism based on Stackelberg game theory is developed to derive the optimal resource allocation under equilibrium, guiding the rational strategy convergence of clients;
3. A confidence-aware smoothing aggregation algorithm is proposed, incorporating a soft filtering function to suppress and recover low-trust updates, effectively balancing robustness and diversity;
4. Extensive experiments under multiple non-IID datasets and adversarial scenarios validate the robustness, fairness, and convergence performance of the proposed method, along with comparative analysis against existing baselines.

The remainder of this paper is organized as follows: Section 2 reviews related works and key methodologies. Section 3 defines the modeling of client behavior in federated systems and formulates the problem. Section 4 elaborates on the trust modeling, algorithm design, and theoretical analysis of TAIM. Section 5 presents the experimental validation and comparative evaluation. Section 6 discusses limitations and future research directions.

2. Related Work

To build a trustworthy federated learning system tailored for heterogeneous edge environments, researchers have extensively explored three major directions in recent years: client incentive mechanisms, trust modeling, and robust aggregation algorithms. As illustrated in Figure 3, this section systematically reviews these research directions, identifies the key limitations in current approaches, and highlights the novelty and distinctiveness of this work.

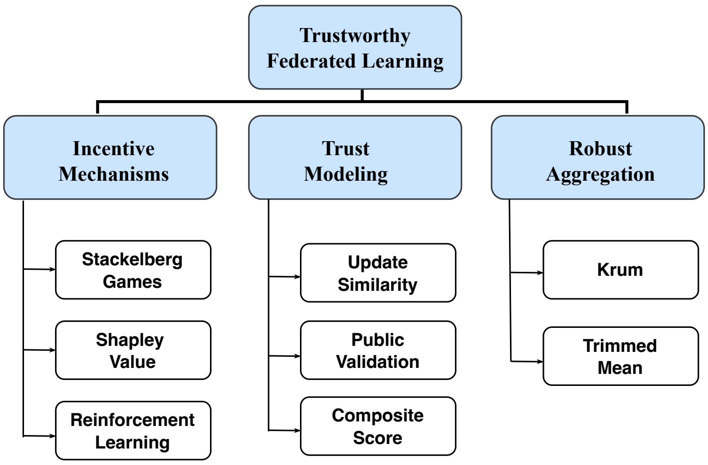

Figure 3. Taxonomy of research directions in trustworthy federated learning.

2.1. Incentive Mechanism Design in Federated Learning

In practical deployments, federated learning involves numerous clients with varying resource capabilities. Due to privacy concerns and resource consumption, clients are often reluctant to participate consistently and contribute high-quality updates [19,20]. Therefore, designing effective incentive mechanisms to enhance participation and contribution has become a key research challenge [21,22]. A comprehensive survey by Zhou [23] categorizes these mechanisms, highlighting the prominence of game theory, auction theory, and contract theory in addressing the challenge of motivating self-interested clients.

Early studies primarily focused on resource-driven incentive models. For example, Zhang et al. [24] proposed a Stackelberg game-based incentive model that rewards clients based on the quality of uploaded models rather than data volume or training time. While such methods can improve system efficiency to some extent, they often ignore heterogeneity in model quality and participation behavior, leading to excessive rewards for strategically behaving clients who appear to contribute.

To enhance fairness and robustness, some works introduced game theory and marginal contribution analysis. Huang et al. [25] and Xia et al. [26] designed demand-based reward allocation strategies using Shapley value estimation to evaluate each client's marginal improvement to the global model. Different game-theoretic approaches have also been explored; for instance, Pang et al. [27] designed an incentive auction for heterogeneous client selection, focusing on creating a market-based environment for efficient resource allocation. While effective, auction mechanisms differ from our approach by emphasizing competitive bidding rather than long-term trust cultivation. However, computing Shapley values is computationally intensive and lacks robustness in adversarial or non-ideal environments.

Recognizing the limitations of static or purely contribution-based metrics, recent works have shifted towards evaluating client behavior over time. For example, Al-Saedi et al. [28] proposed a method to predict client contributions by evaluating past behaviors, aiming to proactively select more reliable participants. This predictive approach complements our reactive trust-scoring mechanism, which assesses credibility after each round to dynamically adjust rewards and aggregation weights. Moreover, some research has explored reinforcement learning for dynamic incentive strategy generation. For instance, Ma et al. [29] introduced a deep reinforcement learning algorithm for incentive-based demand response which continuously optimizes interaction strategies using client states and feedback signals under incomplete information. Although such approaches improve adaptability, they often rely on global reward signals and struggle to capture individual trustworthiness or defend against strategic manipulation. The growing complexity and diversity of these mechanisms also underscore the need for standardized evaluation frameworks, as addressed by platforms like FLWB [30], which facilitate reproducible performance comparisons of FL algorithms.

In summary, current incentive mechanisms lack effective modeling and utilization of client behavioral credibility, resulting in misallocated or abused rewards. This study incorporates trust scores into the game-theoretic incentive function to construct a behavior-driven resource allocation mechanism, aiming to enhance system security and participation stability.

2.2. Trust Modeling and Robust Aggregation in Federated Learning

Security threats in federated learning primarily stem from clients uploading malicious or low-quality updates that degrade global model performance. To mitigate this, trust modeling and robust aggregation have become central research topics.

In trust modeling, various methods have been proposed to evaluate client credibility from different perspectives. FedTrust [31] calculates trust scores based on similarity among uploaded models and adjusts aggregation weights accordingly. TrustFL [32] dynamically adjusts client weights based on performance fluctuations on a public validation set and consistency of feature representations. Lyubchyk et al. [33] constructed a composite trust scoring system using multi-dimensional indicators to reflect long-term behavioral stability and reliability.

At the same time, robust aggregation algorithms provide effective defenses against poisoning attacks. Methods such as the Krum [34] and the Trimmed Mean [35] eliminate outlier gradients or select consistent subsets of updates to enhance robustness. However, most rely on static thresholds or distance-based filtering, which struggle to adapt to dynamic client behaviors and often overlook the strategic interactions among participants.

Recently, some works have attempted to couple trust mechanisms with robust aggregation. Perry et al. [36] introduced update correlation analysis for dynamic detection of collusive poisoning, and Abri et al. [37] modeled the trust learning process as a Markov decision process to recognize potential attack states. However, these approaches still neglect client responses to incentive feedback. Without a proper incentive regulation mechanism, the effectiveness of trust scoring and aggregation strategies can be compromised.

This work proposes a soft trust filtering mechanism that introduces a smoothing suppression function during aggregation to avoid penalizing edge clients with behavioral fluctuations. Moreover, the trust evaluation is coupled with the incentive allocation function to form a "high trust–high incentive–high participation" positive feedback loop, thereby enhancing adaptive defense and strategy stability.

2.3. Federated Modeling Mechanisms for Heterogeneous Edge Environments

In real-world scenarios, federated systems are commonly deployed across heterogeneous edge environments, facing non-ideal conditions such as device heterogeneity, resource imbalance, communication disruptions, and dynamic client availability. These factors significantly amplify challenges in fairness and robustness [38–40].

To address system-level heterogeneity, FedProx [41] introduces a regularization term into the local training objective to limit model divergence and improve global convergence. FedNova [42] normalizes updates to unify contribution scales across clients, while FedCS [43] proposes a bandwidth-aware client selection strategy to optimize training efficiency under communication constraints. Other works have focused on the timeliness of information, proposing Age of Information (AoI)-aware client selection or update weighting schemes to prioritize fresher updates from clients with better connectivity, thereby mitigating the negative impact of stragglers and stale models [22]. While these approaches have achieved progress in system optimization, they largely ignore the dynamics of client participation and strategic evolution, making them less effective in open edge environments with frequent malicious behaviors.

In particular, under the presence of strategic participants, clients may evade detection and manipulate rewards through mimicry attacks, intermittent poisoning, or frequent switching, ultimately undermining long-term stability [44,45]. Therefore, "behavioral trustworthiness" must be considered a core constraint in federated learning systems to enable multi-objective optimization under trustworthy guidance.

This work integrates edge heterogeneity modeling, dynamic trust evaluation, and incentive–response mechanisms to construct a trust-driven game-theoretic regulation framework at the strategic level. A soft suppression strategy is incorporated during aggregation, enabling robustness, incentive compatibility, and resource adaptation. This provides a systematic modeling paradigm for building secure and controllable federated learning systems at the edge.

3. System Model and Problem Formulation

In this section, we first formalize the basic structure of federated learning. We then construct a system modeling framework tailored for heterogeneity in edge computing environments by introducing dynamic trust modeling and incentive allocation mechanisms. Finally, we define a unified optimization objective. The overall system architecture is illustrated in Figure 4, and the notations used throughout the paper are summarized in Table 1.

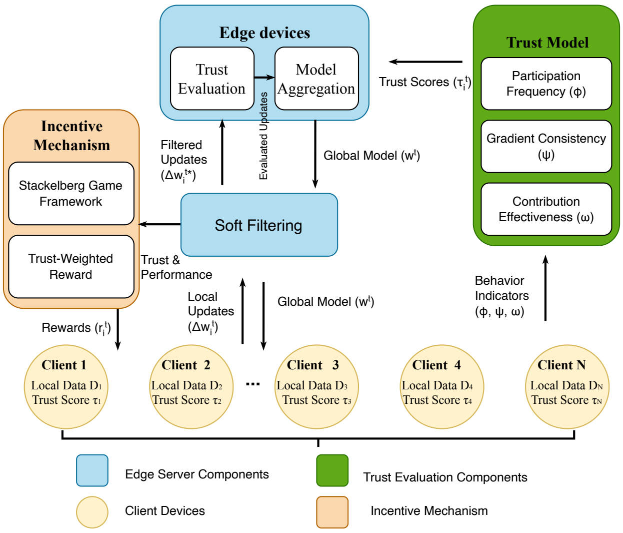

Figure 4. System model.

Table 1. Notations and definitions.

Notation	Definition
$\mathcal{C}$	Set of all clients
c_i	The i-th client
$\mathcal{D}_i$	Local dataset of client c_i
w_t	Global model parameters in round t
Δw_i^t	Model update from client c_i in round t
$\mathcal{S}_t$	Selected subset of clients in round t
τ_i^t	Trust score of client c_i in round t
$\tilde{\tau}_i^t$	Instantaneous trust score of client c_i
ϕ_i^t	Participation frequency of client c_i
ψ_i^t	Gradient consistency of client c_i
ω_i^t	Contribution effectiveness of client c_i
r_i^t	Incentive reward for client c_i in round t
R_t	Total incentive budget in round t
v_i^t	Raw contribution score of client c_i
$\hat{v}_i^t$	Adjusted contribution score based on validation
g_i^t	Validation error reduction from client update
α_i^t	Aggregation weight for client c_i
ζ_i^t	Deviation indicator of client update

3.1. Federated Learning Task Modeling

We consider a cross-device federated learning scenario consisting of a central server and a set of edge clients $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, where each client c_i owns a local dataset $\mathcal{D}_i$ that is typically highly non-IID. The goal of the system is to collaboratively train a robust and high-performing global model without sharing raw data.

During each communication round t, the server selects a subset of clients $\mathcal{S}_t \subseteq \mathcal{C}$ to participate in training. Each selected client performs the following steps:

(1) Downloads the current global model w_t;
(2) Trains locally using its own dataset and computes a model update Δw_i^t;
(3) Uploads Δw_i^t to the server for aggregation.

The server aggregates the received updates via weighted averaging to produce a new global model w_{t+1}, which is then broadcast to all clients, completing one round of training.

3.2. Heterogeneity and Behavior Modeling

Due to the inherent heterogeneity of edge devices, we model client states from three aspects:

Statistical heterogeneity: Differences in data distributions between $\mathcal{D}_i$ and $\mathcal{D}_j$ such as label imbalance and sample shift.

System heterogeneity: Variations in computational capabilities ρ_i and communication latency l_i across clients.

Behavioral heterogeneity: Behavioral anomalies such as unstable participation, gradient manipulation, or strategic uploads.

To capture dynamic behavioral characteristics, we introduce a Trust-Aware Incentive Mechanism to enable behavior perception and adaptive regulation during training.

3.3. Dynamic Trust Score Modeling

Each client c_i is assigned a trust score $\tau_i^t \in [0, 1]$ in round t, representing the overall reliability of its recent behavior. The score is updated via the following exponential decay rule:

$$\tau_i^t = \gamma \cdot \tau_i^{t-1} + (1 - \gamma) \cdot \tilde{\tau}_i^t, \tag{1}$$

where $\gamma \in [0, 1)$ is the memory decay coefficient, and $\tilde{\tau}_i^t$ is the instantaneous score defined as a weighted sum of behavioral indicators where

$$\tilde{\tau}_i^t = \lambda_1 \phi_i^t + \lambda_2 \psi_i^t + \lambda_3 \omega_i^t, \quad \sum_j \lambda_j = 1. \tag{2}$$

Specifically, ϕ_i^t is the participation frequency, or the proportion of active rounds within the past T rounds; ψ_i^t is the gradient consistency, or the cosine similarity between the local update and the global average direction; and ω_i^t is the contribution effectiveness, or the improvement of the validation error brought by the update.

The trust score serves not only as a behavioral descriptor but also as a control variable in both aggregation and incentive allocation.

3.4. Incentive Mechanism and Optimization Objective

Let r_i^t denote the incentive allocated by the server to client c_i in round t, subject to the following budget constraint:

$$\sum_{i \in \mathcal{S}_t} r_i^t \leq R_t, \quad r_i^t \geq 0, \tag{3}$$

Each client responds by investing local resources x_i to complete the training. The client's utility function is defined as follows:

$$U_i(x_i) = \eta \cdot \tau_i^t \cdot \frac{x_i}{\sum_j x_j} - (a_i x_i^2 + b_i x_i), \tag{4}$$

where the first term represents the trust-weighted share of incentives, and the second term captures the cost of resource consumption.

The server's objective is to design $\{r_i^t\}$ and aggregation weights $\{\alpha_i^t\}$ to ensure model quality while promoting high-trust participation and suppressing malicious updates.

4. Trust-Driven Incentive and Aggregation Mechanism

In this section, we systematically present the proposed Trust-Aware Incentive Mechanism (TAIM) and robust aggregation algorithm. The core objective is to achieve a triple control logic in federated learning: incentives should follow trust orientation, aggregation should enhance robustness, and client behaviors should form a positive-feedback convergence driven by incentives. Compared to traditional methods that decouple trust, incentive, and aggregation, our design achieves unified modeling of the three components and introduces corresponding game-theoretic solution strategies and weight adjustment mechanisms.

4.1. Trust-Aware Incentive Allocation Modeling

Incentive mechanism design is critical for motivating clients and ensuring behavioral quality in federated learning systems. Based on trust modeling, we use the trust score τ_i^t and the client's contribution v_i^t as the main factors in reward allocation, avoiding the manipulation that arises from using static metrics such as data volume or training epochs.

Specifically, the raw contribution v_i^t is defined as the normalized L_2 norm of the uploaded update as follows:

$$v_i^t = \frac{\|\Delta w_i^t\|_2}{\sum_{j \in \mathcal{S}_t} \|\Delta w_j^t\|_2}, \tag{5}$$

The incentive reward function is then defined as follows:

$$r_i^t = \frac{R_t \cdot \tau_i^t \cdot v_i^t}{\sum_{j \in \mathcal{S}_t} \tau_j^t \cdot v_j^t}, \tag{6}$$

This function satisfies the incentive compatibility and prioritizes high-trust, high-contribution clients under the budget constraint R_t.

To prevent manipulation through pseudo-contributions (e.g., uploading high-norm but low-benefit updates), we introduce a validation-based actual gain function g_i^t as follows:

$$g_i^t = \mathcal{L}(w_t) - \mathcal{L}(w_t + \Delta w_i^t), \tag{7}$$

The corrected contribution is computed as $\hat{v}_i^t = v_i^t \cdot \frac{g_i^t}{\mathcal{L}(w_t)}$ and is used in both incentive and aggregation processes.

4.2. Stackelberg Game-Based Solution Strategy

To model the strategic interaction between the server and clients, we adopt a Stackelberg game formulation. The server, as the leader, decides the total budget R_t and reward strategy $\{r_i^t\}$; the clients, as followers, choose their resource investment x_i to maximize utility.

Each client's utility function is defined as follows:

$$U_i(x_i) = \eta \cdot \tau_i^t \cdot \frac{x_i}{\sum_{j \in \mathcal{S}_t} x_j} - (a_i x_i^2 + b_i x_i), \tag{8}$$

where the first term denotes the trust-weighted incentive share, and the second term captures the cost of local resource usage.

The quadratic cost function $a_i x_i^2 + b_i x_i$ is widely adopted in economics and resource allocation models [46], as it ensures convexity and reflects diminishing returns—i.e., increasing cost per unit as resource consumption rises. This formulation facilitates closed-form analysis and captures the realistic non-linearity of energy or training time costs on edge devices.

It is worth noting that each client's best response depends on the global term $\sum_j x_j$, which is generally unknown in decentralized settings. We address this by assuming that the server provides an aggregated signal during each communication round. This approximation is consistent with many Stackelberg-based FL mechanisms [47], where clients respond based on coarse-grained information rather than full observability. In future work, distributed best-response estimation or local belief updates could be explored to eliminate this assumption.

This convex utility yields the following closed-form best-response function via the first-order derivative:

$$x_i^* = \frac{\eta \cdot \tau_i^t - b_i \cdot \sum_j x_j}{2a_i \cdot \sum_j x_j}. \tag{9}$$

The server's utility function is defined as the net benefit of the improved model accuracy minus the incentive cost, calculated as follows:

$$U_S = \Delta\mathcal{L}(w_t) - \lambda \cdot R_t, \tag{10}$$

where $\Delta\mathcal{L}(w_t)$ represents the loss reduction after aggregation and λ is a balancing coefficient controlling budget sensitivity. The server's objective is to choose R_t and $\{r_i^t\}$ such that U_S is maximized under budget constraints while also encouraging high-trust participation from clients.

Using backward induction, the server can derive the optimal reward strategy and establish a closed-loop linkage between incentive allocation and client behavior adaptation.

4.3. Trust-Guided Soft Aggregation Mechanism

Traditional robust aggregation methods often adopt rigid techniques such as outlier removal or hard thresholds, which may harm diversity and inclusiveness. To address this, we propose a trust-guided non-linear soft suppression strategy to attenuate the impact of low-trust updates via a continuous weighting function.

We define a sigmoid-based suppression function as follows:

$$\sigma(\tau) = \frac{1}{1 + e^{-k(\tau - \mu)}}, \tag{11}$$

where μ controls the suppression threshold and k controls the steepness. The final aggregation weight is determined by the trust score and corrected contribution as follows:

$$\alpha_i^t = \frac{\sigma(\tau_i^t) \cdot \hat{v}_i^t}{\sum_j \sigma(\tau_j^t) \cdot \hat{v}_j^t}. \tag{12}$$

The global model update becomes the following:

$$w_{t+1} = w_t + \sum_{i \in \mathcal{S}_t} \alpha_i^t \cdot \Delta w_i^t, \tag{13}$$

This aggregation scheme suppresses the influence of malicious updates while allowing low-trust clients to be re-evaluated and regain weight, enhancing long-term fairness and convergence.

To ensure practical deployability, the sigmoid-based suppression function is implemented using precomputed lookup tables or approximate activation functions to avoid runtime overhead. Similarly, trust score updates are server-side vector operations with minimal cost. These design choices ensure that the trust-aware mechanism does not introduce significant delays compared to standard aggregation.

4.4. Robustness Enhancement and Anomaly Detection Mechanisms

To defend against attackers mimicking trustworthy patterns or frequently switching strategies, we introduce the following two robustness enhancement modules:

(1) The Deviation Penalty Mechanism: Here, the relative deviation of the updates is defined as follows:

$$\zeta_i^t = \frac{\|\Delta w_i^t - \bar{\Delta} w_t\|_2}{\|\bar{\Delta} w_t\|_2}, \tag{14}$$

If $\zeta_i^t > \epsilon$, the trust score is penalized as follows:

$$\tau_i^t \leftarrow \tau_i^t \cdot \exp(-\beta \cdot \zeta_i^t), \tag{15}$$

(2) Sliding Window-Based Trust Correction: A sliding window tracks the fluctuation of client trust scores. If a client exhibits drastic, non-monotonic variations, we slow down the growth of its aggregation weight to prevent short-term strategic speculation from receiving high incentives.

These mechanisms improve the behavioral sensitivity and anomaly adaptability of the trust model, forming a layered defense framework for the system.

4.5. Integrated Federated Training Procedure

By combining trust modeling, game-theoretic incentive allocation, and soft aggregation, the TAIM training process is summarized as follows (see Algorithm 1).

Algorithm 1 TAIM: Trust-Aware Incentive and Robust Aggregation Algorithm

```
Require: Initial model $w_0$, total rounds $T$, initialize $\tau_i^0 = 0.5$
for each round $t = 1$ to $T$ do
    Server selects client set $\mathcal{S}_t$ and broadcasts $w_t$
    for each client $c_i \in \mathcal{S}_t$ do
        Train on $\mathcal{D}_i$ to obtain $\Delta w_i^t$
        Upload $\Delta w_i^t$; server computes $v_i^t$, $\omega_i^t$
        Update trust $\tau_i^t$, correct contribution $\hat{v}_i^t$
        Compute reward $r_i^t$, assign aggregation weight $\alpha_i^t$
    end for
    Aggregate: $w_{t+1} = w_t + \sum_i \alpha_i^t \cdot \Delta w_i^t$
end for
return $w_T$
```

This training process maintains the deployability of the standard FedAvg framework while constructing a complete trust–incentive–aggregation feedback loop. It offers enhanced security and strategy adaptiveness, making it particularly suitable for heterogeneous, dynamic, and untrusted open edge environments.

For complexity analysis, the computational overhead introduced by TAIM is manageable and does not alter the overall complexity of the federated learning process. Let $|\mathcal{S}_t|$ be the number of selected clients per round and d be the dimensionality of the model parameters. The primary computations in TAIM include the following: (1) Trust score update: Calculating gradient consistency (cosine similarity) for each client against the average update requires $O(|\mathcal{S}_t| \cdot d)$ operations. Other components are $O(1)$ per client. (2) Incentive allocation: This step involves normalization and summations over $|\mathcal{S}_t|$ clients, resulting in an overhead of $O(|\mathcal{S}_t|)$. (3) Soft aggregation: Computing the aggregation weights also requires $O(|\mathcal{S}_t|)$ operations. The final weighted aggregation of the model updates remains $O(|\mathcal{S}_t| \cdot d)$. Therefore, the total complexity per round is dominated by model-related vector operations, remaining at $O(|\mathcal{S}_t| \cdot d)$. The additional trust and incentive calculations introduce a constant factor increase in server-side computation but do not scale with model complexity in a prohibitive way, making TAIM practical for large-scale deployments, as supported by the empirical overhead analysis in Section 5.5.

5. Experimental Evaluation

To validate the effectiveness and robustness of the proposed TAIM in realistic federated learning environments, this section conducts a comprehensive empirical study across multiple representative datasets, attack types, and baseline methods. The evaluation focuses on the following research questions: (1) Can TAIM improve global model accuracy and convergence efficiency under heterogeneous client participation? (2) Is TAIM more robust against various attack types and capable of identifying adversarial behaviors? (3) Is the overhead of TAIM acceptable under realistic constraints such as limited resources and client uncertainty?

To ensure reproducibility, we describe the experimental setup, attack modeling, evaluation metrics, baseline methods, and result analysis in the following sections.

5.1. Experimental Setup and Datasets

We select three representative federated learning tasks covering image, text, and handwritten character classification under heterogeneous settings. This selection is motivated by the need to validate TAIM's effectiveness across diverse data modalities and types of heterogeneity that mirror real-world challenges. Specifically, the FEMNIST task, provided by the LEAF benchmark, involves handwritten character recognition with a highly non-IID user-based split that naturally models the statistical heterogeneity found in real-world user data. CIFAR-10 is a classical image classification dataset with 10 classes. We generate non-IID partitions using a Dirichlet distribution with $\alpha = 0.3$ to systematically control and evaluate the impact of statistical heterogeneity. Sent140 is a Twitter sentiment analysis task where each client reflects individual language styles, making it an ideal testbed for modeling the behavioral heterogeneity that TAIM is designed to manage. For all experiments, datasets are split into training, validation, and test sets with a ratio of 80:10:10, respectively, unless otherwise specified. This split ensures consistent evaluation across all baseline and proposed methods.

Model Configuration and Training Details

We adopt a CNN for CIFAR-10, a two-layer CNN for FEMNIST, and LSTM for Sent140 as local models. In each round, 10% of clients are randomly selected. Local training runs for five epochs using SGD with a learning rate of 0.01 and momentum 0.9. Trust parameters are set as $\gamma = 0.8$, $\lambda_1 = 0.3$, $\lambda_2 = 0.4$, $\lambda_3 = 0.3$. The sigmoid suppression function uses $k = 10$, $\mu = 0.5$.

5.2. Attack Modeling and Client Behavior Settings

To simulate realistic threats in federated environments, we inject a varying proportion of malicious clients per round and implement four types of adversarial behaviors: a Label Flip attack, which flips a portion of labels (e.g., class 1 to class 9) to mislead convergence; a Gaussian Noise attack, which adds Gaussian noise (mean 0, std 5) to gradients; an On–Off attack, which alternates between honest and malicious behavior to evade trust accumulation; and a Mimic attack, which imitates the gradients of high-trust clients to evade detection while disturbing aggregation.

The client composition is as follows: 10–30% are malicious clients with evenly distributed attack types; 20% are resource-constrained clients with reduced upload frequency; the rest are benign clients.

5.3. Evaluation Metrics

We evaluate all methods from four perspectives: accuracy, robustness, fairness, and detection ability. Final Accuracy (Acc) refers to the test accuracy achieved by the global model after convergence. Robustness Drop (RD) measures the performance degradation caused by adversarial clients. The Gini coefficient is used to quantify inequality in the reward distribution among clients. Recall and False-Positive Rate (FPR) are computed based on the identification of malicious clients. In addition, we report the total Training Time and Communication Volume to assess the system-level overhead and scalability.

To quantify the fairness of incentive distribution, the Gini coefficient G is calculated as follows:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|r_i - r_j|}{2n\sum_{i=1}^{n} r_i}, \tag{16}$$

where r_i is the cumulative reward received by client i, and n is the total number of clients. A lower Gini value indicates a more balanced and equitable incentive distribution.

This metric enables the assessment of whether TAIM's trust-guided reward allocation contributes to reducing reward centralization and maintaining fairness among heterogeneous clients.

5.4. Baseline Methods

We compare TAIM against several mainstream FL strategies: FedAvg (standard averaging) [48], FedProx (adds a proximal term to mitigate heterogeneity) [49], FedTrust (trust-weighted aggregation) [50], and Krum (robust aggregation against outliers) [51]. These baselines serve as the foundation for our comprehensive comparison and analysis.

5.5. Overall Performance Comparison

Figure 5 shows the final test accuracy under 10% and 30% attack ratios on FEMNIST and CIFAR-10. As the attack ratio increases, most methods experience a significant drop in accuracy. TAIM consistently achieves the best performance: 79.5% and 76.1% on FEMNIST; 75.2% and 71.9% on CIFAR-10. Compared to FedAvg and FedProx, TAIM shows up to 9.2% higher accuracy under heavy attacks, demonstrating its effectiveness in suppressing adversarial disturbances through trust modeling and incentive mechanisms.

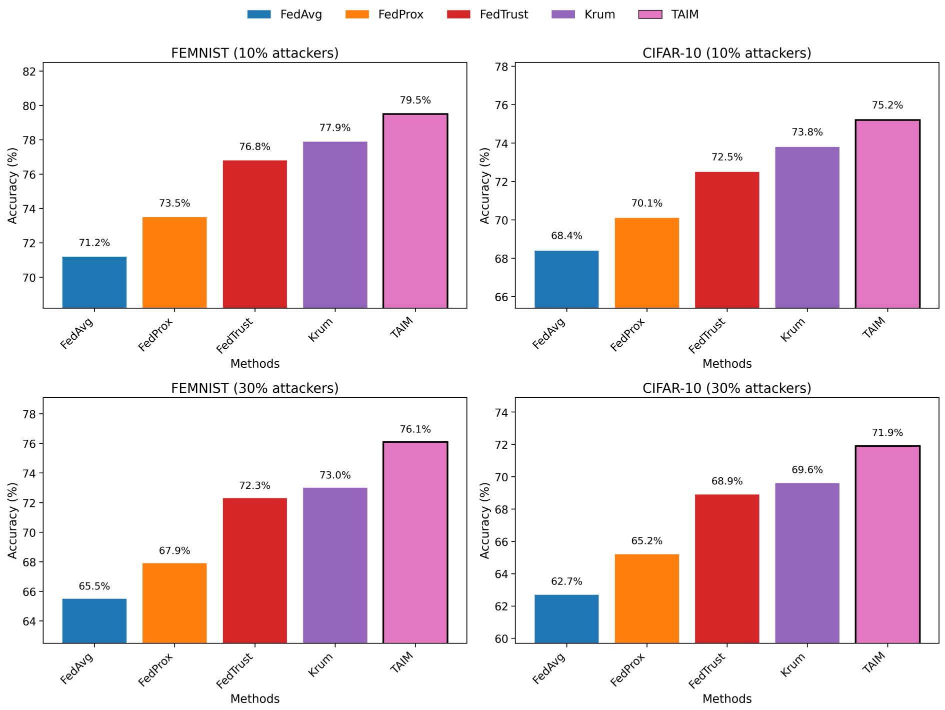

Figure 5. Final Accuracy (%) under different attack ratios.

Figure 6 shows model robustness under On–Off and Mimic attacks. Traditional methods like FedAvg and FedProx exhibit severe performance fluctuation, and even robust methods like FedTrust and Krum are affected by Mimic attacks. In contrast, TAIM maintains stable accuracy with minimal drop, outperforming Krum by 2.5% under Mimic attacks. This indicates that the multi-dimensional trust model in TAIM effectively filters disguised adversaries.

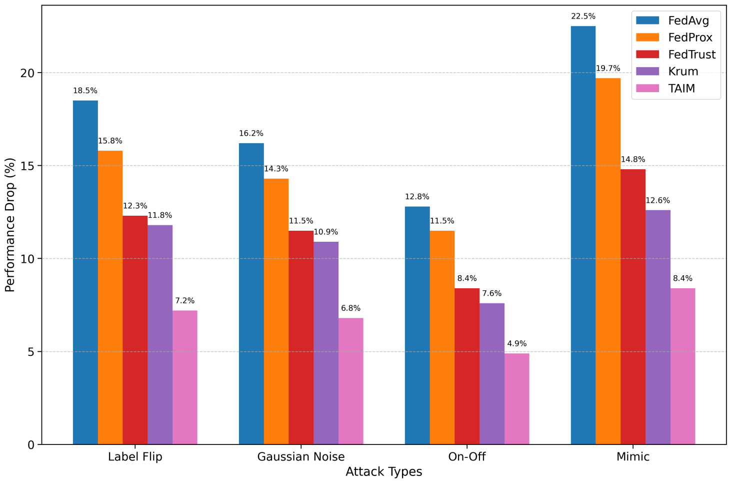

Figure 6. Model robustness under adaptive attacks.

Table 2 quantifies the detection performance. Traditional methods cannot detect malicious clients (denoted by "—"). FedTrust and Krum achieve 82.4% and 85.7% recall, with 9.6% and 12.1% FPR, respectively. TAIM outperforms them with 91.3% recall and only 5.8% FPR, owing to its use of participation, gradient consistency, and contribution effectiveness in trust modeling.

Table 2. Malicious client detection performance.

Method	Recall (%)	FPR (%)
FedAvg	—	—
FedProx	—	—
FedTrust	82.4	9.6
Krum	85.7	12.1
TAIM	91.3	5.8

Figure 7 presents the Gini coefficient during training, reflecting the incentive fairness. FedTrust and FedProx exhibit rising Gini values, indicating reward centralization. TAIM maintains a Gini coefficient < 0.3 throughout thanks to its balanced trust design that considers both participation and quality.

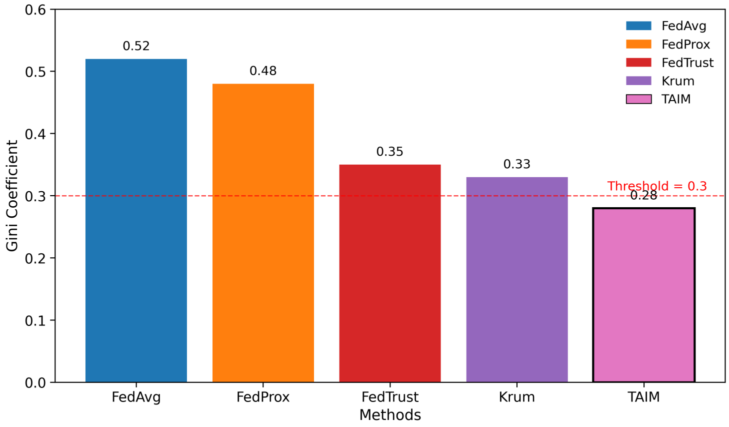

Figure 7. Gini coefficient during training.

Table 3 compares system overhead. Despite additional trust computation and soft aggregation, TAIM's cost remains comparable: only 0.5 MB more in communication and 0.16 s server aggregation time. Compared with Krum, which involves costly distance calculations, TAIM is equally efficient while providing higher robustness and fairness.

Table 3. System resource overhead comparison.

Method	Local Training (s)	Comm. Volume (MB)	Server Aggregation (s)
FedAvg	3.1	8.5	0.03
FedProx	3.4	8.5	0.05
FedTrust	3.9	8.5	0.04
Krum	4.2	8.5	0.15
TAIM	3.9	9.0	0.16

Despite integrating additional trust computation and soft aggregation mechanisms, TAIM remains computationally efficient. This is because (1) trust score updates rely on lightweight operations such as exponential smoothing and cosine similarity, which are low-cost and executed server-side; (2) the soft aggregation function is implemented via a fast sigmoid approximation or lookup table, avoiding expensive distance or sorting operations as in Krum; (3) TAIM does not require additional communication rounds or model retraining, and its trust parameters are updated in line with the standard training flow. Therefore, the overall additional overhead is marginal.

To eliminate the confounding effects caused by differences in model architecture across datasets, we conduct a control experiment with a unified lightweight CNN (two convolutional layers and two fully connected layers, 0.5 M parameters) on both FEMNIST and CIFAR-10. For Sent140, since the data are inherently sequential, we retain the same LSTM-based model for all methods to ensure consistency and fairness in comparison. The corresponding results are presented in Table 4.

Table 4. Accuracy comparison under unified model architecture.

Method	FEMNIST (%)	CIFAR-10 (%)	Sent140 (%)
FedAvg	72.4	69.1	71.0
FedProx	73.2	70.5	72.3
FedTrust	74.8	71.0	73.4
Krum	75.6	70.8	72.7
TAIM	78.5	74.9	75.8

These results show that TAIM consistently outperforms other methods across all three datasets even under unified or consistent model settings. This confirms that the observed performance gains are attributable to the trust-aware mechanism rather than model architecture differences, thereby strengthening the validity and generalizability of our conclusions.

5.6. Ablation Study and Parameter Sensitivity

Table 5 shows the ablation results of the trust components. Removing participation frequency ϕ drops the accuracy to 70.5%; removing gradient consistency ψ further lowers it to 67.9%; excluding contribution effectiveness ω gives 69.4%. This confirms the complementary roles of all three components in accurate trust assessment.

Table 5. Ablation study on trust dimensions (CIFAR-10 accuracy, %).

Configuration	Accuracy
Full TAIM	75.2
w/o ϕ (Participation Frequency)	70.5
w/o ψ (Gradient Consistency)	67.9
w/o ω (Contribution Effectiveness)	69.4

Figure 8 illustrates the sensitivity of the TAIM model to hyperparameter configurations, specifically analyzing the sigmoid suppression steepness parameter k and the trust component weights $(\lambda_1, \lambda_2, \lambda_3)$.

The left subfigure shows how the steepness parameter k in the sigmoid suppression function affects the final model accuracy. When k is set to a small value (e.g., 2.5), the model accuracy drops to 70.8% because the resulting trust scores are overly smooth, which fails to distinguish high- and low-trust clients effectively. As k increases, the accuracy gradually improves, reaching a peak of 75.2% at $k = 10$, then slightly decreases but remains relatively high (e.g., 74.6% at $k = 15$). This suggests that moderately enhancing the steepness helps amplify trust-based differentiation in aggregation, while excessively steep functions may overfit trust estimates and impair generalization. Overall, the system demonstrates robustness to a wide range of k values. The right subfigure examines the influence of different trust component weight configurations. The result shows that, with an equal weight setting (equal: 0.33, 0.33, 0.33), the model achieves 73.8% accuracy, indicating that each trust dimension independently contributes to performance. However, when one dimension is overly emphasized—such as participation frequency ($\lambda_1 = 0.6$)—the accuracy significantly drops to 72.1%. In contrast, emphasizing gradient consistency (consist. heavy: 0.2, 0.6, 0.2) leads to a better result of 74.3%, highlighting the importance of gradient-level behaviors in robust modeling. TAIM's default setting (0.3, 0.4, 0.3) achieves the highest accuracy of 75.2%, confirming the effectiveness of the proposed joint modeling strategy in balancing trust dimensions.

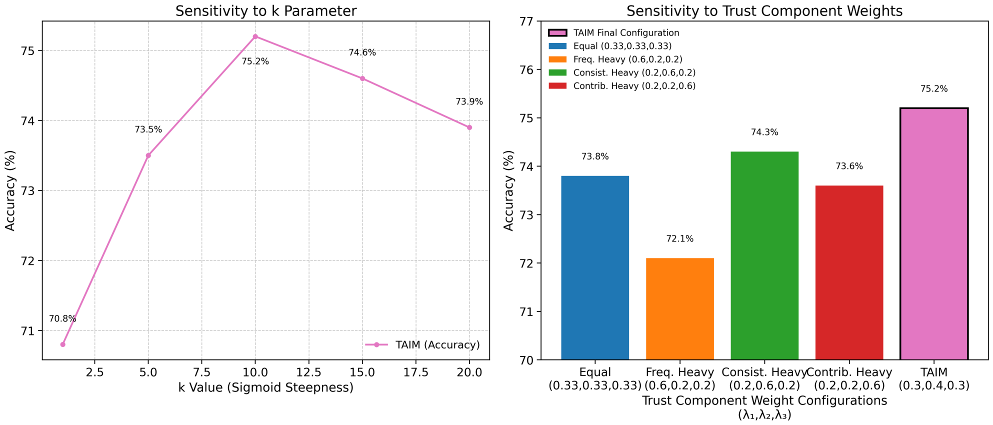

Figure 8. Accuracy impact under different k values and trust component weights.

6. Conclusions and Future Work

In this paper, we propose a unified framework called TAIM (Trust-Aware Incentive Mechanism) to address key challenges in federated learning in edge computing environments, including client heterogeneity, incentive imbalance, and adversarial robustness. TAIM integrates dynamic multi-dimensional trust modeling and incentive game theory, jointly modeling participation reliability, gradient consistency, and contribution effectiveness. Based on these trust scores, a Stackelberg game-based incentive allocation strategy and a trust-guided soft aggregation algorithm are designed.

TAIM achieves a balance between incentive rationality, fairness, and system robustness. Experimental results show that TAIM consistently outperforms baseline methods across various non-IID data and adversarial settings. It improves model accuracy (up to +6.1%), reduces robustness degradation (kept within 3%), and maintains low False-Positive Rates and stable fairness. Moreover, the proposed soft filtering strategy enhances system security while preserving client diversity, enabling long-term trust evolution and reputation recovery.

Despite the progress made, several limitations remain to be addressed in future work:

First, the current trust evaluation process relies on centralized server control, posing potential risks of data linkage leakage and single-point failure. Future research may explore decentralized technologies such as blockchain and secure multi-party computation (SMC) to enhance privacy and system resilience.

Second, the client response function in the Stackelberg game is simplified, assuming fully rational and immediate reactions. It does not account for constrained strategy spaces or delayed behavior. Future extensions may incorporate game learning methods (e.g., Q-learning Stackelberg) or evolutionary game theory to better capture real-world client dynamics.

Third, this work primarily focuses on single-task, unimodal scenarios. The applicability of TAIM to multi-modal federated learning (e.g., joint modeling of vision and language) remains unexplored. Future studies should investigate trust modeling across modalities to support collaborative heterogeneous tasks.

Lastly, the current trust mechanism is not integrated with differential privacy (DP), raising potential privacy leakage concerns. Future work may investigate how to ensure trust computation effectiveness and robustness under DP budget constraints.

In summary, TAIM provides an effective trust-driven solution for building future-ready federated learning systems characterized by openness, dynamism, and strategic participation. Ongoing efforts will focus on enhancing the generality, security, and distributed capability of the mechanism to enable wide deployment of trustworthy federated intelligence.

Author Contributions: Literature review, J.X.; methodology, J.X.; data curation, L.J.; writing—original draft preparation, J.X.; writing—review and editing, C.Z. and C.S.; supervision, C.S.; funding acquisition, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by JSPS Grant-in-Aid for Scientific Research (C) 23K11103.

Data Availability Statement: The datasets used in this study are publicly available. FEMNIST is provided by the LEAF benchmark suite (https://leaf.cmu.edu/, accessed on 20 April 2025). CIFAR-10 is available from the official CIFAR dataset page (https://www.cs.toronto.edu/~kriz/cifar.html, accessed on 20 April 2025). Sent140 can be accessed via the LEAF benchmark or directly at https://www.kaggle.com/kazanova/sentiment140 (accessed on 20 April 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Aminifar, A.; Shokri, M.; Aminifar, A. Privacy-preserving edge federated learning for intelligent mobile-health systems. *Future Gener. Comput. Syst.* **2024**, *161*, 625–637. [CrossRef]
2. Lazaros, K.; Koumadorakis, D.E.; Vrahatis, A.G.; Kotsiantis, S. Federated Learning: Navigating the Landscape of Collaborative Intelligence. *Electronics* **2024**, *13*, 4744. [CrossRef]
3. Ivanovic, M. Influence of Federated Learning on Contemporary Research and Applications. In Proceedings of the 2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Craiova, Romania, 4–6 September 2024 ; IEEE: Piscataway, NJ, USA , 2024; pp. 1–6.

4. Iyer, V.N. A review on different techniques used to combat the non-IID and heterogeneous nature of data in FL. *arXiv* **2024**, arXiv:2401.00809.
5. Hartmann, M.; Danoy, G.; Bouvry, P. FedPref: Federated Learning Across Heterogeneous Multi-objective Preferences. *ACM Trans. Model. Perform. Eval. Comput. Syst.* **2024**, *10*, 1–40. [CrossRef]
6. Chen, Y. Advancing Federated Learning by Addressing Data and System Heterogeneity. In Proceedings of the AAAI Symposium Series, Stanford, CA, USA, 25–27 March 2024; Volume 3, p. 294.
7. Liu, C.; Alghazzawi, D.M.; Cheng, L.; Liu, G.; Wang, C.; Zeng, C.; Yang, Y. Disentangling Client Contributions: Improving Federated Learning Accuracy in the Presence of Heterogeneous Data. In Proceedings of the 2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Wuhan, China, 21–24 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 381–387.
8. Taghiyarrenani, Z.; Alabdallah, A.; Nowaczyk, S.; Pashami, S. Heterogeneous Federated Learning via Personalized Generative Networks. *arXiv* **2023**, arXiv:2308.13265.
9. Ma, C.; Li, J.; Ding, M.; Wei, K.; Chen, W.; Poor, H.V. Federated learning with unreliable clients: Performance analysis and mechanism design. *IEEE Internet Things J.* **2021**, *8*, 17308–17319. [CrossRef]
10. Xia, G.; Chen, J.; Huang, X.; Yu, C.; Zhang, Z. FL-PTD: A Privacy Preserving Defense Strategy Against Poisoning Attacks in Federated Learning. In Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, 26–30 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 735–740.
11. Manzoor, H.U.; Shabbir, A.; Chen, A.; Flynn, D.; Zoha, A. A survey of security strategies in federated learning: Defending models, data, and privacy. *Future Internet* **2024**, *16*, 374. [CrossRef]
12. Zhang, H.; Elsayed, M.; Bavand, M.; Gaigalas, R.; Ozcan, Y.; Erol-Kantarci, M. Federated learning with dual attention for robust modulation classification under attacks. In Proceedings of the ICC 2024-IEEE International Conference on Communications, Denver, CO, USA, 9–13 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 5238–5243.
13. Wang, T.; Zheng, Z.; Lin, F. Federated learning framework based on trimmed mean aggregation rules. *Expert Syst. Appl.* **2025**, *270*, 126354. [CrossRef]
14. Nabavirazavi, S.; Taheri, R.; Shojafar, M.; Iyengar, S.S. Impact of aggregation function randomization against model poisoning in federated learning. In Proceedings of the 22nd IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2023, Exeter, UK, 1–3 November 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024; pp. 165–172.
15. Xiong, A.; Chen, Y.; Chen, H.; Chen, J.; Yang, S.; Huang, J.; Li, Z.; Guo, S. A truthful and reliable incentive mechanism for federated learning based on reputation mechanism and reverse auction. *Electronics* **2023**, *12*, 517. [CrossRef]
16. Han, K.; Zhang, G.; Yang, L.; Bai, J. Client dependability evaluation in federated learning framework. In Proceedings of the Third International Conference on Communications, Information System, and Data Science (CISDS 2024), Nanjing, China, 22–24 November 2024; SPIE: Bellingham, WA, USA, 2025; Volume 13519, pp. 71–79.
17. Chen, Z.; Zhang, H.; Li, X.; Miao, Y.; Zhang, X.; Zhang, M.; Ma, S.; Deng, R.H. FDFL: Fair and discrepancy-aware incentive mechanism for federated learning. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 8140–8154. [CrossRef]
18. Wu, R.; Chen, Y.; Tan, C.; Luo, Y. MDIFL: Robust federated learning based on malicious detection and incentives. *Appl. Sci.* **2023**, *13*, 2793. [CrossRef]
19. Yellampalli, S.S.; Chalupa, M.; Wang, J.; Song, H.J.; Zhang, X.; Yue, H.; Pan, M. Client Selection in Federated Learning: A Dynamic Matching-Based Incentive Mechanism. In Proceedings of the 2024 International Conference on Computing, Networking and Communications (ICNC), Big Island, HI, USA, 19–24 February 2024; IEEE: Piscataway, NJ, USA, 2024, pp. 989–993.
20. Han, J.; Khan, A.F.; Zawad, S.; Anwar, A.; Angel, N.B.; Zhou, Y.; Yan, F.; Butt, A.R. Tiff: Tokenized incentive for federated learning. In Proceedings of the 2022 IEEE 15th International Conference on Cloud Computing (CLOUD), Barcelona, Spain, 10–16 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 407–416.
21. Wenya, L.; Bo, L.; Weiwei, L.; Yuanchao, Y. Survey of incentive mechanism for federated learning. *Comput. Sci.* **2022**, *49*, 7.
22. Ling, X.; Li, R.; Ouyang, T.; Chen, X. Time is Gold: A Time-Dependent Incentive Mechanism Design for Fast Federated Learning. In Proceedings of the 2022 IEEE/CIC International Conference on Communications in China (ICCC), Foshan, China, 11–13 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1038–1043.
23. Zhou, Y. A Survey of Incentive Mechanisms for Federated Learning. *Appl. Comput. Eng.* **2024**, *10*, 1035–1044. [CrossRef]
24. Zhang, W.; Wang, Q.; Zhao, H.; Xia, W.; Zhu, H. Incentivizing Quality Contributions in Federated Learning: A Stackelberg Game Approach. In Proceedings of the 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), Singapore, 24–27 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5.
25. Huang, J.; Hong, C.; Chen, L.Y.; Roos, S. Is shapley value fair? improving client selection for mavericks in federated learning. *arXiv* **2021**, arXiv:2106.10734.

26. Xia, H.; Li, X.; Pang, J.; Liu, J.; Ren, K.; Xiong, L. P-Shapley: Shapley Values on Probabilistic Classifiers. *Proc. VLDB Endow.* **2024**, *17*, 1737–1750. [CrossRef]
27. Pang, J.; Yu, J.; Zhou, R.; Lui, J.C.S. An Incentive Auction for Heterogeneous Client Selection in Federated Learning. *IEEE Trans. Mob. Comput.* **2023**, *22*, 5733–5750. [CrossRef]
28. Al-Saedi, A.A. Contribution prediction in federated learning via client behavior evaluation. *Future Gener. Comput. Syst.* **2024**, *166*, 107639. [CrossRef]
29. Ma, S.; Liu, H.; Wang, N.; Xie, H.; Huang, L.; Li, H. Deep reinforcement learning for an incentive-based demand response model. In Proceedings of the 2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2), Chengdu, China, 11–13 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 246–250.
30. Casalicchio, E.; Esposito, S.; Al-Saedi, A.A. FLWB: A Workbench Platform for Performance Evaluation of Federated Learning Algorithms. In Proceedings of the 2023 IEEE International Workshop on Technologies for Defense and Security (TechDefense) , Rome, Italy, 20–22 November 2023.
31. Hsu, C.F.; Huang, J.L.; Liu, F.H.; Chang, M.C.; Chen, W.C. Fedtrust: Towards building secure robust and trustworthy moderators for federated learning. In Proceedings of the 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), Virtual, 2–4 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 318–323.
32. Zhang, X.; Li, F.; Zhang, Z.; Li, Q.; Wang, C.; Wu, J. Enabling execution assurance of federated learning at untrusted participants. In Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications, Virtual, 6–9 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1877–1886.
33. Lyubchyk, L.; Grinberg, G.; Konokhova, Z.; Yamkovyi, K. Composite Indicators Building Based on Concordant of Expert-Statistical Information Using Biased Ridge Kernel Regression. In Proceedings of the 2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Dortmund, Germany, 7–9 September 2023; IEEE: Piscataway, NJ, USA, 2023; Volume 1, pp. 617–620.
34. Yang, H.; Gu, D.; He, J. A robust and efficient federated learning algorithm against adaptive model poisoning attacks. *IEEE Internet Things J.* **2024**, *11*, 16289–16302. [CrossRef]
35. Yazdinejad, A.; Dehghantanha, A.; Karimipour, H.; Srivastava, G.; Parizi, R.M. A robust privacy-preserving federated learning model against model poisoning attacks. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 6693–6708. [CrossRef]
36. Perry, S.; Jiang, Y.; Zhong, F.; Huang, J.; Gyawali, S. DynaDetect2. 0: Improving Detection Accuracy of Data Poisoning Attacks. In Proceedings of the 2024 Cyber Awareness and Research Symposium (CARS), Grand Forks, ND, USA, 28–29 October 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8.
37. Abri, F.; Zheng, J.; Namin, A.S.; Jones, K.S. Markov decision process for modeling social engineering attacks and finding optimal attack strategies. *IEEE Access* **2022**, *10*, 109949–109968. [CrossRef]
38. Malik, P.; Alirajpurwala, T.; Kaushal, S.; Patidar, T.; Indore, I.I.I.; Padlak, S. Scalability and robustness of federated learning systems: Challenges and solutions. *Int. J. Sci. Res. Eng. Manag. IJSREM* **2024**, *8*. [CrossRef]
39. Abdelmoniem, A.M.; Ho, C.Y.; Papageorgiou, P.; Canini, M. A comprehensive empirical study of heterogeneity in federated learning. *IEEE Internet Things J.* **2023**, *10*, 14071–14083. [CrossRef]
40. Abdelmoniem, A.M.; Ho, C.Y.; Papageorgiou, P.; Canini, M. Empirical analysis of federated learning in heterogeneous environments. In Proceedings of the 2nd European Workshop on Machine Learning and Systems, Rennes France, 5–8 April 2022; pp. 1–9.
41. Yu, X.; He, Z.; Sun, Y.; Xue, L.; Li, R. The Effect of Personalization in FedProx: A Fine-grained Analysis on Statistical Accuracy and Communication Efficiency. *arXiv* **2024**, arXiv:2410.08934.
42. Kang, H.; Kim, M.; Lee, B.; Kim, H. FedAND: Federated learning exploiting consensus ADMM by nulling drift. *IEEE Trans. Ind. Inform.* **2024**, *20*, 9837–9849. [CrossRef]
43. Mahmoud, M.H.; Albaseer, A.; Abdallah, M.; Al-Dhahir, N. Federated learning resource optimization and client selection for total energy minimization under outage, latency, and bandwidth constraints with partial or no CSI. *IEEE Open J. Commun. Soc.* **2023**, *4*, 936–953. [CrossRef]
44. Antonioli, D.; Tippenhauer, N.O.; Rasmussen, K. Bias: Bluetooth impersonation attacks. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 18–20 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 549–562.
45. Ebrahimabadi, M.; Lalouani, W.; Younis, M.; Karimi, N. Countering PUF modeling attacks through adversarial machine learning. In Proceedings of the 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Tampa, FL, USA, 7–9 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 356–361.
46. Wang, B.; Sun, S.; Ren, W. Distributed continuous-time algorithms for optimal resource allocation with time-varying quadratic cost functions. *IEEE Trans. Control Netw. Syst.* **2020**, *7*, 1974–1984. [CrossRef]
47. Javaherian, S.; Turney, B.; Chen, L.; Tzeng, N.F. Incentive-Compatible Federated Learning with Stackelberg Game Modeling. *arXiv* **2025**, arXiv:2501.02662.

48. Collins, L.; Hassani, H.; Mokhtari, A.; Shakkottai, S. Fedavg with fine tuning: Local updates lead to representation learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10572–10586.
49. Mora, A.; Bujari, A.; Bellavista, P. Enhancing generalization in federated learning with heterogeneous data: A comparative literature review. *Future Gener. Comput. Syst.* **2024**, *157*, 1–15. [CrossRef]
50. Chakravarthy, V.; Bell, D.; Bhaskaran, S. Emergent Intrusion Detection System for Fog Enabled Smart Agriculture Using Federated Learning and Blockchain Technology: A Review. In Proceedings of the 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 13–15 April 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–7.
51. Yang, K.; Imam, N. Secure and Private Federated Learning: Achieving Adversarial Resilience through Robust Aggregation. *arXiv* **2025**, arXiv:2505.17226.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.