

Article

A Simulation-Driven Trust-Aware Federated Learning Framework for Robust Intelligent IoT Networks

Manuel J. C. S. Reis ^{1,*}, Carlos Serôdio ² and Frederico Branco ³

¹ Engineering Department and IEETA, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5000-801 Vila Real, Portugal

² Engineering Department and Center ALGORITMI, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5000-801 Vila Real, Portugal; cserodio@utad.pt

³ Engineering Department and INESC-TEC, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5000-801 Vila Real, Portugal; fbranco@utad.pt

* Correspondence: mcabral@utad.pt

Abstract

Federated learning (FL) has emerged as a promising paradigm for enabling distributed intelligence in Internet of Things (IoT) environments while preserving data privacy and reducing the need for centralized data collection. However, the practical deployment of FL in IoT scenarios remains challenging due to heterogeneous data distributions, unreliable communication conditions, and the presence of faulty or malicious edge devices that can disrupt collaborative training. These limitations can significantly degrade convergence stability and predictive performance, particularly in resource-constrained and intermittently connected networks. This paper proposes a simulation-driven trust-aware federated learning framework for robust intelligent IoT networks. The proposed approach incorporates a dynamic trust-based aggregation mechanism that adaptively weights client contributions based on the consistency of their local model updates with the global model state. In addition, a controlled IoT-oriented federated simulation environment is developed to emulate heterogeneous edge conditions, including non-independent and identically distributed (non-IID) data partitioning, adversarial model manipulation, and intermittent client connectivity caused by communication dropouts. Extensive multi-seed experiments were conducted on the UCI Human Activity Recognition (UCI HAR) dataset and complemented with an auxiliary CIFAR-10 convolutional neural network (CNN) validation scenario. The evaluation considered multiple adversarial settings, including sign-flip, Gaussian-noise, scaling, and label-flip attacks, as well as communication-dropout probabilities up to 50%. In contrast with the initial FedAvg-only evaluation, the revised experimental analysis includes comparisons with representative robust aggregation baselines, namely Median, Trimmed Mean, Krum, Multi-Krum, and an auxiliary Bulyan configuration. The experimental results demonstrate that the proposed Trust-FedAvg framework substantially improves robustness over conventional FedAvg and remains competitive with established robust aggregation strategies, particularly under directional model-manipulation attacks and intermittent-connectivity conditions. Under a 20% sign-flip attack on UCI HAR, the proposed method achieved a final test accuracy of 86.2%, whereas conventional FedAvg degraded to approximately 44.7%. Furthermore, under combined adversarial and intermittent-connectivity conditions with 50% communication dropout, Trust-FedAvg maintained a final accuracy of 57.0%, compared with 21.3% for FedAvg, 24.8% for Median, and 14.6% for Trimmed Mean. The additional experiments also show that Trust-FedAvg is not universally superior across all perturbation types: under severe Gaussian-noise attacks, coordinate-wise Median and Multi-Krum provided stronger ro-

Academic Editor: Xianpeng Wang

Received: 2 June 2026

Revised: 2 July 2026

Accepted: 6 July 2026

Published: 8 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

business in some settings. Overall, the results suggest that trust-aware aggregation can improve robustness against unreliable or malicious simulated clients while preserving a relatively simple aggregation procedure. Runtime measurements further indicate that the proposed method introduces only limited round-level overhead compared with FedAvg, while remaining simpler than more complex Byzantine-resilient alternatives. Further validation with real IoT deployments, additional sensor datasets, asynchronous communication models, energy profiling, and communication-overhead measurements is required to fully assess deployment feasibility in real IoT environments. The proposed framework provides a practical, extensible basis for the design and evaluation of resilient AI-enabled IoT networks operating under controlled but practically relevant edge-learning constraints.

Keywords: federated learning; Internet of Things; trust-aware aggregation; edge AI; adversarial robustness; non-IID learning; communication dropout; intelligent IoT networks

1. Introduction

The rapid expansion of the Internet of Things (IoT) has transformed modern digital ecosystems by enabling large-scale, interconnected networks of intelligent devices that can sense, process, and exchange information in real time. IoT technologies are increasingly deployed across domains such as smart manufacturing, healthcare, transportation, environmental monitoring, and industrial automation, where distributed intelligence and autonomous decision-making are becoming essential for scalable, adaptive operation [1,2].

At the same time, recent advances in artificial intelligence (AI) and machine learning (ML) have enabled IoT systems to move beyond simple sensing infrastructures toward intelligent edge-enabled environments capable of local inference and collaborative learning. However, conventional centralized ML approaches typically require large volumes of data to be transmitted from edge devices to centralized cloud servers, raising concerns about privacy, communication overhead, scalability, and energy consumption [3,4].

Federated learning (FL) has emerged as a promising paradigm for addressing these limitations by enabling distributed model training without requiring the direct sharing of raw local data. Instead, participating clients collaboratively train a shared global model by exchanging model updates or gradients with a central aggregation server [3]. In particular, the Federated Averaging (FedAvg) algorithm introduced by McMahan et al. has become the foundational framework for modern FL systems due to its communication efficiency and compatibility with decentralized edge learning scenarios. Since then, FL has attracted significant interest for privacy-preserving intelligent IoT applications, especially in environments characterized by distributed sensing and resource-constrained edge devices [1,5].

Several extensions of the original FedAvg framework have subsequently been proposed to address optimization instability and client heterogeneity under non-independent and identically distributed (non-IID) federated learning conditions. Representative examples include FedProx, which introduces proximal regularization to mitigate client drift; SCAFFOLD, which employs control variates to reduce optimization variance; and FedNova, which normalizes heterogeneous local updates to improve convergence stability under variable client participation [6–8].

Recent surveys have further emphasized the growing relevance of federated learning for edge intelligence, distributed IoT analytics, and privacy-preserving cyber-physical

systems, particularly in the context of large-scale heterogeneous sensor deployments and intelligent edge computing infrastructures [9,10].

Despite its advantages, practical FL deployment in IoT environments remains highly challenging. Unlike conventional centralized learning settings, IoT-based FL systems are inherently heterogeneous and dynamic. Participating edge devices often exhibit substantial variability in sensing quality, computational capabilities, communication reliability, energy availability, and local data distributions. In particular, the non-independent and identically distributed (non-IID) nature of local IoT data can significantly degrade the convergence behavior and stability of federated optimization processes [3,11]. Furthermore, IoT networks frequently operate under intermittent connectivity, where unreliable wireless communication and unstable edge participation can cause clients to fail to upload model updates during federated aggregation rounds.

In addition to these operational challenges, FL systems are also vulnerable to a wide range of adversarial threats because of their distributed, collaborative nature. Malicious or compromised participants may intentionally manipulate local model updates in order to disrupt convergence, poison the global model, or reduce predictive performance. Existing studies have demonstrated that attacks such as sign-flip manipulation, Gaussian perturbation, Byzantine behavior, and model poisoning can severely degrade the robustness and reliability of conventional FL aggregation strategies [12,13]. These vulnerabilities become particularly critical in IoT scenarios involving unreliable edge devices, limited supervision, and heterogeneous communication conditions.

To address these limitations, recent research has increasingly focused on robust, trust-aware federated learning mechanisms that reduce the influence of unreliable or malicious participants during aggregation. Trust-aware FL approaches typically estimate client reliability based on criteria such as behavioral consistency, update similarity, historical participation, or statistical contribution quality, and then adapt the aggregation weights accordingly [14–16]. Such approaches aim to improve convergence stability and adversarial robustness while maintaining the decentralized privacy-preserving characteristics of FL systems.

Nevertheless, several important research gaps remain insufficiently addressed in the current literature. First, many existing studies evaluate FL robustness using simplified experimental settings that assume stable communication conditions and homogeneous client participation, which do not fully reflect intermittently connected IoT deployments. Second, a large portion of prior work focuses primarily on adversarial robustness, without jointly analyzing the combined impact of malicious updates, non-IID client data, and communication dropout. Third, although robust aggregation methods such as Median, Trimmed Mean, Krum, Multi-Krum, and Bulyan provide important Byzantine-resilience mechanisms, they may exhibit different trade-offs in terms of accuracy, convergence stability, computational cost, and suitability under small or dynamically varying client participation. Fourth, trust-aware methods based on update similarity or model-distance consistency require careful evaluation under strongly non-IID settings, since statistically valid benign clients may naturally produce updates that deviate from the current global model. Consequently, there remains a need for lightweight and reproducible frameworks capable of jointly evaluating adversarial robustness, non-IID data heterogeneity, intermittent communication, statistical significance, and runtime overhead in IoT-oriented FL settings.

Motivated by these challenges, this paper proposes a simulation-driven trust-aware federated learning framework for robust intelligent IoT networks. The proposed approach incorporates a dynamic trust-based aggregation mechanism that adaptively weights client contributions according to the consistency of local model updates relative to the current global model state. In addition, a controlled IoT-oriented simulation environment is developed to emulate heterogeneous IoT deployments characterized by non-IID local da-

tasets, adversarial client behavior, and intermittent communication dropout conditions. The term controlled IoT-oriented simulation is used to indicate that the framework reproduces selected edge-learning constraints, namely non-IID client data, probabilistic communication dropout, and adversarial update manipulation. It does not model physical-layer communication, network latency, bandwidth constraints, device energy consumption, hardware failures, asynchronous communication, or deployment on real IoT devices.

Extensive multi-seed experiments are conducted on the UCI Human Activity Recognition (UCI HAR) dataset under multiple attack scenarios, including sign-flip, Gaussian-noise, scaling, and label-flip attacks, as well as varying communication-dropout probabilities. In response to the need for stronger comparative benchmarking, the proposed Trust-FedAvg framework is evaluated against conventional FedAvg and representative robust aggregation baselines, including Median, Trimmed Mean, Krum, Multi-Krum, and an auxiliary Bulyan configuration. Additional experiments analyze sensitivity to different Dirichlet non-IID levels, statistical significance across random seeds, round-level runtime overhead, and an auxiliary CIFAR-10 convolutional neural network (CNN) setting to assess behavior beyond the UCI HAR MLP configuration.

The results show that Trust-FedAvg substantially improves robustness over conventional FedAvg, particularly under directional manipulation attacks and intermittent-connectivity conditions. Under a 20% sign-flip attack on UCI HAR, Trust-FedAvg achieved a final accuracy of 86.2%, compared with 44.7% for FedAvg. Under the most severe combined setting, involving 20% sign-flip malicious clients and 50% communication dropout, Trust-FedAvg achieved a final accuracy of 57.0%, compared with 21.3% for FedAvg, 24.8% for Median, and 14.6% for Trimmed Mean. At the same time, the extended evaluation indicates that Trust-FedAvg is not universally superior across all attack types. In particular, under severe Gaussian-noise perturbations, coordinate-wise Median and Multi-Krum achieved stronger robustness in some settings. These results support a more balanced interpretation of Trust-FedAvg as a lightweight, competitive, and particularly effective aggregation strategy under directional adversarial manipulation and intermittent connectivity, rather than as a universally dominant robust aggregator.

The main contributions of this work can be summarized as follows:

1. A controlled simulation-driven trust-aware federated learning framework for heterogeneous intelligent IoT environments is proposed, incorporating non-IID client data, adversarial update manipulation, and probabilistic communication dropout.
2. A lightweight dynamic trust-based aggregation mechanism is introduced to reduce the influence of unreliable or malicious edge participants during federated optimization by weighting client updates according to their consistency with the current global model state.
3. A systematic robustness evaluation is conducted under multiple adversarial scenarios, including sign-flip, Gaussian-noise, scaling, and label-flip attacks, together with intermittent communication-dropout probabilities up to 50%.
4. The proposed Trust-FedAvg framework is benchmarked against conventional FedAvg and representative robust aggregation baselines, including Median, Trimmed Mean, Krum, Multi-Krum, and an auxiliary Bulyan configuration, using multi-seed experiments and statistical significance testing.
5. A sensitivity analysis is performed across different Dirichlet non-IID levels to examine whether distance-based trust scoring remains effective when benign client updates naturally deviate from the global model trajectory.
6. Runtime measurements and auxiliary CNN-based experiments are included to provide additional evidence regarding computational overhead and behavior beyond the original UCI HAR MLP setting.

Unlike trust-aware FL studies that primarily evaluate robustness under isolated adversarial conditions, the proposed framework jointly analyzes adversarial manipulation, probabilistic communication dropout, and heterogeneous non-IID edge participation within a unified lightweight IoT-oriented simulation setting. The objective is not to introduce a fundamentally new theoretical FL paradigm, but rather to provide a practical, reproducible, and computationally simple trust-aware aggregation framework whose robustness and limitations are evaluated against stronger baselines and under more demanding edge-learning conditions.

2. Related Work

2.1. Federated Learning for Intelligent IoT Systems

The rapid growth of intelligent Internet of Things (IoT) infrastructures has significantly increased the demand for distributed machine learning solutions capable of operating under privacy constraints, limited communication bandwidth, and resource-constrained edge environments. In this context, federated learning (FL) has emerged as one of the most promising paradigms for enabling collaborative intelligence without requiring the direct sharing of raw local data [3]. By allowing edge devices to perform local model training while exchanging only model parameters or gradients with a central server, FL provides an effective compromise between distributed intelligence and data privacy preservation.

Since the introduction of the Federated Averaging (FedAvg) algorithm by McMahan et al. [3], FL has been widely investigated across multiple IoT-related domains, including smart healthcare, industrial IoT, autonomous transportation, environmental sensing, and edge-enabled cyber-physical systems [1,2]. In particular, IoT systems benefit substantially from FL due to their inherently distributed sensing architectures and the increasing need for local intelligence at the network edge.

Several recent surveys have highlighted the growing importance of FL in IoT ecosystems and discussed challenges related to communication efficiency, scalability, energy consumption, and privacy preservation [4,5]. However, practical deployment in IoT-oriented edge environments remains highly challenging due to device heterogeneity, intermittent connectivity, limited computational capabilities, and highly non-independent and identically distributed (non-IID) local datasets. These characteristics often lead to unstable convergence behavior and degraded global model performance when conventional FL aggregation strategies are employed.

Moreover, edge-IoT systems frequently operate under dynamic participation conditions in which devices may temporarily disconnect, fail to upload updates, or exhibit unstable communication behavior due to wireless network limitations. Despite the practical importance of such conditions, many FL studies still assume ideal communication scenarios with stable client participation and reliable synchronization mechanisms.

2.2. Robust Federated Learning

The distributed nature of FL systems also introduces important security and robustness challenges. Because global model updates depend on the collaborative participation of multiple decentralized clients, malicious or compromised devices may intentionally manipulate local model updates in order to disrupt convergence or poison the global model. Consequently, adversarial robustness has become a major research topic within the federated learning literature.

Several attack strategies have been proposed against FL systems, including Byzantine attacks, model poisoning, backdoor attacks, sign-flip manipulation, scaling attacks, label-flip attacks, and gradient perturbation techniques [12,17,18]. Among these, sign-flip at-

tacks represent a particularly effective and computationally inexpensive strategy in which malicious clients invert the direction of local model updates before aggregation, thereby destabilizing the optimization process. Similarly, Gaussian perturbation attacks introduce large amounts of random noise into model parameters, whereas scaling attacks amplify local updates, disproportionately influencing the global aggregation stage. Label-flip attacks represent a simpler form of data poisoning in which malicious clients intentionally corrupt local training labels, thereby producing updates that may appear less extreme than direct parameter manipulation but can still degrade the global model.

Beyond Byzantine manipulation and sign-flip attacks, prior studies have also demonstrated the vulnerability of federated learning systems to model poisoning and backdoor attacks that can introduce targeted malicious behavior into the global model while remaining difficult to detect under standard aggregation procedures [18,19].

To mitigate such vulnerabilities, numerous robust aggregation methods have been proposed. Examples include Median aggregation, Trimmed Mean, Krum, Multi-Krum, Bulyan, and other Byzantine-resilient optimization strategies designed to reduce the influence of anomalous client updates during federated aggregation [17–20]. These approaches generally rely on robust statistical filtering or distance-based selection mechanisms to identify suspicious updates and improve resilience against adversarial behavior.

In particular, Byzantine-resilient aggregation methods such as Krum and Bulyan were designed to tolerate adversarial model updates by selecting client contributions that are statistically consistent during aggregation. In contrast, Trimmed Mean-based strategies aim to suppress extreme anomalous updates through robust statistical filtering [20–22]. Coordinate-wise Median aggregation follows a related robust-statistical rationale by reducing the influence of extreme parameter values independently across model coordinates. These methods therefore represent important baselines for evaluating whether a proposed trust-aware strategy offers advantages beyond conventional robust aggregation.

Although robust aggregation methods have demonstrated important improvements under adversarial conditions, many existing approaches suffer from practical limitations when applied to IoT-oriented edge environments. Some methods exhibit high computational complexity, require strong assumptions regarding the number of malicious participants, or become unstable under highly heterogeneous non-IID data distributions. Furthermore, many studies evaluate robustness only under isolated attack conditions, without simultaneously accounting for communication unreliability and intermittent client participation, which are commonly observed in real-world IoT networks.

2.3. Trust-Aware Federated Learning

More recently, trust-aware federated learning approaches have emerged as an alternative to improve robustness and reliability in distributed learning environments. Instead of relying solely on hard anomaly filtering, trust-aware methods estimate the reliability or behavioral consistency of participating clients and dynamically adjust aggregation weights based on trust-related metrics.

Existing trust-aware FL approaches employ a variety of reliability indicators, including historical contribution quality, statistical consistency, gradient similarity, reputation scores, participation frequency, and behavioral stability [14,15]. In general, these methods aim to reduce the influence of unreliable or malicious participants while preserving the collaborative and decentralized nature of federated optimization.

For example, several studies have proposed adaptive trust-scoring mechanisms that estimate client reliability based on the similarity between local and global model updates [16]. Other approaches incorporate blockchain-based trust management, incentive mechanisms, or reputation systems designed to improve transparency and accountability in

distributed FL environments [14]. Trust-aware aggregation has shown promising results in terms of adversarial robustness, convergence stability, and resilience against unreliable edge participation.

Nevertheless, important limitations remain in the current trust-aware FL literature. Many proposed methods introduce substantial computational overhead or require complex trust-management infrastructures that may not be suitable for lightweight IoT deployments. Additionally, a significant portion of prior work evaluates trust mechanisms only under simplified experimental conditions, often without considering communication instability, intermittent client connectivity, or heterogeneous edge participation. Another important limitation is that distance- or similarity-based trust mechanisms may unintentionally penalize benign clients under strongly non-IID data distributions, since legitimate local updates can naturally deviate from the current global model trajectory. This issue motivates explicit sensitivity analysis across different non-IID levels.

Table 1 summarizes representative prior studies in federated learning robustness and trust-aware aggregation, highlighting the main characteristics and limitations of existing approaches in IoT-oriented intelligent environments. Note that the classification in this table is qualitative and literature-based. It indicates whether each method has been reported to address a given aspect in prior work, not whether all methods were experimentally benchmarked under the same conditions in this study. The labels “Yes”, “Partial”, “Limited”, and “No” are used as qualitative indicators derived from the primary design objective and reported evaluation scope of each method: “Yes” indicates that the aspect is explicitly addressed, “Partial” indicates that it is considered only indirectly or under limited conditions, “Limited” indicates that the aspect is only weakly supported or not central to the method, and “No” indicates that it is not addressed in the original formulation or typical evaluation setting.

Table 1. Qualitative comparison of representative federated learning, robust aggregation, and trust-aware aggregation approaches with respect to adversarial robustness, communication dropout, non-IID evaluation, IoT orientation, and computational lightweightness. The labels indicate whether each aspect is explicitly addressed, partially considered, weakly supported, or not covered in the original formulation or typical evaluation setting.

Work	Trust-Aware	Adversarial Robustness	Communication Dropout	Non-IID Evaluation	IoT-Oriented	Lightweight
FedAvg [3]	No	Limited	No	Partial	Partial	Yes
FedProx [6]	No	Limited	No	Yes	Partial	Yes
SCAFFOLD [7]	No	Limited	No	Yes	Partial	Yes
Krum/Multi-Krum [20]	No	Yes	No	Limited	No	No
Bulyan [22]	No	Yes	No	Limited	No	No
Trust-based FL [14]	Yes	Partial	No	Partial	Yes	Partial
Reputation-aware FL [15]	Yes	Partial	Limited	Partial	Yes	Partial
Median/Trimmed Mean [21]	No	Yes	No	Limited	Partial	Yes
Proposed Framework	Yes	Yes	Yes	Yes	Yes	Yes

As summarized in Table 1, existing robustness approaches for federated learning frequently address only isolated aspects of IoT-oriented intelligent deployments. Several Byzantine-resilient aggregation methods primarily focus on adversarial robustness, neglecting communication instability and intermittent client participation. Conversely, many trust-aware approaches emphasize reputation management or behavioral consistency without systematically evaluating robustness under heterogeneous, non-IID conditions and probabilistic communication dropout. Furthermore, several robust aggregation strategies introduce substantial computational overhead, potentially limiting scala-

bility in resource-constrained edge-IoT environments. These limitations motivate the development of lightweight trust-aware federated learning frameworks capable of jointly addressing adversarial robustness, communication unreliability, and heterogeneous distributed optimization.

2.4. Research Gap and Positioning of This Work

Despite significant progress in federated learning, robust aggregation, and trust-aware distributed intelligence, several important research gaps remain inadequately addressed in the current literature.

First, many existing FL robustness studies assume ideal communication conditions with stable, synchronous client participation, which does not accurately reflect the realities of IoT deployments, characterized by unreliable wireless communication and intermittent edge connectivity. Second, numerous robust aggregation studies focus exclusively on adversarial robustness while neglecting the combined impact of communication instability and heterogeneous participation dynamics. Third, although trust-aware FL approaches have demonstrated promising robustness, relatively few works integrate lightweight trust-based aggregation with controlled IoT-oriented simulation frameworks that incorporate non-IID data heterogeneity, adversarial manipulation, and probabilistic communication dropout.

In addition, several existing trust-management approaches rely on complex architectures or computationally intensive procedures, which may limit scalability and practical applicability in resource-constrained edge environments. Moreover, comparatively few studies evaluate trust-aware aggregation against multiple robust baselines while also reporting statistical significance, runtime overhead, non-IID sensitivity, and behavior under intermittent client connectivity. Consequently, there remains a need for computationally lightweight, extensible trust-aware FL frameworks that operate effectively in IoT-oriented edge-learning environments.

Motivated by these limitations, the present work proposes a simulation-driven trust-aware federated learning framework specifically designed for robust intelligent IoT networks. The proposed approach combines dynamic trust-aware aggregation with a controlled IoT-oriented simulation environment that incorporates heterogeneous, non-IID data partitioning, adversarial model manipulation, and intermittent communication dropout. Unlike many previous studies, the proposed framework jointly evaluates adversarial robustness and communication instability across multiple seeds, providing a more reproducible and practically relevant basis for analyzing resilient distributed intelligence in edge-IoT systems. In the revised experimental evaluation, the proposed method is also compared with representative robust aggregation baselines, including Median, Trimmed Mean, Krum, Multi-Krum, and an auxiliary Bulyan configuration. In summary, the specific gap addressed in this work is the lack of a unified, lightweight simulation setting in which trust-aware aggregation is evaluated jointly under non-IID client partitions, adversarial update manipulation, probabilistic communication dropout, and multi-seed experimental repetition while also considering baseline robustness, statistical significance, non-IID sensitivity, and round-level runtime overhead.

3. Proposed Framework

3.1. System Architecture

The proposed framework was designed to emulate controlled IoT-oriented federated learning environments characterized by heterogeneous edge devices, non-independent and identically distributed (non-IID) local datasets, intermittent communication conditions, and the possible presence of unreliable or malicious participants. The overall archi-

texture follows a centralized federated learning paradigm in which multiple distributed IoT clients collaboratively train a shared global model under the coordination of a central aggregation server.

The proposed system architecture is composed of four main components:

1. heterogeneous IoT edge clients;
2. a central federated aggregation server;
3. a trust evaluation module;
4. a communication-dropout simulation layer.

Each IoT client stores its own private dataset locally and trains its own model without transmitting raw data to the server. This configuration preserves data locality and aligns with the privacy-preserving principles of federated learning. The clients are assumed to represent heterogeneous edge devices with different local data distributions and varying communication reliability conditions.

During each federated round, a subset of clients is selected to participate in collaborative training. Selected clients receive the current global model from the central server and perform local optimization using their respective local datasets. After local training, model parameters are transmitted back to the server for aggregation. Unlike conventional federated learning approaches, the proposed framework incorporates a trust evaluation mechanism that dynamically estimates the reliability of participating clients based on the consistency of their local model updates with the global model state.

The trust evaluation module operates before global aggregation and assigns adaptive trust scores to client updates. Clients exhibiting highly inconsistent or anomalous updates receive lower aggregation weights, thereby reducing their influence on the global optimization process. This mechanism is particularly relevant under adversarial conditions involving malicious or unreliable participants capable of disrupting federated convergence.

To emulate intermittent IoT communication conditions, the framework also includes a probabilistic communication dropout simulation layer. Under this model, selected clients may fail to upload their local model updates during a federated round due to intermittent connectivity or communication instability. This behavior reflects practical edge-IoT environments in which wireless communication quality, energy limitations, or temporary device unavailability may prevent successful participation in federated aggregation. The dropout layer is intentionally modeled at the federated-learning protocol level and does not attempt to reproduce physical-layer communication, bandwidth allocation, routing, or packet-level network dynamics.

Globally, the proposed architecture combines privacy-preserving distributed learning, trust-aware aggregation, adversarial robustness, and communication instability simulation within a unified, extensible experimental framework suitable for evaluating resilient AI-enabled IoT systems.

Figure 1 illustrates the overall architecture and operational workflow of the proposed simulation-driven trust-aware federated learning framework for heterogeneous intelligent IoT environments.

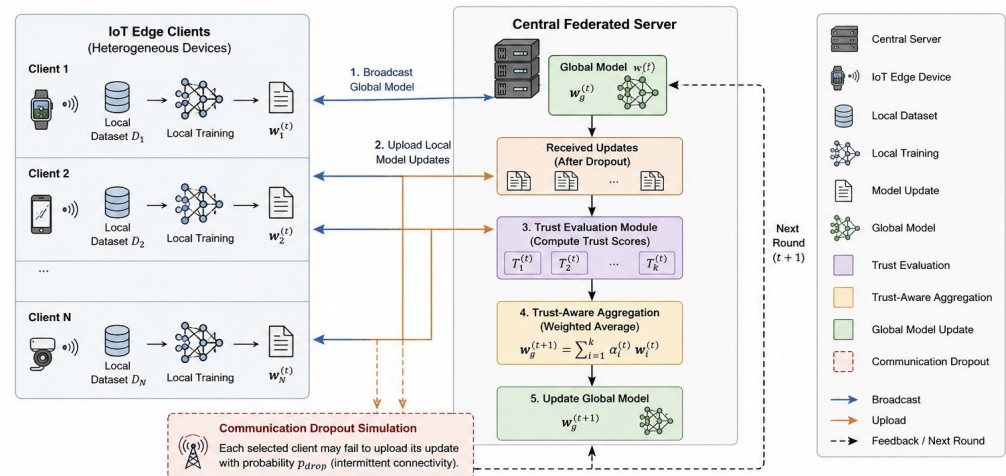


Figure 1. Architecture and operational workflow of the proposed Trust-FedAvg framework for IoT-oriented federated learning. Heterogeneous edge clients perform local training using private datasets, after which communication dropout is simulated before the server receives available model updates. The server then evaluates update consistency, assigns trust scores, performs trust-aware weighted aggregation, and updates the global model for the next federated round. The orange dotted lines indicate that client uploads are subject to the communication dropout simulation, whereby each selected client may fail to upload its local model update with probability p_{drop} . The ellipsis (...) denotes omitted intermediate clients, received updates, or trust scores for visual clarity.

3.2. Federated Learning Workflow

The proposed federated learning framework follows an iterative server-client collaborative optimization procedure. Let N denote the total number of available IoT clients participating in the distributed learning process. During each federated communication round t , the central server selects a subset of clients to participate in local training and model aggregation.

Initially, the server broadcasts the current global model parameters, $w_g^{(t)}$, to the selected clients. Each participating client then performs local training using its own private local dataset, D_i , producing an updated local model $w_i^{(t)}$. Local optimization is performed independently on each device using stochastic gradient-based optimization without sharing raw training data with other clients or the central server.

After local training, clients attempt to transmit their updated model parameters back to the server. However, unlike idealized federated learning settings, the proposed framework incorporates probabilistic communication dropout to emulate intermittent edge-network connectivity. Consequently, some clients may fail to upload their updates during a communication round, resulting in dynamic and potentially reduced participation.

The overall federated workflow can therefore be summarized as follows:

1. client selection by the central server;
2. broadcast of the global model to selected clients;
3. local training using private edge datasets;
4. communication-dropout evaluation;
5. trust estimation of received client updates;
6. trust-aware global aggregation;
7. update of the global model parameters.

Unlike conventional FedAvg aggregation, the proposed framework introduces an additional trust-evaluation stage prior to aggregation. The trust mechanism estimates the consistency of each client update relative to the current global model and adaptively adjusts aggregation weights accordingly. Clients producing highly inconsistent updates re-

ceive lower trust scores and therefore exert reduced influence on the global optimization process.

Algorithm 1 summarizes the complete Trust-FedAvg workflow implemented in the proposed framework. It explicitly shows the ordering of client selection, local training, adversarial manipulation, communication-dropout simulation, trust-score computation, and trust-aware aggregation.

Algorithm 1. Trust-FedAvg workflow under adversarial and communication-dropout conditions.

Input: initial global model $w_g^{(0)}$, client set $\mathcal{C}$, number of rounds T , participation fraction q , local datasets D_i , dropout probability p_{drop} , minimum trust threshold $T_{\min}$, and numerical constant ϵ .

For each communication round $t = 0, 1, \dots, T - 1$:

1. The server samples a subset $\mathcal{S}^{(t)} \subseteq \mathcal{C}$ according to the participation fraction q .
2. The server broadcasts the current global model $w_g^{(t)}$ to all selected clients $i \in \mathcal{S}^{(t)}$.
3. Each selected client trains the received model locally on its private dataset D_i , producing $w_i^{(t)}$.
4. If client i is malicious, the corresponding adversarial transformation is applied to $w_i^{(t)}$.
5. Each selected client attempts to upload its local model. The upload succeeds with probability $1 - p_{\text{drop}}$.
6. For each successfully received update, the server computes the distance $d_i^{(t)} = \|w_i^{(t)} - w_g^{(t)}\|_2$.
7. The server computes the trust score $T_i^{(t)} = \max(T_{\min}, 1/(d_i^{(t)} + \epsilon))$.
8. The aggregation coefficient is computed as $\alpha_i^{(t)} = n_i T_i^{(t)} / \sum_{k \in \mathcal{R}^{(t)}} n_k T_k^{(t)}$, where $\mathcal{R}^{(t)}$ is the set of clients whose updates were successfully received.
9. The global model is updated as $w_g^{(t+1)} = \sum_{i \in \mathcal{R}^{(t)}} \alpha_i^{(t)} w_i^{(t)}$. If no update is received, the global model remains unchanged for that round.

Output: final global model $w_g^{(T)}$.

This workflow enables the framework to simultaneously address several IoT-oriented challenges associated with intelligent federated learning systems, including:

1. heterogeneous local data distributions;
2. unreliable communication conditions;
3. intermittent client participation;
4. malicious or adversarial client behavior;
5. distributed privacy-preserving optimization.

Furthermore, the proposed workflow remains computationally lightweight and scalable, making it suitable for simulation-driven evaluation of edge-IoT environments involving resource-constrained devices and dynamic network conditions.

3.3. Trust-Aware Aggregation Mechanism

Conventional federated learning aggregation strategies, such as Federated Averaging (FedAvg), assume that all participating clients contribute reliable, statistically consistent local model updates during collaborative optimization. However, this assumption is often unrealistic in intelligent IoT environments characterized by heterogeneous edge devices, unreliable communication, noisy sensing data, and the potential presence of malicious participants. Under such conditions, anomalous or adversarial local updates may disproportionately influence global model aggregation, leading to degraded convergence stability and reduced predictive performance.

To address these limitations, the proposed framework incorporates a lightweight, trust-aware aggregation mechanism that dynamically adjusts each client's contribution based on the consistency of its local model update with the current global model state. The underlying intuition is that benign clients participating in collaborative optimization tend to produce updates that remain reasonably consistent with the global optimization trajectory, whereas malicious or unreliable participants are more likely to generate highly divergent updates. This assumption is not universal under strongly non-IID conditions, and this limitation is explicitly evaluated through the non-IID sensitivity analysis reported in the experimental section.

Let $w_g^{(t)}$ denote the global model parameters at federated communication round t , and let $w_i^{(t)}$ represent the local model parameters produced by client i after local training. The discrepancy between the local and global models is quantified using the Euclidean distance:

$$d_i^{(t)} = \|w_i^{(t)} - w_g^{(t)}\|_2$$

where $d_i^{(t)}$ represents the deviation of client i relative to the current global model state. Clients producing highly inconsistent updates will therefore exhibit larger distances.

Based on this deviation measure, a trust score is assigned to each participating client according to an inverse-distance formulation:

$$T_i^{(t)} = \frac{1}{d_i^{(t)} + \epsilon}$$

where $T_i^{(t)}$ denotes the trust score associated with client i at communication round t , and ϵ is a small positive constant introduced to avoid numerical instability when the distance approaches zero.

Under this formulation, clients that generate updates closer to the global optimization trajectory receive higher trust scores, whereas highly divergent updates receive lower trust scores. To prevent the complete exclusion of clients and maintain aggregation stability under highly heterogeneous non-IID conditions, a minimum trust threshold is enforced during aggregation.

$$T_i^{(t)} \leftarrow \max(T_{\min}, T_i^{(t)})$$

The trust scores are subsequently incorporated into the aggregation process through adaptive trust-aware weighting coefficients defined as:

$$\alpha_i^{(t)} = \frac{n_i T_i^{(t)}}{\sum_{k \in \mathcal{R}^{(t)}} n_k T_k^{(t)}}$$

where n_i denotes the number of local training samples associated with client i , and $\mathcal{R}^{(t)}$ denotes the set of clients whose updates were successfully received during communication round t .

Finally, the global model is updated using the trust-aware weighted aggregation rule:

$$w_g^{(t+1)} = \sum_{i \in \mathcal{R}^{(t)}} \alpha_i^{(t)} w_i^{(t)}$$

If $\mathcal{R}^{(t)}$ is empty because all selected clients experience communication dropout, aggregation is skipped and the global model is carried forward unchanged, i.e., $w_g^{(t+1)} = w_g^{(t)}$.

In the experiments, ϵ was set to 1×10^{-8} to avoid numerical instability. The minimum trust score was set to 0.05. Trust values were computed from the Euclidean distance between each received local model and the current global model, considering all floating-

point trainable parameters as a single equivalent parameter vector. No layer-wise normalization was applied; therefore, all parameters contributed to the Euclidean distance according to their raw numerical scale. After applying the minimum trust threshold, aggregation weights were obtained by multiplying each client's trust score by its local dataset size and renormalizing the resulting weights across the available clients.

Unlike conventional FedAvg aggregation, which weights clients solely by dataset size, the proposed Trust-FedAvg mechanism jointly considers both local data contributions and behavioral consistency relative to the global optimization process. Consequently, the influence of unreliable or malicious clients is substantially reduced during aggregation.

An important advantage of the proposed approach is its relatively low computational complexity compared with more sophisticated Byzantine-resilient aggregation methods. The trust computation requires only distance evaluation between local and global model parameters, making the framework computationally lightweight and suitable for resource-constrained edge-IoT deployments. In contrast, robust aggregation methods such as Krum, Multi-Krum, and Bulyan generally require pairwise distance calculations among client updates or multi-stage selection procedures, which can increase server-side aggregation cost as the number of participating clients grows.

Furthermore, the proposed mechanism does not require prior knowledge of the number or identity of malicious participants, thereby improving its practical applicability in dynamic intelligent IoT environments where adversarial behavior may vary over time. The trust-aware formulation also remains compatible with intermittent client participation and probabilistic communication dropout, making it particularly suitable for controlled IoT-oriented FL simulations with unstable connectivity and heterogeneous edge participation.

A potential limitation of distance-based trust scoring is that benign clients with highly non-IID local data may produce updates that deviate substantially from the global model trajectory. In such cases, the trust mechanism may partially down-weight statistically valid but distributionally distinct client contributions. This limitation motivates the sensitivity analysis on non-IID settings and should be considered when interpreting the robustness results.

3.4. Adversarial Attack Model

To evaluate the robustness of the proposed trust-aware federated learning framework under adversarial edge-learning conditions, several attack strategies were incorporated into the simulation environment. These attacks emulate malicious or unreliable IoT participants that can intentionally manipulate local model updates before transmission to the central aggregation server. The selected attacks were chosen for their relevance in the federated learning literature and their ability to represent different forms of adversarial behavior commonly observed in distributed edge-learning environments.

In the proposed framework, a fraction of participating clients is designated as malicious during federated training. Malicious clients follow the standard local training procedure but intentionally alter their local model updates before transmitting them to the aggregation server. Let $M \subseteq N$ denote the subset of malicious clients participating in a given communication round.

Four adversarial attack strategies were considered in this work:

1. sign-flip attack;
2. Gaussian noise attack;
3. scaling attack;
4. label-flip attack.

For each experimental seed, the malicious client set was fixed across all communication rounds. This design choice was adopted to emulate persistent device compromise, in which the same compromised IoT devices remain adversarial throughout the federated training process. However, malicious clients only affected a given round when they were selected for participation and successfully uploaded their manipulated updates.

3.4.1. Sign-Flip Attack

The sign-flip attack is among the most widely studied adversarial strategies in federated learning due to its simplicity and effectiveness. In this attack, malicious clients invert the direction of local model updates before sending them to the server. More specifically, if $w_i^{(t)}$ denotes the local model parameters generated by client i , the manipulated update is defined as:

$$\tilde{w}_i^{(t)} = -w_i^{(t)}$$

This manipulation causes malicious clients to push the global optimization process in the opposite direction of legitimate gradient descent updates, potentially destabilizing convergence and degrading predictive performance.

The sign-flip attack is particularly relevant in IoT federated learning environments because it can be implemented with minimal computational overhead and without requiring access to other clients' data or model updates. Consequently, it represents a practical threat model for compromised edge devices operating in distributed intelligent IoT systems.

3.4.2. Gaussian Noise Attack

In the Gaussian noise attack, malicious clients inject large, random perturbations into the local model parameters before transmitting them to the aggregation server. The manipulated model update is defined as:

$$\tilde{w}_i^{(t)} = w_i^{(t)} + \mathcal{N}(0, \sigma^2)$$

where $\mathcal{N}(0, \sigma^2)$ denotes additive Gaussian noise with zero mean and variance σ^2 .

This attack emulates unreliable or corrupted IoT edge devices producing unstable or highly noisy model updates due to hardware faults, sensing instability, communication corruption, or malicious perturbation. Unlike sign-flip attacks, Gaussian perturbations do not necessarily reverse the optimization direction; instead, they introduce stochastic instability into the aggregation process.

3.4.3. Scaling Attack

The scaling attack amplifies local model updates by multiplying model parameters by a predefined scaling factor γ :

$$\tilde{w}_i^{(t)} = \gamma w_i^{(t)}$$

where $\gamma > 1$ controls the attack intensity.

The objective of this attack is to disproportionately increase the influence of malicious participants during aggregation by artificially increasing the magnitudes of updates. Conventional FedAvg aggregation strategies may be particularly vulnerable to this type of manipulation, as large-magnitude updates can dominate the weighted averaging process.

Scaling attacks are especially relevant in heterogeneous IoT environments where naturally occurring variability in device behavior may partially mask adversarial manipulation, making anomaly detection more difficult.

3.4.4. Label-Flip Attack

The label-flip attack was included as an additional data-poisoning scenario. In this case, malicious clients do not directly manipulate model parameters after training. Instead, they corrupt their local training labels before local optimization, thereby causing the resulting local model update to be learned from intentionally incorrect supervision. Let y denote the original class label and K the number of classes. In the implemented label-flip setting, the malicious label transformation is defined as:

$$\tilde{y} = (y + 1) \bmod K$$

where $\tilde{y}$ is the poisoned label used locally by malicious clients during training. This attack is less extreme than direct sign-flip or scaling manipulation because the transmitted model parameters are still produced by a valid local training process. However, it represents an important form of data poisoning in which compromised clients attempt to bias the global model through corrupted local supervision.

3.4.5. Adversarial Configuration

To ensure reproducible evaluation conditions, adversarial participation was simulated probabilistically by assigning a predefined fraction of clients to be malicious in each experiment. In the performed experiments, 20% of clients were configured as adversarial participants, representing partially compromised IoT environments.

The attack mechanisms were applied after local training and before communication with the aggregation server, thereby emulating adversarial manipulation at the edge device level. For the label-flip attack, the malicious transformation was instead applied to local labels before client-side training. Importantly, the central server was assumed to have no prior knowledge of the identities or numbers of malicious participants, reflecting practical distributed IoT deployment conditions.

Globally, the selected attack models provide complementary adversarial scenarios capable of evaluating the robustness of federated aggregation under:

1. directional manipulation;
2. stochastic perturbation;
3. aggregation dominance attacks;
4. local data-poisoning behavior.

This diversity of adversarial conditions enables a more comprehensive evaluation of the proposed trust-aware federated learning framework under IoT-oriented adversarial simulation conditions.

3.5. Communication Dropout Simulation

In practical intelligent Internet of Things (IoT) environments, communication reliability cannot generally be assumed to be stable or deterministic. Edge devices frequently operate under constrained wireless conditions characterized by intermittent connectivity, variable latency, packet loss, limited transmission power, and temporary device unavailability. Consequently, participating clients may fail to upload local model updates during federated learning communication rounds, resulting in dynamic and irregular participation patterns.

To emulate intermittent edge-network conditions at the FL protocol level, the proposed framework incorporates a probabilistic communication dropout mechanism. The objective of this module is to evaluate the robustness and convergence stability of federated learning algorithms under intermittent client participation, a condition commonly observed in practical IoT deployments.

Let p_{drop} denote the communication dropout probability for each selected client during a federated communication round. After local training is completed, each participating client attempts to upload its local model update to the central aggregation server. However, the upload succeeds only with probability:

$$P(\text{successful upload}) = 1 - p_{\text{drop}}$$

Consequently, the probability of communication failure is given by:

$$P(\text{dropout}) = p_{\text{drop}}$$

If a communication dropout occurs, the corresponding client update is discarded and excluded from the aggregation stage for that communication round. Therefore, the number of effective participating clients may vary over time, depending on network conditions and probabilistic communication failures.

This mechanism emulates several high-level IoT communication phenomena, including:

- intermittent wireless connectivity;
- unstable edge-network participation;
- packet transmission failures;
- temporary edge-device unavailability;
- energy-related communication interruptions.

Unlike conventional federated learning studies, which frequently assume synchronous, fully reliable communication, the proposed simulation environment explicitly models dynamic participation variability during collaborative optimization. This characteristic is particularly relevant for intelligent IoT systems operating in large-scale distributed environments where communication reliability may fluctuate significantly over time.

An important consequence of communication dropout is the reduction in the effective aggregation population during federated rounds. In extreme cases, all selected clients may fail to upload updates during a communication round, causing the aggregation process to be temporarily skipped. The proposed framework explicitly models these conditions to evaluate the resilience of federated optimization under severe communication instability.

To assess robustness under varying communication conditions, multiple dropout probabilities were evaluated experimentally, including:

$$p_{\text{drop}} \in \{0.0, 0.1, 0.3, 0.5\}$$

These configurations represent progressively more challenging edge-network environments ranging from ideal communication conditions to highly unstable IoT connectivity scenarios.

The integration of communication-dropout simulation with adversarial client behavior and trust-aware aggregation enables the proposed framework to reproduce more practically relevant IoT-oriented FL simulation conditions than many conventional FL evaluation settings. In particular, the combined analysis of adversarial robustness and intermittent communication provides important insights regarding the practical resilience of distributed AI systems operating under heterogeneous edge constraints.

Furthermore, the probabilistic dropout model remains computationally lightweight and scalable, making it suitable for large-scale simulation-driven evaluation of federated learning algorithms in controlled IoT-oriented environments.

4. Experimental Setup

4.1. Datasets

To evaluate the effectiveness and robustness of the proposed trust-aware federated learning framework, experiments were conducted using the publicly available UCI Human Activity Recognition (UCI HAR) dataset. This dataset was selected due to its widespread adoption in edge intelligence, federated learning, and IoT-oriented machine learning research, particularly for human activity recognition in distributed sensing environments and resource-constrained edge devices. Although the UCI HAR dataset is a well-established benchmark for sensor-based activity recognition, it represents only a single IoT-related application domain and relies on pre-engineered features rather than raw sensor streams. Consequently, the obtained results should be interpreted as evidence of robustness within a controlled sensor-learning benchmark, rather than as definitive proof of general applicability across all intelligent IoT scenarios and edge-learning tasks.

The UCI HAR dataset contains inertial sensor measurements acquired from smartphone-based accelerometers and gyroscopes during the execution of multiple human activities. The dataset includes recordings collected from 30 subjects performing six daily activities:

1. walking;
2. walking upstairs;
3. walking downstairs;
4. sitting;
5. standing;
6. laying.

The raw sensor signals were preprocessed using fixed-width sliding windows and transformed into a feature representation composed of 561 engineered features extracted from time-domain and frequency-domain signals. The resulting dataset provides a widely used sensor-based benchmark for evaluating distributed learning under heterogeneous sensing conditions.

The dataset contains:

1. 7352 training samples;
2. 2947 testing samples;
3. 561 input features;
4. 6 activity classes.

In the proposed federated learning environment, the centralized training dataset was partitioned among multiple simulated IoT clients using a non-independent and identically distributed (non-IID) allocation strategy based on Dirichlet sampling. This approach emulates edge-learning conditions in which participating devices observe heterogeneous local data distributions rather than balanced global datasets.

More specifically, a Dirichlet distribution with concentration parameter $\alpha = 0.5$ was employed to generate heterogeneous local partitions across clients in the main experimental campaign. Lower values of (α) lead to greater statistical heterogeneity and class imbalance among participating edge devices, thereby increasing the difficulty of federated optimization. To explicitly assess the sensitivity of the proposed distance-based trust mechanism to client heterogeneity, additional experiments were conducted with $\alpha \in \{0.1, 0.3, 0.5, 1.0\}$.

The use of non-IID client partitions is particularly important in IoT federated learning because real-world edge devices typically collect environment- or user-specific data that exhibit substantial distributional variability.

In addition to the UCI HAR experiments, an auxiliary CIFAR-10 experiment was conducted using a compact convolutional neural network (CNN). This auxiliary experiment

was not intended to replace the IoT-oriented UCI HAR evaluation or to claim direct IoT deployment validity. Instead, it was included to address the limitation of relying exclusively on a single MLP-based sensor benchmark and to examine whether the aggregation behavior remains consistent in a higher-dimensional CNN setting.

Table 2 summarizes the main characteristics of the datasets employed in the experimental evaluation.

Table 2. Main characteristics of the datasets employed in the federated learning experiments.

Dataset	Samples (Train/Test)	Input Representation	Classes	Application Domain	Role in This Study
UCI HAR	7352/2947	561 engineered sensor features	6	Human activity recognition	Main IoT-oriented sensor benchmark
CIFAR-10	50,000/10,000	32 × 32 RGB images	10	Image classification	Auxiliary CNN validation

4.2. Federated Learning Configuration

The federated learning experiments were conducted using a centralized server-client architecture composed of multiple simulated IoT edge devices collaboratively training a shared global model. The overall experimental configuration was designed to emulate controlled IoT-oriented environments characterized by heterogeneous client participation and intermittent communication.

A total of 10 federated clients were included in the main UCI HAR training configuration. During each communication round, 50% of clients were randomly selected to participate in local training and aggregation. This configuration reflects practical IoT scenarios in which only a subset of edge devices may be available or selected during a given communication cycle.

The MLP consisted of an input layer with 561 features, followed by two hidden layers with 256 and 128 neurons, respectively. Each hidden layer used a ReLU activation function followed by dropout regularization with a dropout rate of 0.2. The output layer contained six neurons corresponding to the activity classes. The model contained 177,542 trainable parameters. Training used cross-entropy loss, Adam with a learning rate of 0.001, batch size 32, and one local epoch per communication round.

The federated training process was executed for 30 communication rounds. Multi-seed evaluation was employed to improve statistical robustness and reduce sensitivity to random initialization and client selection variability. More specifically, experiments were repeated using five independent random seeds:

1. 42;
2. 123;
3. 456;
4. 789;
5. 999.

For the auxiliary CIFAR-10 experiment, a compact CNN was used, consisting of two convolutional layers with 32 and 64 filters, max-pooling operations, a fully connected hidden layer with 128 neurons, dropout regularization with a rate of 0.3, and a 10-class output layer. The model contained 545,098 trainable parameters. The CIFAR-10 auxiliary experiments used 10 clients, 50% client participation, $\alpha = 0.5$, batch size 64, 20 communication rounds, and three random seeds: 42, 123, and 456.

The main federated learning configuration parameters are summarized in Table 3.

Table 3. Federated learning configuration parameters used throughout the experimental evaluation.

Parameter	Value
Number of clients	10
Client participation fraction	0.5
Communication rounds	30
Local epochs	1
Batch size	32
Optimizer	Adam
Learning rate	0.001
Dirichlet parameter (α)	0.5
Number of random seeds	5

Table 4 summarizes the additional experimental configurations used for the auxiliary CNN and Bulyan evaluations.

Table 4. Auxiliary experimental configurations used for extended validation.

Experiment	Dataset	Model	Clients	Client Participation	Rounds	Seeds	Purpose
CNN auxiliary validation	CIFAR-10	Compact CNN	10	0.5	20	42, 123, 456	Assess behavior beyond the UCI HAR MLP setting
Bulyan auxiliary comparison	UCI HAR	MLP	20	1.0	30	42, 123, 456, 789, 999	Evaluate Bulyan under a valid full-participation setting

The Bulyan experiment was performed separately because Bulyan requires a sufficient number of participating clients relative to the assumed number of Byzantine clients. The main UCI HAR configuration uses 10 clients with 50% participation, resulting in approximately five effective clients per round before dropout, which is not an appropriate setting for a stable Bulyan evaluation. Therefore, Bulyan was evaluated in an auxiliary 20-client full-participation configuration under clean and sign-flip conditions.

4.3. Aggregation Methods and Robust Baselines

To address the need for stronger comparative evaluation, the proposed Trust-FedAvg method was compared against conventional FedAvg and several representative robust aggregation baselines. The evaluated aggregation methods were:

1. FedAvg;
2. Trust-FedAvg;
3. coordinate-wise Median;
4. Trimmed Mean;
5. Krum;
6. Multi-Krum;
7. Bulyan, in an auxiliary configuration.

FedAvg was used as the conventional baseline aggregation method. Median and Trimmed Mean were included as robust statistical aggregation strategies that reduce the influence of extreme coordinate-wise update values. Krum and Multi-Krum were included as distance-based Byzantine-resilient aggregation methods that select client updates according to pairwise consistency. Bulyan was included as an auxiliary stronger Byzantine-resilient baseline, evaluated separately under a 20-client full-participation setting.

For Trimmed Mean, the trimming ratio was set to 0.2. For Krum, Multi-Krum, and Bulyan, the assumed number of Byzantine clients was derived from the malicious-client ratio used in the corresponding experiment. For Trust-FedAvg, the trust-score numerical constant was set to $\epsilon = 1 \times 10^{-8}$, and the minimum trust score was set to 0.05.

The main SOTA comparison on UCI HAR evaluated FedAvg, Trust-FedAvg, Median, Trimmed Mean, Krum, and Multi-Krum under clean, sign-flip, Gaussian-noise, scaling, and label-flip conditions. The communication-dropout campaign focused on FedAvg, Trust-FedAvg, Median, and Trimmed Mean, since Krum-type methods may become unstable or theoretically inappropriate when the number of successfully received client updates is too small under high dropout.

4.4. Adversarial Configuration

To evaluate robustness against malicious or unreliable participants, adversarial behavior was introduced into the federated learning environment through multiple attack strategies applied to local client updates before aggregation.

Four attack mechanisms were considered:

1. sign-flip attack;
2. Gaussian noise attack;
3. scaling attack;
4. label-flip attack.

In all adversarial experiments, 20% of participating clients were randomly designated as malicious edge devices. Malicious clients followed the standard local training process but intentionally manipulated their local model updates before transmitting them to the aggregation server. For the label-flip attack, the malicious behavior was introduced before local training by cyclically modifying local class labels.

For Gaussian perturbation attacks, additive noise with standard deviation

$$\sigma = 5$$

was applied to local model parameters.

For scaling attacks, manipulated updates were amplified using a scaling factor

$$\gamma = 5$$

The attack parameters were selected to generate sufficiently disruptive adversarial conditions while maintaining optimization dynamics suitable for comparative robustness analysis.

Table 5 summarizes the adversarial configurations employed in the experimental evaluation.

Table 5. Adversarial attack configurations considered in the robustness evaluation experiments.

Attack Type	Description	Parameters
Sign-flip	Inversion of model update direction	$w_i \rightarrow -w_i$
Gaussian noise	Additive random perturbation	$\sigma = 5$
Scaling	Amplification of local updates	$\gamma = 5$
Label-flip	Cyclic corruption of local training labels on malicious clients	$\tilde{y} = (y + 1) \bmod K$

4.5. Communication Dropout Configuration

To emulate intermittent edge-network connectivity conditions, probabilistic communication dropout was incorporated into the federated learning process. During each communication round, some selected clients may fail to upload local model updates due to a predefined dropout probability.

The following communication-dropout probabilities were evaluated:

$$p_{\text{drop}} \in \{0.0, 0.1, 0.3, 0.5\}$$

These values represent progressively more unstable communication conditions ranging from ideal fully connected environments to highly unreliable edge-IoT networks with substantial communication failure rates.

Communication dropout was evaluated under both clean federated learning conditions and adversarial attack scenarios to assess the combined impact of malicious behavior and intermittent connectivity on the robustness of federated optimization. In the dropout-focused campaign, the sign-flip attack was used as the primary adversarial condition because it produced substantial degradation in conventional FedAvg while remaining computationally simple and representative of directional model-update manipulation.

4.6. Evaluation Metrics and Statistical Analysis

The proposed framework was evaluated using several performance and robustness metrics designed to characterize convergence quality, predictive performance, and resilience under heterogeneous IoT conditions.

The primary evaluation metric was classification accuracy computed on the global test dataset after each federated communication round. In addition, test loss values were monitored to assess optimization stability and convergence behavior. To provide a more complete classification assessment, macro-averaged F1-score, macro-averaged precision, and macro-averaged recall were also computed. Macro-averaged metrics were selected because they assign equal importance to all classes, which is particularly relevant under non-IID client partitions and potential class imbalance across local datasets.

To improve statistical reliability, results were aggregated across multiple random seeds, and both the mean and standard deviation were reported. The following metrics were analyzed throughout the experiments:

1. mean test accuracy;
2. best test accuracy;
3. final test accuracy;
4. mean test loss;
5. final macro-F1;
6. final macro-precision;
7. final macro-recall;
8. convergence stability across seeds.

For communication-dropout experiments, additional metrics related to effective client participation were also evaluated, including:

1. average number of effective participating clients;
2. number of skipped aggregation rounds.

Runtime profiling was also performed by measuring the wall-clock time of each communication round and the total runtime of each seed-level experiment. These measurements were used to quantify the server-side and training-loop overhead introduced by trust-aware aggregation and robust aggregation baselines under the same hardware and software environment.

Statistical comparisons were performed using paired tests across common random seeds. Trust-FedAvg was treated as the reference method, and its final accuracy and macro-F1 were compared with the corresponding values obtained by each baseline under the same dataset, attack type, non-IID parameter, dropout probability, and seed. Both paired (*t*)-tests and Wilcoxon signed-rank tests were computed. Given the limited number

of seeds, the statistical tests were interpreted as supporting evidence rather than as definitive distributional proof.

Globally, the selected evaluation methodology enables comprehensive analysis of:

1. adversarial robustness;
2. convergence stability;
3. communication resilience;
4. heterogeneous participation dynamics;
5. trust-aware aggregation effectiveness;
6. statistical significance;
7. runtime overhead.

In addition to accuracy and loss, we report the relative accuracy degradation with respect to the clean setting and the absolute robustness gain over FedAvg under the same adversarial and dropout configuration. These metrics allow a clearer comparison of robustness across attack types and communication conditions.

4.7. Implementation Details

All experiments were implemented in Python 3.10.0 using the PyTorch 2.8.0 (CPU build) deep learning framework. The federated learning simulation environment was developed using custom Python-based scripts executed under Windows 11 Pro on a standard consumer-grade laptop equipped with an Intel Core i9-185H processor, 32 GB RAM, and Intel Arc integrated graphics. All experiments were performed on the CPU, without GPU acceleration.

The experimental framework incorporated NumPy 2.2.6, Pandas 2.2.3, Matplotlib 3.10.0, and Scikit-learn libraries 1.6.1 for data processing, statistical analysis, and visualization. To improve reproducibility and reduce stochastic variability, all experimental scenarios were evaluated using multiple independent random seeds. The complete simulation pipeline included automated multi-seed execution, adversarial attack injection, communication-dropout simulation, robust aggregation baselines, trust-aware aggregation, runtime profiling, and statistical result aggregation.

To support reproducibility, the source code, configuration files, random seeds, and aggregated experimental logs are publicly available at <https://github.com/mcabralreis/trust-fedavg-iot-applsci> (accessed on 7 June 2026) and archived on Zenodo at <https://doi.org/10.5281/zenodo.21131862> (accessed on 7 June 2026). The repository includes the scripts and aggregated results required to reproduce the main UCI HAR experiments, the non-IID sensitivity analysis, the communication-dropout robustness experiments, the auxiliary Bulyan comparison, and the auxiliary CIFAR-10 CNN validation. The implementation used Python 3.10 and the PyTorch deep learning framework, together with NumPy, Pandas, Scikit-learn, Matplotlib, PyYAML, tqdm, and torchvision for data processing, model training, statistical analysis, and visualization. The complete software environment and dependency list are provided in the released repository. The repository also includes the scripts used to reproduce the reviewer-revision campaigns, including the UCI HAR SOTA comparison, non-IID sensitivity analysis, dropout robustness analysis, auxiliary CIFAR-10 CNN experiment, and auxiliary Bulyan comparison.

5. Results and Discussion

5.1. Baseline Performance and Robust Aggregation Benchmarking

The first set of experiments evaluated the behavior of the proposed Trust-FedAvg framework against conventional FedAvg and representative robust aggregation baselines under the main UCI HAR configuration. The objective was to establish not only whether

Trust-FedAvg improves over standard FedAvg, but also whether it remains competitive with established Byzantine-resilient and robust-statistical aggregation strategies. The evaluated methods included FedAvg, Trust-FedAvg, coordinate-wise Median, Trimmed Mean, Krum, and Multi-Krum. All results in this subsection correspond to the UCI HAR dataset, $\alpha = 0.5$, 10 clients, 50% client participation, 30 communication rounds, and five independent random seeds.

Under clean training conditions, all lightweight averaging-based methods achieved high final accuracy. FedAvg reached 0.896 ± 0.035 , Trimmed Mean reached 0.897 ± 0.036 , and Trust-FedAvg achieved the highest clean performance, 0.906 ± 0.015 . Median aggregation achieved 0.885 ± 0.050 , whereas Krum and Multi-Krum produced lower clean accuracies, 0.659 ± 0.082 and 0.845 ± 0.047 , respectively. These results indicate that the proposed trust-aware weighting does not degrade clean convergence under the main non-IID UCI HAR configuration.

The adversarial results show a more nuanced behavior. Under sign-flip attacks, conventional FedAvg degraded severely, reaching only 0.447 ± 0.149 final accuracy and 0.345 ± 0.148 macro-F1. In contrast, Trust-FedAvg achieved 0.862 ± 0.032 final accuracy and 0.854 ± 0.039 macro-F1, outperforming Median, Trimmed Mean, Krum, and Multi-Krum in this setting. The improvement over FedAvg was 41.4 percentage points in final accuracy. This confirms that the proposed distance-based trust mechanism is particularly effective against directional model-update manipulation.

Under label-flip attacks, Trust-FedAvg also achieved the strongest performance, reaching 0.888 ± 0.030 final accuracy and 0.884 ± 0.037 macro-F1. This result suggests that trust-aware aggregation can also mitigate a milder data-poisoning scenario in which malicious clients corrupt local supervision rather than directly manipulating transmitted model parameters.

However, the results also show that Trust-FedAvg is not universally superior across all perturbation types. Under Gaussian-noise attacks, coordinate-wise Median achieved the strongest robustness, with 0.868 ± 0.072 final accuracy and 0.858 ± 0.088 macro-F1. Multi-Krum also outperformed Trust-FedAvg in this scenario. Trust-FedAvg still substantially improved over FedAvg and Trimmed Mean, but its final accuracy, 0.695 ± 0.094 , remained below Median and Multi-Krum. This indicates that severe high-magnitude stochastic perturbations can be better handled by coordinate-wise robust statistics or client-client consistency filtering than by simple client-global distance-based trust weighting.

Under scaling attacks, Median and Trust-FedAvg achieved very similar final performance. Median obtained 0.907 ± 0.021 accuracy, whereas Trust-FedAvg obtained 0.902 ± 0.021 . The difference between these two methods was not statistically significant. These results support a balanced interpretation: Trust-FedAvg is highly effective against sign-flip and label-flip attacks, competitive under scaling attacks, but less robust than Median and Multi-Krum under severe Gaussian-noise perturbations.

Statistical testing further supports these observations. Under sign-flip attacks, Trust-FedAvg significantly outperformed FedAvg in final accuracy according to the paired t -test ($p = 0.0042$) and also outperformed Trimmed Mean ($p = 0.0003$) and Krum ($p = 0.0498$). The difference between Trust-FedAvg and Median under sign-flip was not statistically significant ($p = 0.4344$), although Trust-FedAvg achieved higher mean accuracy. Under Gaussian-noise attacks, Trust-FedAvg significantly improved over FedAvg ($p = 0.0002$) but was significantly worse than Median ($p = 0.0013$) and Multi-Krum ($p = 0.0413$). Under scaling attacks, Trust-FedAvg significantly improved over FedAvg ($p = 0.0040$), while the difference relative to Median was not significant ($p = 0.5206$).

Globally, the main robust aggregation benchmark indicates that the proposed method should not be interpreted as a universally dominant robust aggregator. Instead,

Trust-FedAvg provides a lightweight and competitive aggregation strategy that is particularly effective against directional and label-based manipulation, while retaining clear limitations under severe stochastic perturbations. This interpretation is consistent with the expected behavior of distance-based trust scoring: updates that move strongly against the global optimization trajectory are heavily down-weighted, whereas high-dimensional Gaussian perturbations may be better suppressed by coordinate-wise robust statistics.

5.2. Convergence Behavior Under Sign-Flip Attacks

To illustrate the convergence behavior of the proposed method, Figure 2 shows a single-seed comparison between clean FedAvg, FedAvg under 20% sign-flip attack, and Trust-FedAvg under the same attack condition. This figure is intended as an illustrative example of round-level behavior rather than as the primary statistical evidence, which is provided by the multi-seed results in Table 6 and Figure 3.

Table 6. Main UCI HAR robust aggregation comparison under clean and adversarial conditions. Values correspond to mean $\pm$ standard deviation across five independent random seeds.

Method	Attack	Final Accuracy	Macro-F1
FedAvg	Clean	0.896 $\pm$ 0.035	0.895 $\pm$ 0.036
Median	Clean	0.885 $\pm$ 0.050	0.876 $\pm$ 0.066
Trimmed Mean	Clean	0.897 $\pm$ 0.036	0.892 $\pm$ 0.044
Krum	Clean	0.659 $\pm$ 0.082	0.614 $\pm$ 0.112
Multi-Krum	Clean	0.845 $\pm$ 0.047	0.825 $\pm$ 0.074
Trust-FedAvg	Clean	0.906 $\pm$ 0.015	0.905 $\pm$ 0.015
FedAvg	Sign-flip	0.447 $\pm$ 0.149	0.345 $\pm$ 0.148
Median	Sign-flip	0.835 $\pm$ 0.062	0.822 $\pm$ 0.080
Trimmed Mean	Sign-flip	0.419 $\pm$ 0.059	0.290 $\pm$ 0.082
Krum	Sign-flip	0.705 $\pm$ 0.118	0.664 $\pm$ 0.141
Multi-Krum	Sign-flip	0.790 $\pm$ 0.092	0.756 $\pm$ 0.121
Trust-FedAvg	Sign-flip	0.862 $\pm$ 0.032	0.854 $\pm$ 0.039
FedAvg	Gaussian noise	0.401 $\pm$ 0.049	0.328 $\pm$ 0.065
Median	Gaussian noise	0.868 $\pm$ 0.072	0.858 $\pm$ 0.088
Trimmed Mean	Gaussian noise	0.545 $\pm$ 0.164	0.496 $\pm$ 0.206
Krum	Gaussian noise	0.707 $\pm$ 0.089	0.669 $\pm$ 0.088
Multi-Krum	Gaussian noise	0.799 $\pm$ 0.061	0.774 $\pm$ 0.077
Trust-FedAvg	Gaussian noise	0.695 $\pm$ 0.094	0.667 $\pm$ 0.114
FedAvg	Scaling	0.870 $\pm$ 0.022	0.869 $\pm$ 0.024
Median	Scaling	0.907 $\pm$ 0.021	0.905 $\pm$ 0.023
Trimmed Mean	Scaling	0.884 $\pm$ 0.021	0.882 $\pm$ 0.023
Krum	Scaling	0.705 $\pm$ 0.118	0.664 $\pm$ 0.141
Multi-Krum	Scaling	0.790 $\pm$ 0.092	0.756 $\pm$ 0.121
Trust-FedAvg	Scaling	0.902 $\pm$ 0.021	0.900 $\pm$ 0.022
FedAvg	Label-flip	0.846 $\pm$ 0.070	0.834 $\pm$ 0.086
Median	Label-flip	0.849 $\pm$ 0.036	0.840 $\pm$ 0.040
Trimmed Mean	Label-flip	0.830 $\pm$ 0.050	0.817 $\pm$ 0.061
Krum	Label-flip	0.683 $\pm$ 0.089	0.616 $\pm$ 0.123
Multi-Krum	Label-flip	0.775 $\pm$ 0.094	0.735 $\pm$ 0.126
Trust-FedAvg	Label-flip	0.888 $\pm$ 0.030	0.884 $\pm$ 0.037

The single-seed convergence curves show that FedAvg under sign-flip attack exhibits large fluctuations and repeated performance degradation throughout training. In contrast, Trust-FedAvg follows a trajectory closer to clean FedAvg and reaches substantially

higher final test accuracy. This behavior suggests that the trust mechanism reduces the influence of adversarial updates that reverse the optimization direction.

To reduce the dependence on a single stochastic run, Figure 3 reports the averaged convergence behavior across five random seeds. The shaded regions represent the standard deviation across seeds.

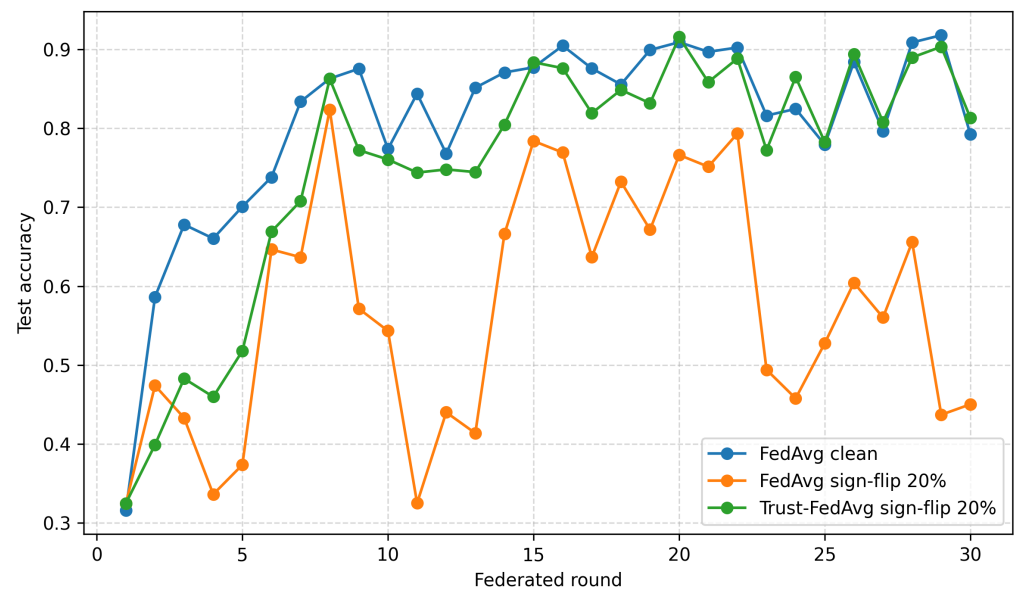


Figure 2. Illustrative single-seed convergence comparison between clean FedAvg, FedAvg under a 20% sign-flip attack, and Trust-FedAvg under the same sign-flip attack on the UCI HAR dataset.

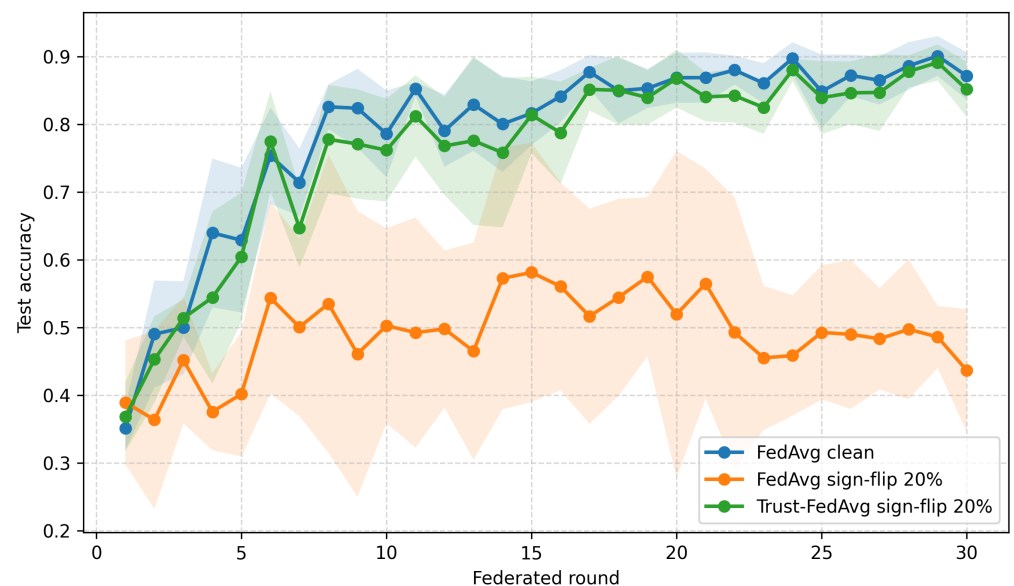


Figure 3. Multi-seed convergence behavior of FedAvg and Trust-FedAvg under a 20% sign-flip attack scenario on UCI HAR. Shaded regions represent the standard deviation across five independent random seeds.

The multi-seed results confirm the pattern observed in the illustrative single-seed example. FedAvg under sign-flip attack remains unstable and converges to substantially lower accuracy, whereas Trust-FedAvg maintains higher average accuracy and a more stable convergence trajectory. These results are consistent with the final metrics reported

in Table 6, where Trust-FedAvg improved final accuracy from 0.447 to 0.862 relative to FedAvg under sign-flip attack.

Figure 4 summarizes the final accuracy of the original FedAvg-versus-Trust-FedAvg adversarial comparison. The figure is retained to visualize the direct effect of the proposed trust mechanism against the conventional FedAvg baseline. The complete robust aggregation comparison, including Median, Trimmed Mean, Krum, Multi-Krum, and label-flip attacks, is provided in Table 6.

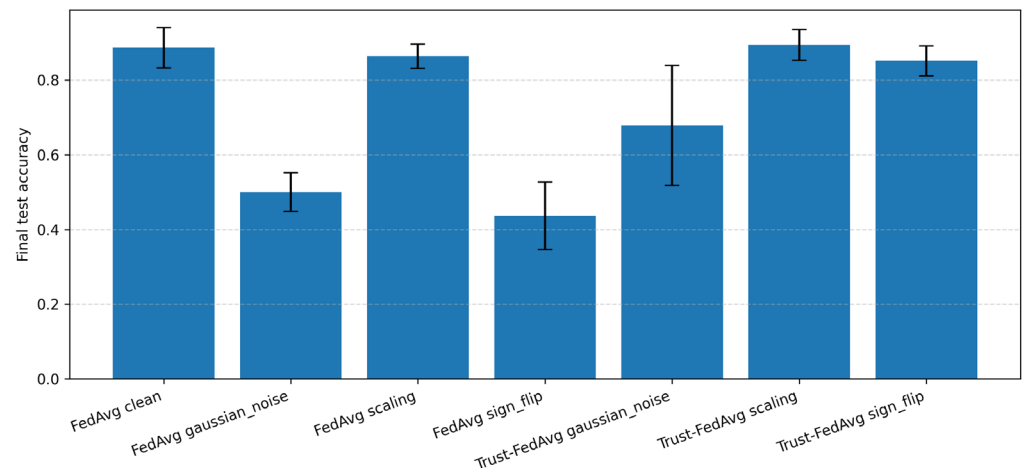


Figure 4. Final accuracy comparison between conventional FedAvg and Trust-FedAvg under selected adversarial attacks on the UCI HAR dataset. Error bars represent standard deviation across five independent random seeds. The broader robust aggregation comparison is reported in Table 6.

The direct FedAvg-versus-Trust-FedAvg comparison shows that trust-aware aggregation substantially improves robustness under sign-flip and Gaussian-noise attacks and provides similar or slightly improved behavior under scaling attacks. However, the expanded benchmark in Table 6 shows that Median and Multi-Krum can outperform Trust-FedAvg under Gaussian-noise attacks. Therefore, Figure 4 should be interpreted as a direct comparison with the conventional FedAvg baseline, not as evidence of universal superiority over all robust aggregation methods.

5.3. Trust-Score Behavior

To better understand the mechanism behind the observed robustness improvements, Figure 5 presents the distribution of round-wise mean trust scores assigned to benign and malicious clients under sign-flip, Gaussian-noise, and scaling attacks. The trust-score analysis was conducted across five random seeds and all communication rounds in which the corresponding clients participated and successfully produced model updates.

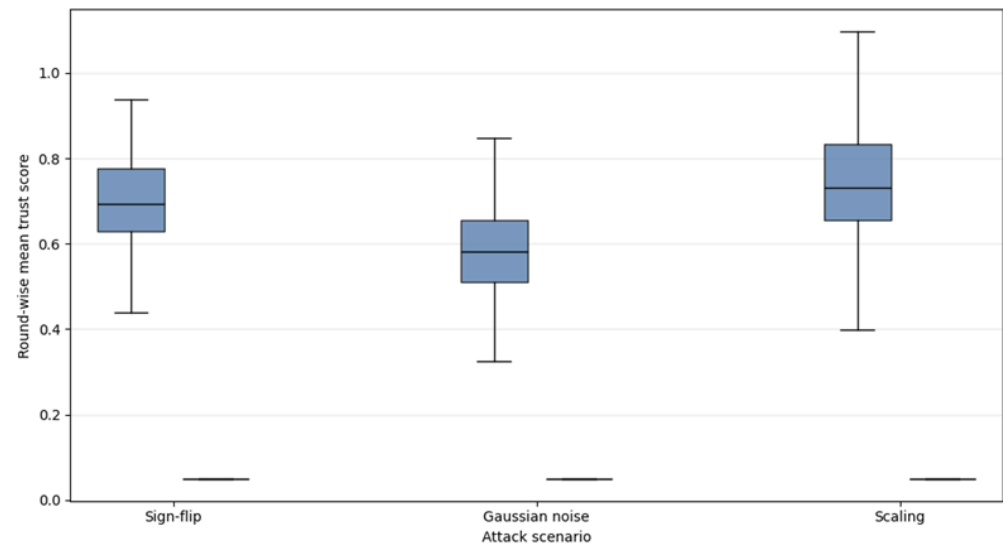


Figure 5. Distribution of round-wise mean trust scores assigned to benign and malicious clients under different adversarial attacks. Boxplots summarize results across five random seeds and all communication rounds in which the corresponding clients participated and successfully uploaded model updates.

Across all three attack scenarios, malicious clients consistently received trust values at or very close to the minimum trust threshold of 0.05. In contrast, benign clients received substantially higher trust scores. Averaged across rounds and seeds, the mean benign-client trust was approximately 0.687 under sign-flip attacks, 0.588 under Gaussian-noise attacks, and 0.752 under scaling attacks, whereas malicious-client trust remained approximately 0.05. These results confirm that the inverse-distance trust mechanism can effectively distinguish between strongly inconsistent adversarial updates and benign updates in the evaluated settings.

The trust-score distribution also helps explain why Trust-FedAvg performs particularly well under sign-flip attacks. Sign-flip manipulation produces updates that strongly diverge from the global optimization trajectory, causing malicious clients to be down-weighted aggressively. By contrast, Gaussian-noise attacks introduce high-dimensional stochastic perturbations whose effect may vary across parameters and rounds. Although Trust-FedAvg still improves substantially over FedAvg under Gaussian noise, the stronger performance of Median and Multi-Krum in Table 6 indicates that coordinate-wise filtering and client-client consistency mechanisms may be more effective against this perturbation type.

These observations are important for positioning the proposed approach relative to existing robust aggregation methods. Trust-FedAvg offers a simple client-global consistency mechanism that is effective when malicious updates are clearly inconsistent with the current global model. However, its effectiveness depends on the relationship between benign update diversity, attack magnitude, and the geometry of the model-parameter space. This limitation motivates the non-IID sensitivity analysis presented in the next subsection.

5.4. Sensitivity to Non-IID Client Heterogeneity

Because the proposed trust mechanism relies on the distance between local and global model parameters, it is important to evaluate whether benign clients with strongly heterogeneous local data may be penalized by the trust-scoring rule. To address this issue, a dedicated non-IID sensitivity analysis was conducted using Dirichlet concentration pa-

rameters $\alpha \in \{0.1, 0.3, 0.5, 1.0\}$ (Table 7). Lower values of α correspond to stronger heterogeneity and more imbalanced local class distributions.

Table 7. Sensitivity to Dirichlet non-IID heterogeneity on UCI HAR under clean and sign-flip conditions. Values correspond to final accuracy mean $\pm$ standard deviation across five seeds.

Dirichlet α	FedAvg Clean	Trust-FedAvg Clean	FedAvg Sign-flip	Multi-Krum Sign-Flip	Trust-FedAvg Sign-Flip
0.1	0.673 $\pm$ 0.045	0.706 $\pm$ 0.063	0.278 $\pm$ 0.177	0.508 $\pm$ 0.083	0.483 $\pm$ 0.188
0.3	0.880 $\pm$ 0.052	0.890 $\pm$ 0.036	0.260 $\pm$ 0.131	0.669 $\pm$ 0.145	0.723 $\pm$ 0.141
0.5	0.896 $\pm$ 0.035	0.906 $\pm$ 0.015	0.447 $\pm$ 0.149	0.790 $\pm$ 0.092	0.862 $\pm$ 0.032
1.0	0.919 $\pm$ 0.018	0.919 $\pm$ 0.020	0.495 $\pm$ 0.204	0.896 $\pm$ 0.031	0.901 $\pm$ 0.008

The clean results show that Trust-FedAvg does not substantially degrade performance relative to FedAvg across the evaluated non-IID levels. In fact, Trust-FedAvg slightly improved final accuracy under $\alpha = 0.1$, $\alpha = 0.3$, and $\alpha = 0.5$, while matching FedAvg under $\alpha = 1.0$. This suggests that the minimum trust threshold and dataset-size-aware renormalization help prevent excessive suppression of benign heterogeneous clients under clean conditions.

Under sign-flip attacks, Trust-FedAvg substantially improved over FedAvg for all values of α . However, the most heterogeneous setting, $\alpha = 0.1$, revealed an important limitation: Multi-Krum achieved 0.508 ± 0.083 , while Trust-FedAvg achieved 0.483 ± 0.188 . This indicates that under extreme statistical heterogeneity, client-global distance may become less reliable as a sole trust indicator because benign clients can naturally deviate from the current global model trajectory. For $\alpha \geq 0.3$, Trust-FedAvg outperformed Multi-Krum under sign-flip attacks, reaching 0.723, 0.862, and 0.901 final accuracy for $\alpha = 0.3, 0.5$, and 1.0 , respectively.

These results directly address the concern that distance-based trust scoring may penalize honest but distributionally distinct clients. The proposed method remains effective across a broad range of non-IID settings, but its advantage is reduced under the most extreme heterogeneity condition. This supports the inclusion of future temporal trust accumulation, layer-wise normalization, or validation-assisted trust scoring to better distinguish malicious deviation from legitimate non-IID client behavior.

5.5. Robustness Under Intermittent Connectivity

To evaluate the resilience of the proposed framework under intermittent edge-network conditions, additional experiments were conducted with probabilistic communication dropout. These experiments emulate situations in which selected clients fail to upload local model updates due to unstable wireless connectivity, temporary device unavailability, or energy-related communication interruptions. Dropout probabilities $p_{\text{drop}} \in \{0.0, 0.1, 0.3, 0.5\}$ were evaluated under clean and sign-flip conditions.

Figure 6 illustrates the robustness trend for FedAvg and Trust-FedAvg under increasing communication dropout. The complete dropout comparison, including Median and Trimmed Mean, is summarized in Table 8.

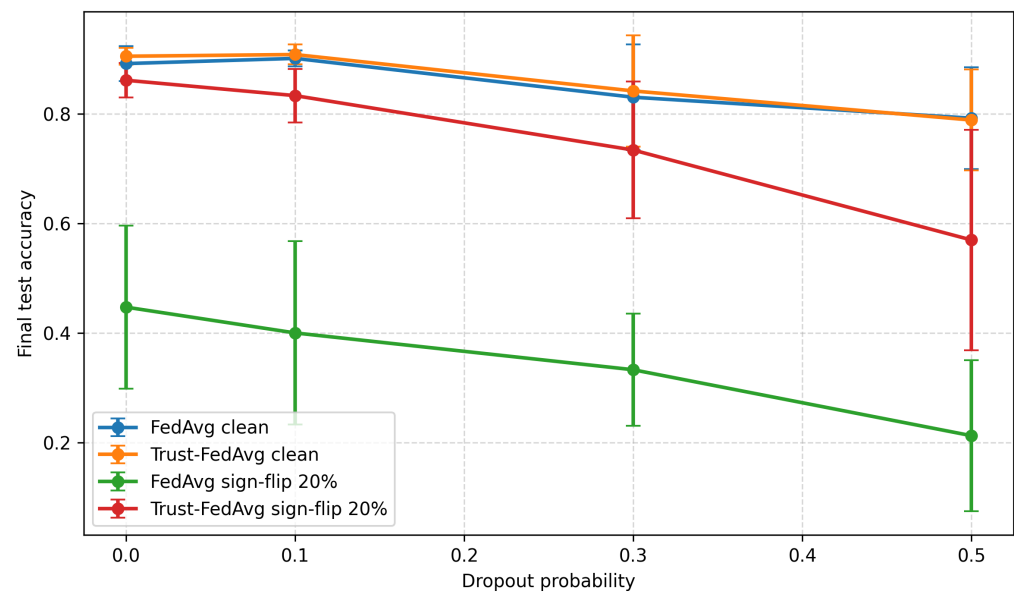


Figure 6. Robustness of FedAvg and Trust-FedAvg under intermittent communication-dropout conditions on the UCI HAR dataset. Error bars represent standard deviation across five independent random seeds.

Under clean conditions, FedAvg and Trust-FedAvg exhibited similar behavior as dropout increased. Trust-FedAvg achieved 0.906 ± 0.015 with no dropout and 0.811 ± 0.112 under 50% dropout, while FedAvg achieved 0.896 ± 0.035 and 0.819 ± 0.114 , respectively. This indicates that communication dropout alone affects both methods similarly when adversarial manipulation is absent.

The impact of dropout became much more severe under sign-flip attacks. FedAvg degraded from 0.447 ± 0.149 with no dropout to 0.213 ± 0.138 under 50% dropout. Trimmed Mean also degraded substantially, reaching 0.146 ± 0.063 under 50% dropout. Median was effective with no dropout, achieving 0.835 ± 0.062 , but became unstable when dropout was introduced, dropping to 0.170 ± 0.017 at 10% dropout and 0.248 ± 0.143 at 50% dropout. In contrast, Trust-FedAvg maintained substantially higher robustness across all dropout levels, achieving 0.862 ± 0.032 , 0.834 ± 0.049 , 0.734 ± 0.125 , and 0.570 ± 0.201 for dropout probabilities of 0.0, 0.1, 0.3, and 0.5, respectively.

Table 8. Federated learning robustness under sign-flip attacks and intermittent communication dropout. Values correspond to final accuracy and macro-F1 mean $\pm$ standard deviation across five seeds.

Method	Dropout	Final Accuracy	Macro-F1	Mean Effective Clients	Skipped Rounds
FedAvg	0.0	0.447 ± 0.149	0.345 ± 0.148	5.00	0.0
FedAvg	0.1	0.400 ± 0.167	0.276 ± 0.178	4.39	0.0
FedAvg	0.3	0.333 ± 0.103	0.201 ± 0.114	3.50	0.2
FedAvg	0.5	0.213 ± 0.138	0.097 ± 0.073	2.48	0.6
Median	0.0	0.835 ± 0.062	0.822 ± 0.080	5.00	0.0
Median	0.1	0.170 ± 0.017	0.048 ± 0.004	4.39	0.0
Median	0.3	0.193 ± 0.069	0.070 ± 0.052	3.50	0.2
Median	0.5	0.248 ± 0.143	0.146 ± 0.167	2.48	0.6
Trimmed Mean	0.0	0.419 ± 0.059	0.290 ± 0.082	5.00	0.0
Trimmed Mean	0.1	0.252 ± 0.140	0.131 ± 0.134	4.39	0.0
Trimmed Mean	0.3	0.233 ± 0.094	0.102 ± 0.062	3.50	0.2
Trimmed Mean	0.5	0.146 ± 0.063	0.043 ± 0.015	2.48	0.6

Trust-FedAvg	0.0	0.862 ± 0.032	0.854 ± 0.039	5.00	0.0
Trust-FedAvg	0.1	0.834 ± 0.049	0.822 ± 0.056	4.39	0.0
Trust-FedAvg	0.3	0.734 ± 0.125	0.685 ± 0.175	3.50	0.2
Trust-FedAvg	0.5	0.570 ± 0.201	0.494 ± 0.256	2.48	0.6

Under the most severe evaluated condition, combining 20% sign-flip malicious clients with 50% communication dropout, Trust-FedAvg improved final accuracy by 35.7 percentage points over FedAvg, from 0.213 to 0.570. It also outperformed Median by 32.2 percentage points and Trimmed Mean by 42.4 percentage points. Statistical testing confirmed that, under this severe dropout condition, Trust-FedAvg significantly outperformed FedAvg in final accuracy according to the paired *t*-test ($p = 0.0293$), as well as Median ($p = 0.0125$) and Trimmed Mean ($p = 0.0075$).

These results indicate that communication dropout can amplify the destabilizing impact of adversarial updates by reducing the effective number of benign clients available for aggregation. Under such conditions, fixed robust-statistical rules may become unstable because the number and composition of received updates vary from round to round. Trust-FedAvg remains more resilient because it evaluates each received update relative to the current global model and then continuously reweights the available client contributions. Nevertheless, the method does not eliminate the impact of severe dropout: Trust-FedAvg still degraded from (0.811) under clean 50% dropout to (0.570) under sign-flip plus 50% dropout.

5.6. Auxiliary Bulyan Comparison

Because Bulyan requires a sufficient number of participating clients relative to the assumed number of Byzantine clients, it was evaluated in a separate auxiliary setting with 20 clients and full client participation. This configuration differs from the main UCI HAR setup, where only approximately five clients participate per round before dropout. The auxiliary Bulyan experiment was therefore designed to provide a fairer comparison under conditions in which Bulyan is more appropriate (Table 9)

Table 9. Auxiliary Bulyan comparison on UCI HAR under a 20-client full-participation setting with $\alpha = 1.0$. Values correspond to mean $\pm$ standard deviation across five seeds.

Method	Attack	Final Accuracy	Macro-F1	Mean Round Time
FedAvg	Clean	0.902 ± 0.008	0.900 ± 0.009	0.558 s
Trust-FedAvg	Clean	0.900 ± 0.013	0.898 ± 0.014	0.509 s
Bulyan	Clean	0.902 ± 0.008	0.900 ± 0.009	0.508 s
FedAvg	Sign-flip	0.483 ± 0.025	0.348 ± 0.035	0.601 s
Trust-FedAvg	Sign-flip	0.883 ± 0.014	0.879 ± 0.016	0.506 s
Bulyan	Sign-flip	0.895 ± 0.017	0.893 ± 0.018	0.644 s

Under clean conditions, FedAvg, Trust-FedAvg, and Bulyan achieved nearly identical performance. Under sign-flip attacks, Bulyan achieved the highest final accuracy, 0.895 ± 0.017 , followed closely by Trust-FedAvg with 0.883 ± 0.014 . The difference between Bulyan and Trust-FedAvg was not statistically significant across the five seeds, whereas Trust-FedAvg substantially outperformed FedAvg. In terms of runtime, Bulyan required approximately 0.644 s per communication round under sign-flip attack, compared with 0.506 s for Trust-FedAvg, corresponding to approximately 27% higher round-level runtime.

This auxiliary result reinforces the balanced interpretation of the proposed method. Bulyan can provide slightly stronger robustness when its client-participation assumptions

are satisfied, but Trust-FedAvg remains competitive while preserving a simpler aggregation structure and lower round-level runtime in the evaluated setting.

5.7. Auxiliary CIFAR-10 CNN Validation

To address the limitation of relying exclusively on UCI HAR and an MLP model, an auxiliary CIFAR-10 experiment was conducted using a compact convolutional neural network. This experiment was not intended to represent an IoT sensor deployment. Instead, it was included to evaluate whether the aggregation behavior observed on UCI HAR remains partially consistent in a higher-dimensional CNN setting. The experiment used three seeds, 10 clients, 50% client participation, $\alpha = 0.5$, and 20 communication rounds (Table 10).

Table 10. Auxiliary CIFAR-10 CNN validation. Values correspond to mean $\pm$ standard deviation across three seeds.

Method	Attack	Final Accuracy	Macro-F1
FedAvg	Clean	0.573 $\pm$ 0.022	0.562 $\pm$ 0.028
Trimmed Mean	Clean	0.521 $\pm$ 0.012	0.499 $\pm$ 0.015
Multi-Krum	Clean	0.451 $\pm$ 0.013	0.402 $\pm$ 0.026
Trust-FedAvg	Clean	0.575 $\pm$ 0.016	0.564 $\pm$ 0.019
FedAvg	Sign-flip	0.158 $\pm$ 0.052	0.073 $\pm$ 0.052
Trimmed Mean	Sign-flip	0.165 $\pm$ 0.018	0.083 $\pm$ 0.031
Multi-Krum	Sign-flip	0.384 $\pm$ 0.072	0.345 $\pm$ 0.080
Trust-FedAvg	Sign-flip	0.410 $\pm$ 0.029	0.362 $\pm$ 0.039
FedAvg	Gaussian noise	0.100 $\pm$ 0.004	0.049 $\pm$ 0.018
Trimmed Mean	Gaussian noise	0.099 $\pm$ 0.002	0.033 $\pm$ 0.006
Multi-Krum	Gaussian noise	0.374 $\pm$ 0.050	0.325 $\pm$ 0.055
Trust-FedAvg	Gaussian noise	0.100 $\pm$ 0.000	0.018 $\pm$ 0.000

The auxiliary CNN results show that Trust-FedAvg achieved clean performance comparable to FedAvg and the highest sign-flip robustness among the evaluated methods. Under sign-flip attack, Trust-FedAvg achieved 0.410 $\pm$ 0.029 final accuracy, compared with 0.158 $\pm$ 0.052 for FedAvg and 0.384 $\pm$ 0.072 for Multi-Krum. However, under Gaussian-noise attack, Trust-FedAvg degraded to near-random performance, whereas Multi-Krum retained substantially better accuracy.

This auxiliary result is important because it clarifies the scope of the proposed approach. Trust-FedAvg can generalize its sign-flip robustness behavior beyond the UCI HAR MLP setting, but it does not provide universal robustness in higher-dimensional CNN models, especially under severe stochastic parameter perturbations. This limitation supports the need for future hybrid trust-aware and robust-statistical aggregation mechanisms, particularly for high-dimensional deep-learning architectures.

5.8. Computational Complexity, Runtime, and Scalability

In addition to predictive robustness, computational cost is an important requirement for practical IoT-oriented FL systems. Table 11 summarizes the qualitative aggregation complexity of representative methods. Conventional FedAvg has linear aggregation complexity because it performs weighted averaging. Median and Trimmed Mean require coordinate-wise sorting or order-statistic operations. Krum, Multi-Krum, and Bulyan require client-client distance computations and are therefore more expensive as the number of participating clients increases. Trust-FedAvg requires only client-global distance computation, preserving a simple aggregation structure while introducing trust-aware weighting.

Table 11. Qualitative comparison of aggregation complexity and scalability characteristics of representative federated learning aggregation methods.

Method	Aggregation Complexity	Client–Client Distance Computation	Client–Global Distance Computation	Additional Trust Management	IoT Suitability
FedAvg [3]	$O(N)$	No	No	No	High
Median	$O(N \log N)$	No	No	No	High
Trimmed Mean [21]	$O(N \log N)$	No	No	No	High
Krum [20]	$O(N^2)$	Yes	No	No	Moderate
Multi-Krum [20]	$O(N^2)$	Yes	No	No	Moderate
Bulyan [22]	$O(N^2)$	Yes	No	No	Moderate
Trust-FedAvg	$O(N)$	No	Yes	Lightweight	High

Empirical runtime measurements were collected as end-to-end round-level wall-clock times on a CPU-only Windows 11 laptop. These timings include local training, attack injection, aggregation, evaluation, and logging, and therefore should be interpreted as simulation-level runtime rather than isolated server-side aggregation time. In the dropout campaign under sign-flip attack and no dropout, FedAvg required 0.310 s per round, whereas Trust-FedAvg required 0.322 s per round, corresponding to an approximate 3.9% increase. In the auxiliary Bulyan experiment, Trust-FedAvg required 0.506 s per round under sign-flip attack, while Bulyan required 0.644 s per round.

These results suggest that the additional trust computation introduces limited overhead in the evaluated simulation setting, especially compared with more complex Byzantine-resilient methods when their assumptions are satisfied. However, the current runtime measurements are not a substitute for detailed deployment profiling. Future work should include isolated server-side aggregation timing, memory profiling, communication-cost analysis, latency measurements, and energy consumption evaluation on real or emulated IoT hardware.

5.9. Discussion

The obtained results demonstrate that IoT-oriented federated learning environments require aggregation mechanisms capable of simultaneously handling adversarial manipulation, statistical heterogeneity, and unstable communication conditions. Conventional FedAvg performs adequately under clean conditions, but becomes highly vulnerable when malicious participants and intermittent connectivity are simultaneously present. Communication dropout substantially amplifies the destabilizing effect of adversarial updates by reducing the effective aggregation population and increasing the relative influence of unreliable participants during model synchronization.

The main strength of Trust-FedAvg lies in its ability to down-weight updates that strongly diverge from the current global model trajectory. This explains its strong performance under sign-flip attacks, where malicious updates intentionally reverse the optimization direction. On UCI HAR, Trust-FedAvg improved final accuracy under sign-flip attack from 0.447 with FedAvg to 0.862. Under combined sign-flip and 50% communication dropout, it improved final accuracy from 0.213 with FedAvg to 0.570. These results support the practical value of lightweight trust-aware aggregation in intermittently connected edge-learning scenarios.

At the same time, the extended benchmark demonstrates that robust aggregation behavior is attack-dependent. Median aggregation was stronger under Gaussian-noise attacks on UCI HAR, while Multi-Krum was clearly superior to Trust-FedAvg under Gaussian-noise perturbations in the auxiliary CIFAR-10 CNN setting. This confirms that different attacks induce different failure modes and that no single lightweight aggregation mechanism should be assumed to dominate across all adversarial scenarios. In this re-

spect, the present results are consistent with the broader robust FL literature, where Byzantine-resilient, coordinate-wise robust, and trust-aware methods often exhibit different strengths depending on the attack model, data heterogeneity, and number of participating clients [20–22].

The non-IID sensitivity analysis further clarifies the conditions under which distance-based trust scoring is effective. Trust-FedAvg remained stable across the evaluated α values and substantially improved over FedAvg under sign-flip attacks. However, at the most heterogeneous setting, $\alpha = 0.1$, Multi-Krum slightly outperformed Trust-FedAvg. This suggests that client-global distance alone may be insufficient to distinguish malicious deviation from legitimate non-IID client diversity under extreme heterogeneity. Future trust-aware mechanisms should therefore consider temporal trust evolution, layer-wise normalization, validation-assisted scoring, or hybrid combinations with robust-statistical aggregation.

The auxiliary Bulyan experiment also provides an important positioning result. Bulyan achieved slightly higher sign-flip accuracy than Trust-FedAvg in the 20-client full-participation setting, but the difference was not statistically significant and Bulyan required higher round-level runtime. This suggests that stronger Byzantine-resilient methods can be highly effective when their assumptions are satisfied, whereas Trust-FedAvg offers a simpler and competitive alternative that is easier to integrate into lightweight IoT-oriented simulations.

From a practical perspective, the proposed framework is best interpreted as a lightweight simulation-driven trust-aware aggregation framework rather than as a complete deployment-ready IoT FL system. The simulation includes non-IID client partitions, adversarial update manipulation, and probabilistic communication dropout, but it does not model physical-layer communication, bandwidth constraints, asynchronous aggregation, real-device energy consumption, hardware faults, or on-device execution. Therefore, the results support robustness within controlled edge-learning simulations, but do not yet constitute real-world IoT deployment validation.

Regarding internal validity, the results depend on the selected trust formulation, attack intensity, client-partition strategy, and number of participating clients. Regarding external validity, the primary evaluation is based on UCI HAR and a lightweight MLP, while the CIFAR-10 experiment is only an auxiliary CNN validation rather than an IoT-specific deployment scenario. Regarding deployment validity, detailed real-device profiling, energy measurement, communication overhead evaluation, and asynchronous network behavior remain outside the scope of the present study.

Nevertheless, the revised experimental evaluation substantially strengthens the original study. The proposed method is now compared against multiple robust baselines, evaluated under additional label-flip attacks, tested across different non-IID levels, analyzed under communication dropout, assessed with statistical tests, compared with Bulyan in an auxiliary valid setting, and complemented with a CNN-based auxiliary experiment. These additions provide a more balanced and defensible assessment of the robustness, limitations, and computational trade-offs of Trust-FedAvg.

FLTrust was not included because it requires a trusted server-side root dataset, which is outside the privacy and data-availability assumptions adopted in the present simulation framework. Additionally, backdoor attacks were not included in the present revision because the primary UCI HAR benchmark relies on pre-engineered features rather than raw sensor streams, making trigger design less straightforward; this remains an important direction for future work.

Future research directions may therefore include:

1. evaluation using additional IoT sensor datasets and industrial edge-learning tasks;

2. integration with larger deep-learning architectures, including convolutional, recurrent, graph-based, and transformer-based models;
3. hybrid aggregation mechanisms combining trust-aware weighting with coordinate-wise robust statistics;
4. temporal trust accumulation, historical reputation modeling, and adaptive behavioral profiling;
5. decentralized and hierarchical trust-aware federated learning architectures;
6. blockchain-assisted trust management and auditable reputation mechanisms;
7. communication-efficient trust-aware optimization strategies;
8. explicit latency, memory, energy, and communication-overhead benchmarking;
9. large-scale intelligent edge-IoT deployment studies under asynchronous and dynamic network conditions.

6. Conclusions

This paper presented a simulation-driven trust-aware federated learning framework for robust intelligent IoT networks operating under heterogeneous, unreliable, and adversarial edge conditions. The proposed approach integrates a lightweight dynamic trust evaluation mechanism into the federated aggregation process, enabling adaptive weighting of local client updates according to behavioral consistency and optimization reliability.

A comprehensive experimental evaluation was conducted using the UCI HAR dataset under non-independent and identically distributed (non-IID) federated learning conditions. Multiple adversarial scenarios were considered, including sign-flip, Gaussian-noise, scaling, and label-flip attacks, as well as intermittent communication dropout conditions that emulate unstable IoT connectivity. The evaluation was further extended with robust aggregation baselines, including Median, Trimmed Mean, Krum, Multi-Krum, and an auxiliary Bulyan configuration, together with non-IID sensitivity analysis, statistical testing, runtime profiling, and an auxiliary CIFAR-10 CNN validation scenario.

The obtained results demonstrated that conventional FedAvg aggregation remains highly vulnerable to malicious or unreliable client participation, particularly under sign-flip and noisy-update attacks. In contrast, the proposed Trust-FedAvg framework substantially improved robustness over FedAvg and remained competitive with established robust aggregation strategies, especially under directional model-manipulation attacks and intermittent-connectivity conditions.

Under sign-flip attacks involving 20% malicious participants, the proposed framework achieved a final accuracy of 86.2% on UCI HAR, compared with 44.7% for conventional FedAvg. The communication-dropout experiments further demonstrated that trust-aware aggregation becomes increasingly beneficial under unstable edge-network conditions where reduced client participation amplifies the relative influence of malicious updates. Under the most severe evaluated setting, combining 20% sign-flip malicious clients with 50% communication dropout, Trust-FedAvg achieved 57.0% final accuracy, compared with 21.3% for FedAvg, 24.8% for Median, and 14.6% for Trimmed Mean.

An important advantage of the proposed framework is its computational simplicity and practical applicability in controlled IoT-oriented edge-learning simulations. Unlike several Byzantine-resilient aggregation strategies that require pairwise client-update comparisons or stronger assumptions regarding the number of malicious participants, the proposed method relies on lightweight client-global model-consistency analysis. The runtime measurements indicated only limited round-level overhead relative to FedAvg in the evaluated CPU-based simulation setting, while the auxiliary Bulyan experiment showed that Trust-FedAvg remained competitive with a stronger Byzantine-resilient baseline at lower round-level runtime.

At the same time, the extended evaluation also identified important limitations. Trust-FedAvg was not universally superior across all attack types. In particular, coordinate-wise Median and Multi-Krum achieved stronger robustness under severe Gaussian-noise perturbations in some settings, and Multi-Krum slightly outperformed Trust-FedAvg under the most heterogeneous non-IID condition ($\alpha = 0.1$). These findings indicate that distance-based trust scoring is especially effective against directional and label-based manipulation, but may be less suitable as a standalone defense against high-dimensional stochastic perturbations or extreme benign-client heterogeneity.

Globally, the results presented here suggest that trust-aware federated aggregation provides an effective and scalable approach to improving the robustness of distributed learning in intelligent IoT ecosystems characterized by heterogeneous sensing conditions, unreliable communication, and adversarial behavior. However, the proposed framework should be interpreted as a lightweight and reproducible simulation-driven evaluation framework rather than as a fully deployment-validated IoT FL system. The current simulation does not model physical-layer communication, asynchronous aggregation, bandwidth constraints, device energy consumption, hardware failures, or execution on real IoT devices.

The public release of the implementation, configuration files, seeds, and aggregated experimental logs further supports reproducibility and facilitates independent verification of the reported results.

Future work will focus on:

1. extending the framework to additional IoT sensor datasets, industrial edge-learning tasks, and heterogeneous application domains;
2. developing hybrid aggregation strategies that combine trust-aware weighting with coordinate-wise robust statistics or Byzantine-resilient selection mechanisms;
3. incorporating temporal trust accumulation and reputation-aware mechanisms;
4. evaluating decentralized, hierarchical, and blockchain-assisted trust-aware federated architectures;
5. investigating communication-efficient trust-aware optimization strategies;
6. performing detailed latency, memory, energy, and communication-overhead benchmarking;
7. validating the proposed framework in large-scale real-world IoT deployments.

Author Contributions: Conceptualization, M.J.C.S.R.; Methodology, M.J.C.S.R., C.S. and F.B.; Software, M.J.C.S.R. and F.B.; Validation, C.S. and F.B.; Formal analysis, M.J.C.S.R. and C.S.; Investigation, C.S.; Data curation, F.B.; Writing—original draft, M.J.C.S.R.; Writing—review and editing, C.S. and F.B.; Visualization, M.J.C.S.R. and F.B.; Funding acquisition, F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw datasets used in this study are publicly available from their original sources. The UCI Human Activity Recognition dataset is available from the UCI Machine Learning Repository, and CIFAR-10 is available through standard public machine-learning dataset repositories and the torchvision interface. The source code, configuration files, random seeds, and aggregated experimental logs supporting the revised experiments are publicly available at <https://github.com/mcabralreis/trust-fedavg-iot-applsci> (accessed on 7 June 2026).and archived on Zenodo at <https://doi.org/10.5281/zenodo.21131862> (accessed on 7 June 2026). The raw datasets are not redistributed in the repository.

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

References

1. Dritsas, E.; Trigka, M. Federated Learning for IoT: A Survey of Techniques, Challenges, and Applications. *J. Sens. Actuator Netw.* **2025**, *14*, 9. <https://doi.org/10.3390/jsan14010009>.
2. Kaur, I.; Jadhav, A.J. Federated Learning in IoT: A Survey from a Resource-Constrained Perspective. In Proceedings of the 2023 International Conference on Artificial Intelligence Robotics, Signal and Image Processing (AIROSIP), Yogyakarta, Indonesia, 9–10 August 2023; pp. 376–381.
3. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*; PMLR; JMLR, Inc.: Norfolk, MA, USA, 2017; Volume 54, pp. 1273–1282.
4. Gugueoth, V.; Safavat, S.; Shetty, S. Security of Internet of Things (IoT) Using Federated Learning and Deep Learning—Recent Advancements, Issues and Prospects. *ICT Express* **2023**, *9*, 941–960. <https://doi.org/10.1016/j.ict.2023.03.006>.
5. Hameed, R.T.; Mohamad, O.A. Federated Learning in IoT: A Survey on Distributed Decision Making. *Babylon. J. Internet Things* **2023**, *2023*, 1–7. <https://doi.org/10.58496/BJIoT/2023/001>.
6. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
7. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning*; JMLR, Inc.: Norfolk, MA, USA, 2020; Volume 119, pp. 5132–5143.
8. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7611–7623.
9. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Seneviratne, A.; Li, J.; Vincent Poor, H. Federated Learning for Internet of Things: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1622–1658. <https://doi.org/10.1109/COMST.2021.3075439>.
10. Hoffpauir, K.; Simmons, J.; Schmidt, N.; Pittala, R.; Briggs, I.; Makani, S.; Jararweh, Y. A Survey on Edge Intelligence and Lightweight Machine Learning Support for Future Applications and Services. *J. Data Inf. Qual.* **2023**, *15*, 20:1–20:30. <https://doi.org/10.1145/3581759>.
11. Huang, W.; Ye, M.; Shi, Z.; Wan, G.; Li, H.; Du, B.; Yang, Q. Federated Learning for Generalization, Robustness, Fairness: A Survey and Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9387–9406. <https://doi.org/10.1109/TPAMI.2024.3418862>.
12. Nguyen, T.D.; Nguyen, T.; Nguyen, P.L.; Pham, H.H.; Doan, K.D.; Wong, K.-S. Backdoor Attacks and Defenses in Federated Learning: Survey, Challenges and Future Research Directions. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107166. <https://doi.org/10.1016/j.engappai.2023.107166>.
13. Farid Babar, F.; Khan, S.; Parkinson, S. Defending Federated Learning against Adversarial Attacks: A Systematic Literature Review. *Cyber Secur. Appl.* **2026**, *4*, 100127. <https://doi.org/10.1016/j.csa.2026.100127>.
14. Tariq, A.; Serhani, M.A.; Sallabi, F.M.; Barka, E.S.; Qayyum, T.; Khater, H.M.; Shuaib, K.A. Trustworthy Federated Learning: A Comprehensive Review, Architecture, Key Challenges, and Future Research Prospects. *IEEE Open J. Commun. Soc.* **2024**, *5*, 4920–4998. <https://doi.org/10.1109/OJCOMS.2024.3438264>.
15. Sánchez Sánchez, P.M.; Huertas Celdrán, A.; Xie, N.; Bovet, G.; Martínez Pérez, G.; Stiller, B. FederatedTrust: A Solution for Trustworthy Federated Learning. *Future Gener. Comput. Syst.* **2024**, *152*, 83–98. <https://doi.org/10.1016/j.future.2023.10.013>.
16. Xu, J.; Zhang, C.; Jin, L.; Su, C. A Trust-Aware Incentive Mechanism for Federated Learning with Heterogeneous Clients in Edge Computing. *J. Cybersecurity Priv.* **2025**, *5*, 37. <https://doi.org/10.3390/jcp5030037>.
17. Zhao, S.; Pu, J.; Fu, X.; Liu, L.; Dai, F. Byzantine-Robust Federated Learning with Ensemble Incentive Mechanism. *Future Gener. Comput. Syst.* **2024**, *159*, 272–283. <https://doi.org/10.1016/j.future.2024.05.017>.
18. Bhagoji, A.N.; Chakraborty, S.; Mittal, P.; Calo, S. Analyzing Federated Learning through an Adversarial Lens. In *Proceedings of the 36th International Conference on Machine Learning*; PMLR; JMLR, Inc.: Norfolk, MA, USA, 2019; Volume 97, pp. 634–643.
19. Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; Shmatikov, V. How To Backdoor Federated Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*; PMLR; JMLR, Inc.: Norfolk, MA, USA, 2020; Volume 108, pp. 2938–2948.
20. Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; NIPS'17; Curran Associates Inc.: Long Beach, California, USA, 2017; pp 118–128.

21. Yin, D.; Chen, Y.; Kannan, R.; Bartlett, P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR; JMLR, Inc.: Norfolk, MA, USA, 2018; Volume 80, pp. 5650–5659.
22. El-Mhamdi, E.-M.; Guerraoui, R.; Rouault, S. The Hidden Vulnerability of Distributed Learning in Byzantium. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR; JMLR, Inc.: Norfolk, MA, USA, 2018; Volume 80, pp. 3521–3530.

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.