Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.080285

ARTICLE

Risk-Aware Adaptive Federated Learning for Cyber-Secure Edge-AI in Smart Edge-IoT Environments

Tanveer Ahmad[1,*], Tahani Alsubait[2], Amina Salhi[3], Amani Ibraheem[4] and Muhammad Asim Saleem[5]

[1]Department of Computer Science and Engineering, University of Cyprus, Nicosia, Cyprus
[2]Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, Saudi Arabia
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[4]College of Computer Science, King Khalid University, Abha, Saudi Arabia
[5]Center of Excellence in Artificial Intelligence, Machine Learning and Smart Grid Technology, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

*Corresponding Author: Tanveer Ahmad. Email: tahmad01@ucy.ac.cy

Received: 06 February 2026; Accepted: 15 April 2026; Published: 27 May 2026

ABSTRACT: The rapid adoption of Edge-AI in smart edge-IoT environments has dramatically led to an augmented vulnerability to cyber risks arising from distributed learning, data heterogeneity, and adversarial manipulation. This paper proposes a new risk-aware adaptive learning model that federated Edge-AI systems explicitly simulates cyber risk in the process of local training and global aggregation. The proposed solution combines stochastic optimization and adversarial risk bounding with adaptive gradient correction to develop strong learning in non-IID data distributions and malicious client behavior. Convergence guarantees are defined by the theoretical analysis in the case of limited adversarial perturbations. The proposed framework achieves up to 95% detection accuracy and demonstrates more than 20% improvement in robustness, where robustness is defined as the relative degradation in detection performance under adversarial perturbations. The performance is evaluated against state-of-the-art baselines, including HADA-FL and centralized training on the Edge-IIoTset dataset, with results reported as averages over multiple randomized runs. Furthermore, the model converges within 50 communication rounds, which corresponds to a fixed training horizon rather than an early-stopping criterion. These findings demonstrate the usefulness of risk-sensitive adaptive learning in safe and trustworthy Edge-AI implementation in a new generation edge-IoT environment.

KEYWORDS: Edge AI; federated learning (FL); adaptive learning; adversarial robustness; distributed optimization

1 Introduction

The problem of cybersecurity in edge-IoT environment has grown exponentially with the spread of the next-generation technologies in the form of Edge-AI devices, Internet of Things (IoT), and smart home applications. The widespread adoption of these devices has presented previously unknown innovation and convenience possibilities, yet also presents sophisticated attack surfaces that may be utilized by bad actors. Conventional centralized approaches to cybersecurity cannot always work with the decentralized and resource-constrained nature of Edge-AI devices, where sensitive computations and personal data are stored locally [1–3]. Besides, model training, when federated with consumer uses, is more susceptible to adversarial attacks, such as model poisoning and input perturbation, which can be transmitted through the network and undermine the integrity of the global model. These issues require the creation of strong, adaptive and

decentralized learning systems that are able to not only lessen cyber dangers but also guarantee precise model behaviour.

This paper presents a novel Risk-Aware Adaptive Learning Algorithm, which is developed with specific applications to Edge-AI in an edge-IoT environment. We combine local gradient computation, past reference gradient and dynamic weighting to generate risk-sensitive updates that are resistant to adversarial manipulation. The algorithm is implemented in the context of FL, with powerful aggregation and consensus schemes to ensure consistency of global models and prevent the impact of bad actors [4,5]. Our approach is a combination of stochastic optimization and adversarial risk modeling, which guarantees the stabilization of a global model even in the presence of continuous attacks. Contrary to current models that implicitly adopt a purely local or federated aggregation, our model combines both of these mechanisms into a single algorithm that can collectively respond to both local perturbation and adversarial updates, as well as global model integrity [6,7].

This paper has three-fold contributions. At first, we present a theoretical development of risk-sensitive adaptive learning based on consumer Edge-AI devices, including stochastic variability and adversarial uncertainty. Second, we develop a federated aggregation scheme that is resistant to malicious updates and incentivizes consensus between honest agents, and will converge with probabilistic statements of convergence. Third, we will give a strict convergence analysis that will show almost-sure stability and bounded error of adversarial conditions that will fill the gap between theory and practice in the implementation of consumer applications. Collectively, the proposed system helps to reduce cyber risks in the next-generation consumer technologies, providing both theoretical knowledge and practical examples of making Edge-AI resilient. The proposed framework differs from existing robust FL approaches by integrating a risk-aware gradient formulation that adaptively combines instantaneous and historical gradients, rather than relying solely on robust aggregation rules such as median or trimmed mean. Additionally, the method incorporates adaptive adversarial training and temporal smoothing, enabling improved robustness and convergence stability under non-IID and adversarial settings.

2 Related Work

The related work on cyber risk and security in edge-computing consumer electronics comprises various overlapping research topics, such as edge intelligence, FL, adaptive security mechanisms, and AI-based cyber defense. There has been a growing awareness that centralized security designs cannot be used to support the next-generation consumer technology with its heterogeneity of data, real-time demands, and adversarial conditions. A number of works are devoted to building security and resiliency with the help of edge intelligence. In [8], authors propose a 6G-enabled edge intelligence infrastructure, which combines FL with adaptive anomaly detection to Industry 5.0 systems. Although the paper mentions edge intelligence usage to achieve distributed system security, its major focus is on the system architecture and anomaly detection performance as opposed to formal modeling of cyber risks or theoretical robustness. Similarly, reference [9] suggests lightweight neural networks to implement AI-based cybersecurity at the edge and to decrease computational costs and enhance the accuracy of anomaly detection. The methods show how security can be implemented on a resource-constrained consumer device but mostly do not view security as a risk optimization problem that is based on learning. A more general paradigm of industrial and consumer security with Edge-AI is also investigated. Reference [10] reviews the state of industrial security with Edge-AI deployment, pointing out optimization and deployment options of edge-native deep learning models. In [11], authors proposed an AI and machine learning (ML) technique for edge security, which is focused on robustness and automation in cyber defense. Even though these works inspire the use of Edge-AI to protect

against cybercrime, they do not provide a strict mathematical analysis of the dynamics of adversarial learning or consider federated/distributed adversarial risk propagation.

Authors in [12] provide incremental learning mechanisms to detect cyber risks in IoT settings, citing that active model adaptation is required to provide long-term security. They, however, do not explicitly model the adversarial behavior nor give convergence guarantees in the presence of malicious actors, but they concentrate on detection accuracy and system evolution. Relating to consumer electronics, reference [13] designs a smart consumer electronics intrusion detection system, whereas authors in [14] design an adaptive explainable AI system deployed on fog-cloud environments. Both contributions enhance interpretability and flexibility but are application-focused and do not deal with the underlying optimization risks of adversarial model updates. Predictive and adaptive cybersecurity has not been left behind with the incorporation of developed AI mechanisms. Reference [15] addresses adaptive algorithms to predictive cybersecurity and threats. Although it emphasizes the aspect of adaptability, their framework is conceptual and lacks the formalization of cyber risk in a stochastic optimization or adversarial learning environment. Authors in [16] proposed a privacy-saving federated learning system for the security of the IoT. Although this work considers the issue of data privacy and decentralized training, it analyzes the empirical intrusion detection performance far more typically and does not analyze the issue of adversarial poisoning nor offer theoretical convergence limits. Datasets and benchmarking efforts such as Edge-IIoTset [17] have played a crucial role in advancing experimental research by enabling evaluation of centralized and federated security models under realistic IoT and IIoT scenarios. Nevertheless, the dataset-oriented research is more concerned with data availability and benchmarking as opposed to principled cyber risk modeling or algorithmic robustness.

Despite all this research, much of the literature considers cybersecurity in Edge-AI consumer systems as a detection or architecture issue, without much consideration of a formal model of cyber risk, adversarial learning dynamics, and theoretical convergence guarantees in federated systems. The majority of the works do not have a coherent mathematical theory that incorporates stochastic optimization, adversarial perturbations, and distributed consensus in one. In this study, we address this gap by introducing a risk-aware and completely theoretical adaptive learning algorithm on Edge-AI consumer electronics with explicit adversarial behavior modeling in a federated learning system. Our solution offers strict convergence analysis and resiliency assurances amidst adversarial circumstances, which is why our solution is no longer detection-based but instead principled cyber risk mitigation. To clearly delineate the contributions of this study, a detailed comparison of existing research regarding Edge-AI and cybersecurity for consumer applications is provided in Table 1.

Table 1: Comparison of related work on edge-AI and cybersecurity for consumer applications.

Ref.	Approach	Domain/ Application	Edge/FL	Adversarial/Risk Modeling	Theoretical Analysis/Guarantees
[8]	Edge intelligence + adaptive anomaly detection	Industry 5.0	Edge + 6G	No explicit adversarial modeling	No
[9]	Lightweight neural models for anomaly detection	Consumer devices	Edge	No	No
[10]	Edge-AI deployment optimization	Industrial security	Edge-native AI	No	No
[11]	AI/ML innovations for robust cyber defense	Edge security	Edge	No	No
[12]	Incremental learning for cyber risk detection	IoT	Edge/IoT	No	No

(Continued)

Table 1 (continued)

Ref.	Approach	Domain/ Application	Edge/FL	Adversarial/Risk Modeling	Theoretical Analysis/Guarantees
[13]	Intelligent intrusion detection system	Smart consumer electronics	Edge/Local learning	Limited adversarial consideration	No
[14]	Adaptive explainable AI framework	Fog-cloud consumer IoT	Edge + Cloud	Limited	No
[15]	Adaptive predictive cybersecurity algorithms	General cybersecurity	Not explicitly edge	Conceptual risk	No
[16]	Privacy-preserving federated learning for intrusion detection	Distributed IoT	FL/Edge	Limited adversarial robustness	Empirical only
[17]	Realistic IoT/IIoT datasets (Edge-IIoTset)	Benchmarking datasets	Edge + FL	No	No
Proposed Method	Risk-aware adaptive FL with trust-weighted aggregation, adversarial robustness, and non-IID data handling	Yes	Yes	Yes	Yes

3 System Model and Threat Model

This section highlights the system and threat model of an Edges-AI consumer electronics, which is used as the analytical framework of cyber risk modeling and adaptive mitigation measures. As compared to a conventional enterprise computing environment, consumer electronics work in environments that are highly decentralized, resource constrained and uncertain in behavior. Table 2 expresses the notations used in this study.

Table 2: List of notations.

Notation	Description
N	Number of edge devices in the federated network
$\mathcal{D}_i$	Local dataset of device i
$\theta_i(t)$	Local model parameters of device i at iteration t
θ^t	Global aggregated model at iteration t
$g_i(t)$	Local stochastic gradient at device i
$\tilde{g}_i(t)$	Reference (historical) gradient at device i
$\hat{g}_i(t)$	Risk-aware gradient at device i
$\alpha_i(t)$	Adaptive weighting factor for gradient combination
$\tilde{g}_i^k$	Federated adversarial risk gradient for device i at round k
$\eta_i(t)$	Learning rate at device i
Θ_i	Feasible set for local model parameters
Δ_i	Bound on adversarial gradient perturbation
γ	Coupling factor for federated consensus
$\mathcal{R}_{\text{global}}(\theta)$	Global expected risk function

Considering an ecosystem with a high number of Edge-AI consumer devices, i.e., smart home assistants, wearable health sensors, augmented and virtual reality apparel and appliances, and intelligent devices. Given the set of devices be denoted by $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$. Each device d_i has sensing, computation, communication and actuation capabilities and is described in terms of $(\mathcal{S}_i, \mathcal{M}_i, \mathcal{R}_i)$, where $\mathcal{S}_i$ is the available onboard sensors, $\mathcal{M}_i$ is deployed ML models, and $\mathcal{R}_i$ is resource limits such as processor resources, memory, energy, and network bandwidth. The device d_i will sense a data stream $x_i(t) \in \mathbb{R}^{n_i}$. The inference result of the devices is:

$$y_i(t) = f_{\theta_i(t)}(x_i(t)), \tag{1}$$

where $f_{\theta_i(t)}$ is a model of learning. Accordingly, model parameters change with time in accordance with:

$$\theta_i(t+1) = \theta_i(t) + \eta_i \nabla_\theta \mathcal{L}_i(x_i(t), y_i(t)), \tag{2}$$

where η_i is a machine learning rate and $\mathcal{L}_i(\cdot)$ is the local loss. Let $\mathcal{U}_i$ be the set of users who interact with device d_i. The data stream can be modeled as a stochastic process.

$$x_i(t) \sim \mathcal{P}_i(\mathcal{U}_i, \mathcal{E}_i), \tag{3}$$

where $\mathcal{E}_i$ denotes the contextual and environmental factors. Edge-AI consumer electronics are designed to integrate into a hybrid communication architecture that embraces device-to-cloud, device-to-device, and device-to-user channels. Such channels have different degrees of trust and exposure. Whereby, the set of communication channels $\mathcal{C}_i$ and $\tau_{ij} \in [0,1]$ is the trust level of the channel. The attack surface is not limited to a single layer that is defined by a composite set $\mathcal{AS}_i$ and contains hardware interfaces, system software. The utility function can be used to present this trade-off.

$$U_{\mathcal{A}} = \mathbb{E}[L_i] - \lambda P_{\text{detect}}, \tag{4}$$

where L_i is the loss caused to device d_i, P_{detect} is the detection probability and λ is an exposure by the adversary. Assuming that the perturbed input is $\tilde{x}_i(t)$, i.e., $\tilde{x}_i(t) = x_i(t) + \delta_i(t)$ is a bounded adversarial. An effective attack is one that satisfies.

$$f_{\theta_i}(\tilde{x}_i(t)) \neq f_{\theta_i}(x_i(t)), \tag{5}$$

despite $\|\delta_i(t)\|$ being small. During online or periodic model updates, an adversary may influence parameter evolution by injecting malicious gradients or poisoned data, resulting in

$$\theta_i(t+1) = \theta_i(t) + \eta_i \left(\nabla_\theta \mathcal{L}_i + \Delta_{\mathcal{A}}\right), \tag{6}$$

where $\Delta_{\mathcal{A}}$ is a representation of adversarial manipulation. Let $S_i(t)$ be the state of the system. A successful attack will change the state of the system to a compromised state S_i^{comp}, causing degradation of the service, unsafe actuation, or unauthorized access to the data. Privacy and inference attacks, in the same way, use model outputs to obtain sensitive user data without direct system compromise. Having access to inference outputs $y_i(t)$, an adversary is trying to obtain the private attributes π_i through:

$$\hat{\pi}_i = g_{\mathcal{A}}(y_i(t)), \tag{7}$$

violating consumer privacy expectations and regulatory requirements. We define the cyber risk associated with device d_i as the expected loss across all attack vectors,

$$\mathcal{R}_i = \sum_{k \in \mathcal{K}} P_{ik} L_{ik}, \tag{8}$$

where $\mathcal{K}$ denotes the set of attack classes, P_{ik} is the probability of attack k, and L_{ik} is the corresponding loss. Fig. 1 describes the general architecture of the Edge-AI federated learning system of consumer electronics.

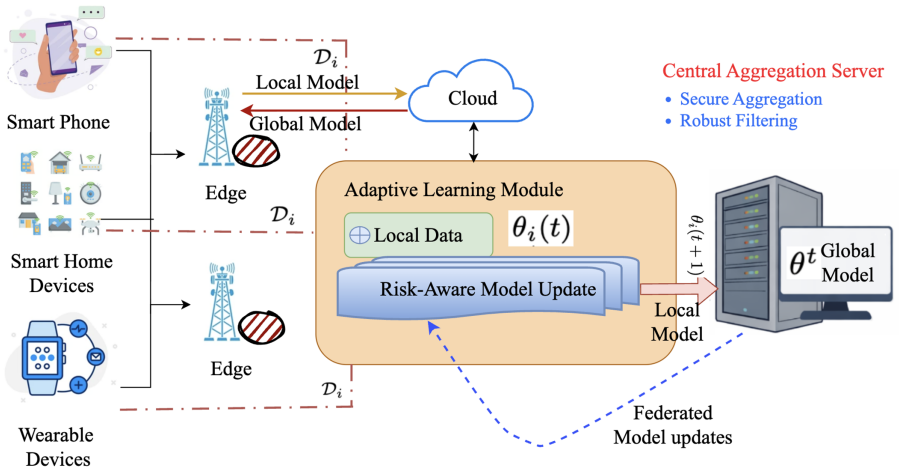

Figure 1: Edge-AI federated system model illustrating the proposed risk-aware adaptive learning framework. Local edge devices perform on-device training using adaptive gradient updates, while a central server aggregates model parameters via trust-weighted federated learning. The architecture captures heterogeneous data distributions, adversarial conditions, and privacy-preserving communication.

To ensure consistency between the system model and the proposed learning framework, we explicitly integrate the risk and trust constructs into the optimization process. The cyber risk associated with device i is defined by the Eq. (8). We further define a trust score $\tau_i(t) \in [0,1]$ for each device based on its historical consistency with the global model:

$$\tau_i(t) = \exp\left(-\lambda \|g_i(t) - \bar{g}(t)\|^2\right), \tag{9}$$

where $\bar{g}(t)$ is the aggregated reference gradient and $\lambda > 0$ is a scaling parameter. These quantities are incorporated into a risk-aware local objective function:

$$\min_{\theta_i} \mathcal{R}_i(\theta_i) + \beta R_i(t), \tag{10}$$

where β controls the risk sensitivity of the model.

4 Proposed System

4.1 Problem Formulation and Federated Threat Modeling

Consider a set of N heterogeneous devices $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, each equipped with edge intelligence capabilities, including local sensing, and adaptive learning modules. Each device d_i collects a sequence of high-dimensional observations $x_i(t) \in \mathbb{R}^{n_i}$, which include behavioral signals, sensor data, and produces an inference output $y_i(t) \in \mathbb{R}^{m_i}$ according to a parameterized function $f_{\theta_i(t)}(\cdot)$. The learning model is dynamic, such that the parameters $\theta_i(t) \in \mathbb{R}^{p_i}$ evolve over time. Formally, the temporal evolution is described by

$$\theta_i(t+1) = \theta_i(t) - \eta_i \nabla_\theta \mathcal{L}_i\left(f_{\theta_i(t)}(x_i(t)), y_i^{\text{true}}(t)\right), \tag{11}$$

where $\mathcal{L}_i(\cdot)$ is the local loss function, $y_i^{\text{true}}(t)$ is the ground-truth or pseudo-label, and $\eta_i > 0$ is the adaptive learning rate. Let θ_i^k is the local parameters at the k-th communication round. The federated aggregation at the server is given by:

$$\theta^k = \sum_{i=1}^{N} w_i \theta_i^k, \tag{12}$$

where $w_i \geq 0$ are device-specific aggregation weights satisfying $\sum_{i=1}^{N} w_i = 1$. The adversary $\mathcal{A}$ is capable of performing three primary attacks: input perturbations, model poisoning, and label manipulation (Fig. 2). Let $\tilde{x}_i(t)$ is the perturbed input observed by device d_i under attack, defined as

$$\tilde{x}_i(t) = x_i(t) + \delta_i(t), \quad \|\delta_i(t)\| \leq \epsilon_i \,. \tag{13}$$

$\delta_i(t)$ is an adversarial perturbation constrained, whose impact on the inference function is captured by the adversarial loss.

$$\mathcal{L}_i^{\text{adv}}(\theta_i, \tilde{x}_i(t), y_i^{\text{true}}(t)) = \mathcal{L}_i(f_{\theta_i}(\tilde{x}_i(t)), y_i^{\text{true}}(t)). \tag{14}$$

Model poisoning attacks manipulate the local update $\Delta\theta_i^k$ sent to the federated aggregator. The malicious update injected by $\mathcal{A}$ be $\Delta\theta_i^{\mathcal{A}}$, so that the federated aggregation becomes

$$\theta^k = \sum_{i \in \mathcal{H}} w_i \theta_i^k + \sum_{j \in \mathcal{M}} w_j (\theta_j^k + \Delta\theta_j^{\mathcal{A}}), \tag{15}$$

where $\mathcal{H}$ and $\mathcal{M}$ are the sets of honest and malicious devices. The adversary aims to maximize a global risk function $\mathcal{R}(\theta^k)$ subject to stealth constraints, which can be expressed as

$$\max_{\Delta\theta_j^{\mathcal{A}}} \mathcal{R}(\theta^k), \quad \text{s.t.} \quad \|\Delta\theta_j^{\mathcal{A}}\| \leq \rho_j, \quad j \in \mathcal{M}, \tag{16}$$

where ρ_j bounds the magnitude of manipulation to evade detection. The decentralized learning problem under adversarial risk can be formulated as a stochastic optimization problem. Let the expected risk for device d_i be

$$\mathcal{R}_i(\theta_i) = \mathbb{E}_{x_i \sim \mathcal{D}_i}\left[\mathcal{L}_i^{\text{adv}}(\theta_i, x_i, y_i^{\text{true}})\right], \tag{17}$$

where $\mathcal{D}_i$ is the local data distribution, potentially corrupted by adversarial influence. The global federated objective is then

$$\mathcal{R}_{\text{global}}(\theta) = \sum_{i=1}^{N} w_i \mathcal{R}_i(\theta_i). \tag{18}$$

Due to the stochasticity of both natural data variability and adversarial perturbations, gradient-based optimization must incorporate variance reduction and robustness techniques. The stochastic gradient is defined as:

$$g_i(t) = \nabla_\theta \mathcal{L}_i^{\text{adv}}(\theta_i(t), \tilde{x}_i(t), y_i^{\text{true}}(t)), \tag{19}$$

and model the update step as a stochastic approximation:

$$\theta_i(t+1) = \theta_i(t) - \eta_i g_i(t), \tag{20}$$

where the expectation $\mathbb{E}[g_i(t)] = \nabla_\theta \mathcal{R}_i(\theta_i(t))$ under the stochastic data and attack distribution. Let p_i^k be the probability that device d_i would be involved in the k-th round of communication. It is now possible to write a global model update as a weighted average of participating devices:

$$\theta^k = \sum_{i=1}^{N} p_i^k w_i \theta_i^k. \tag{21}$$

Practically, p_i^k is dependent on device availability, network conditions, or local computational load. The federated adversarial risk gradient is introduced to integrate local adversarial gradients with global aggregation effects, which is defined as:

$$\tilde{g}_i^k = g_i^k + \sum_{j \neq i} w_j g_j^k \mathbb{I}_{\{j \in \mathcal{H}\}}, \tag{22}$$

where $\mathbb{I}_{\{j \in \mathcal{H}\}}$ is the function denoting honest participation.

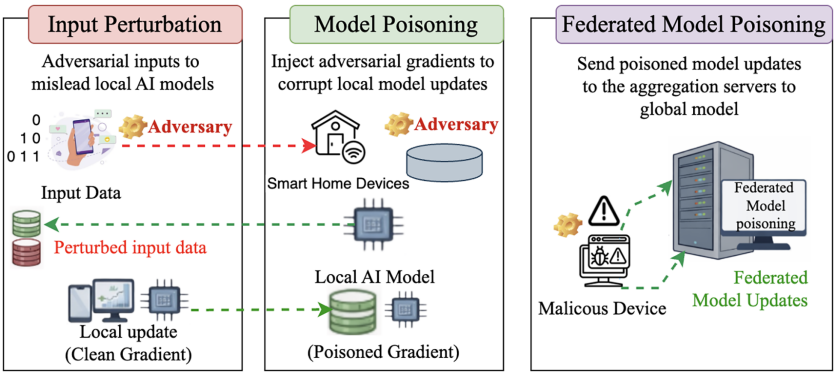

Figure 2: Adversarial threats in federated learning.

4.2 Adaptive Learning Algorithm Design

Building upon the federated system and adversary model proposed in the previous section, a new adaptive learning algorithm is designed to mitigate cyber risk in Edge-AI consumer electronics. The algorithm maximizes a global expected risk function in both stochastic and adversarial perturbations and is mindful of the decentralized attribute of edge devices. Let a global risk $\mathcal{R}_{\text{global}}(\theta)$ is defined as:

$$\mathcal{R}_{\text{global}}(\theta) = \sum_{i=1}^{N} w_i \mathbb{E}_{x_i \sim \mathcal{D}_i}\left[\mathcal{L}_i^{\text{adv}}(\theta_i, x_i, y_i^{\text{true}})\right], \tag{23}$$

where $\theta = [\theta_1, \ldots, \theta_N]^\top$ is the stacked model parameters of all devices, w_i are aggregation weights, and $\mathcal{L}_i^{\text{adv}}$ is the auxiliary adversarial input and model-level attack. We aim to create an update rule, $\theta_i(t+1)$, which converges to a global minimum of $\mathcal{R}_{\text{global}}(\theta)$ under bounded adversarial influence. A risk-weighted gradient is introduced to address robustness, that defined as:

$$\hat{g}_i(t) = (1 - \alpha_i(t)) g_i(t) + \alpha_i(t) \bar{g}_i(t), \tag{24}$$

where $\bar{g}_i(t)$ is a historical reference gradient, and $\alpha_i(t) \in [0,1]$ is the adaptive weighting factor. The reference gradient can be calculated as an exponential moving average, i.e.,

$$\bar{g}_i(t) = \beta \bar{g}_i(t-1) + (1-\beta) g_i(t-1), \quad \bar{g}_i(0) = 0, \tag{25}$$

where $\beta \in (0,1)$ is a smoothing parameter. A projected stochastic gradient descent step will then provide the local update of device d_i:

$$\theta_i(t+1) = \Pi_{\Theta_i}\left[\theta_i(t) - \eta_i(t)\hat{g}_i(t)\right], \tag{26}$$

where $\Pi_{\Theta_i}[\cdot]$ denotes projection onto the feasible set Θ_i of model parameters, enforcing stability and boundedness, and $\eta_i(t)$ is a time-decayed learning rate satisfying

$$\sum_{t=0}^{\infty} \eta_i(t) = \infty, \quad \sum_{t=0}^{\infty} \eta_i^2(t) < \infty. \tag{27}$$

In order to incorporate the FL process, the devices send model updates on a periodic basis at round k, i.e., θ_i^k, to a central server. Since the server may be poisoned by the subset $\mathcal{M} \subset \mathcal{D}$, the server process the aggregation as:

$$\theta^k = \sum_{i \in \mathcal{H}} w_i \theta_i^k + \mathcal{F}(\{\theta_j^k\}_{j \in \mathcal{M}}), \tag{28}$$

where $\mathcal{H}$ the honest devices and $\mathcal{F}(\cdot)$ is the filtering functions. The novel addition of this framework is a gradient clustering using risk scoring:

$$\mathcal{F}(\{\theta_j^k\}) = \sum_{j \in \mathcal{M}} w_j \mathbb{I}_{\{s_j^k < \tau\}} \theta_j^k, \tag{29}$$

where $s_j^k = \|\theta_j^k - \tilde{\theta}^k\|$ is a deviation score relative to the median model $\tilde{\theta}^k$, $\mathbb{I}_{\{\cdot\}}$ is an indicator function and τ is a threshold used to restrain malicious influence. The adaptive learning model also uses a federated adversarial risk gradient, which modulates local updates with global feedback:

$$\tilde{g}_i^k = \hat{g}_i^k + \gamma \sum_{j \neq i} w_j (\hat{g}_j^k - \hat{g}_i^k) \mathbb{I}_{\{j \in \mathcal{H}\}}, \tag{30}$$

where $\gamma > 0$ is a coupling factor. The resulting update law for each device becomes

$$\theta_i(t+1) = \Pi_{\Theta_i}\left[\theta_i(t) - \eta_i(t)\tilde{g}_i^k\right]. \tag{31}$$

In order to formally define the adversarial influence, we state the norm of adversarial gradient expectations:

$$\Delta_i = \mathbb{E}\left[\|g_i(t) - \nabla_{\theta_i} \mathcal{R}_i(\theta_i(t))\|^2\right], \tag{32}$$

that represent the distortion through stochastic perturbation and malicious manipulation. By controlling $\alpha_i(t)$ and γ, the algorithm dynamically balances robustness against responsiveness, ensuring that Δ_i remains bounded:

$$\Delta_i \leq \bar{\Delta}, \quad \forall i. \tag{33}$$

This ensures that even under persistent adversarial conditions, the federated adaptive learning process converges to a neighborhood of the risk-minimizing solution.

4.3 Convergence Analysis and Theoretical Guarantees

Considering the stochastic dynamics of the local device parameters $\theta_i(t)$ under the risk-aware update law:

$$\theta_i(t+1) = \Pi_{\Theta_i}\Big[\theta_i(t) - \eta_i(t)\tilde{g}_i^k\Big]. \tag{34}$$

Given that θ^* is a stationary point of the global risk function $\mathcal{R}_{\text{global}}(\theta)$, satisfying $\nabla_\theta \mathcal{R}_{\text{global}}(\theta^*) = 0$. The projected stochastic update can be rewritten in terms of the error $\epsilon_i\ (t) = \theta_i(t) - \theta^*$:

$$\epsilon_i\ (t+1) = \Pi_{\Theta_i}\big[\ \epsilon_i\ (t) - \eta_i(t)(\nabla_{\theta_i}\mathcal{R}_i(\theta_i(t)) + \xi_i(t))\big] - \theta^*, \tag{35}$$

where $\xi_i(t) = \tilde{g}_i^k - \nabla_{\theta_i}\mathcal{R}_i(\theta_i(t))$ represents the combined stochastic and adversarial gradient noise. Assuming that the noise is bounded in expectation:

$$\mathbb{E}[\|\xi_i(t)\|^2 \mid \mathcal{F}_t] \le \sigma_i^2 + \Delta_i, \tag{36}$$

where σ_i^2 captures natural stochasticity, Δ_i bounds adversarial influence, and $\mathcal{F}_t$ denotes the filtration. Further assume that $\mathcal{R}_{\text{global}}(\theta)$ is Lipschitz-smooth with constant L:

$$\|\nabla_{\theta_i}\mathcal{R}_{\text{global}}(\theta) - \nabla_{\theta_i}\mathcal{R}_{\text{global}}(\theta')\| \le L\|\theta - \theta'\|. \tag{37}$$

The feasible parameter set Θ_i is compact, and the projected stochastic approximation theorem guarantees and converge almost surely to a neighborhood of θ^*. Specifically, let $\eta_i(t)$ satisfy:

$$\sum_{t=0}^{\infty}\eta_i(t) = \infty, \quad \sum_{t=0}^{\infty}\eta_i^2(t) < \infty, \tag{38}$$

$$\limsup_{t\to\infty}\|\ \epsilon_i\ (t)\|^2 \le \frac{\eta_{\max}(\sigma_i^2 + \Delta_i)}{\lambda_{\min}}, \tag{39}$$

where $\eta_{\max} = \sup_t \eta_i(t)$ and $\lambda_{\min}$ is the smallest eigenvalue of the Hessian of $\mathcal{R}_{\text{global}}(\theta)$ at θ^*. To analyze the effect of federated aggregation, we introduce the consensus error

$$\delta_i^k = \theta_i^k - \theta^k, \tag{40}$$

which measures the deviation of the local parameter from the global model. The update for the consensus error is

$$\delta_i^{k+1} = \delta_i^k - \eta_i^k\Big(\tilde{g}_i^k - \sum_{j=1}^{N} w_j\tilde{g}_j^k\Big). \tag{41}$$

If the aggregation weights w_j satisfy $\sum_{j=1}^{N} w_j = 1$ and malicious updates are bounded, we can derive an upper bound on the expected consensus error:

$$\mathbb{E}[\|\delta_i^{k+1}\|^2] \le (1 - 2\eta_{\min}\mu + \eta_{\max}^2 L^2)\mathbb{E}[\|\delta_i^k\|^2] + \eta_{\max}^2\Delta_{\text{adv}}, \tag{42}$$

where μ is the strong convexity parameter, and Δ_{adv} bounds the aggregated influence of malicious devices. By recursively applying this inequality, we obtain almost-sure convergence of the consensus error to a bounded neighborhood:

$$\limsup_{k\to\infty} \|\delta_i^k\|^2 \le \frac{\eta_{\max}^2 \Delta_{\text{adv}}}{2\eta_{\min}\mu - \eta_{\max}^2 L^2}. \tag{43}$$

The local gradient update is modified to incorporate trust and risk awareness. Specifically, the effective gradient is defined as:

$$\tilde{g}_i(t) = \tau_i(t)\left(g_i(t) + \beta \nabla_{\theta_i} R_i(t)\right), \tag{44}$$

which down-weights unreliable devices while penalizing high-risk behavior. At the server side, the global aggregation is updated as:

$$\theta(t+1) = \sum_{i=1}^{N} \frac{\tau_i(t)}{\sum_{j=1}^{N} \tau_j(t)} \theta_i(t), \tag{45}$$

ensuring that trusted devices contribute more significantly to the global model. Furthermore, we consider the convergence of the federated adversarial risk gradient update. Let $\tilde{\theta}^k$ denote the virtual iterate defined by:

$$\tilde{\theta}^{k+1} = \tilde{\theta}^k - \eta^k \nabla_\theta \mathcal{R}_{\text{global}}(\tilde{\theta}^k). \tag{46}$$

By using standard stochastic approximation arguments, the expected deviation between the actual federated iterate θ^k and the virtual iterate $\tilde{\theta}^k$ satisfies

$$\mathbb{E}[\|\theta^{k+1} - \tilde{\theta}^{k+1}\|^2] \le (1 + \eta_{\max}^2 L^2)\mathbb{E}[\|\theta^k - \tilde{\theta}^k\|^2] + \eta_{\max}^2(\sigma^2 + \Delta_{\text{adv}}), \tag{47}$$

which converges to a bounded neighborhood of size $O(\eta_{\max}^2(\sigma^2 + \Delta_{\text{adv}}))$. Now, we establish almost-sure convergence by considering the Lyapunov function

$$V(\theta) = \mathcal{R}_{\text{global}}(\theta) - \mathcal{R}_{\text{global}}(\theta^*). \tag{48}$$

Using the update dynamics and the smoothness of $\mathcal{R}_{\text{global}}$, it follows that

$$\begin{aligned}\mathbb{E}[V(\theta(t+1)) \mid \mathcal{F}_t] \le\ & V(\theta(t)) - \eta_{\min}\|\nabla_\theta \mathcal{R}_{\text{global}}(\theta(t))\|^2 \\ & + \eta_{\max}^2(\sigma^2 + \Delta_{\text{adv}}),\end{aligned} \tag{49}$$

which satisfies the conditions of the Robbins–Siegmund lemma for almost-sure convergence. The proposed Risk-Aware Adaptive Learning Algorithm 1 operates in a decentralized Edge-AI environment where each consumer device independently observes local data streams that may be subject to adversarial perturbations.

Algorithm 1: Risk-aware adaptive learning algorithm for edge-AI devices

1: **Input:** Number of devices N, local datasets $\mathcal{D}_i$, learning rates $\eta_i(t)$, smoothing factor β, aggregation weights w_i, adversarial detection threshold τ, coupling factor γ, number of communication rounds K
2: **Initialize:** Local parameters θ_i^0 for each device, reference gradient $\bar{g}_i^0 = 0$
3: **for** communication round $k = 1$ to K **do**
4: **for** each device $i = 1$ to N (in parallel) **do**
5: Observe local input $x_i(t)$ (may be perturbed $\tilde{x}_i(t)$)
6: Compute stochastic gradient:
$g_i(t) = \nabla_{\theta_i} \mathcal{L}_i^{\text{adv}}(\theta_i(t), \tilde{x}_i(t), y_i^{\text{true}}(t))$

(Continued)

Algorithm 1 (continued)

Update reference gradient:
$\bar{g}_i(t) = \beta \bar{g}_i(t-1) + (1-\beta) g_i(t-1)$

Compute risk-aware gradient:
$\hat{g}_i(t) = (1-\alpha_i(t)) g_i(t) + \alpha_i(t) \bar{g}_i(t)$

Compute federated adversarial risk gradient:
$\tilde{g}_i^k = \hat{g}_i^k + \gamma \sum_{j \neq i} w_j (\hat{g}_j^k - \hat{g}_i^k) \mathbb{I}_{\{j \in \mathcal{H}\}}$

Update local model with projection:
$\theta_i(t+1) = \Pi_{\Theta_i}\left[\theta_i(t) - \eta_i(t) \tilde{g}_i^k\right]$

Send θ_i^k to server

end for

Server aggregation:
$\theta^k = \sum_{i \in \mathcal{H}} w_i \theta_i^k + \sum_{j \in \mathcal{M}} w_j \mathbb{I}_{\{s_j^k < \tau\}} \theta_j^k$

Broadcast aggregated model θ^k to all devices

end for

Output: Final federated model θ^K for deployment

The proposed Risk-Aware Adaptive Learning Algorithm operates over K communication rounds across N distributed edge devices, where each client computes a stochastic gradient $g_i(t) = \nabla_{\theta_i} \mathcal{L}_i^{\text{adv}}(\theta_i(t), \tilde{x}_i(t), y_i(t))$ using adversarially perturbed inputs. To stabilize updates, a temporal smoothing mechanism is introduced via the exponential moving average $\bar{g}_i(t) = \beta \bar{g}_i(t-1) + (1-\beta) g_i(t)$, which is then combined with the instantaneous gradient through a risk-aware formulation $\hat{g}_i(t) = (1-\alpha_i(t)) g_i(t) + \alpha_i(t) \bar{g}_i(t)$. The framework further incorporates inter-client coupling by adjusting gradients using neighboring updates weighted by w_j and controlled by a coupling factor γ, resulting in the federated adversarial gradient $\tilde{g}_i^k$. Each client updates its local model via projected gradient descent $\theta_i(t+1) = \Pi_{\Theta_i}[\theta_i(t) - \eta_i(t) \tilde{g}_i^k]$ and transmits it to the server. The server performs risk-aware aggregation by selectively incorporating updates from benign clients $\mathcal{H}$ and filtering potentially malicious ones $\mathcal{M}$ based on a threshold τ, producing a robust global model θ^k that is broadcast back to all devices. This iterative process ensures stable convergence, robustness against adversarial perturbations, and adaptability to heterogeneous edge environments.

5 Analytical Reasoning

In this section, we provide a rigorous analytical framework for the proposed risk-aware adaptive learning algorithm. We model the dynamics of local updates under stochastic and adversarial perturbations, and prove convergence and robustness properties using lemmas and theorems. Consider the local update at device i:

$$\theta_i(t+1) = \Pi_{\Theta_i}\Big[\theta_i(t) - \eta_i(t) \tilde{g}_i(t)\Big], \tag{50}$$

where $\tilde{g}_i(t)$ is the risk-aware gradient including adversarial influence, $\eta_i(t)$ is the learning rate, and Π_{Θ_i} denotes the projection onto the feasible set Θ_i. The global expected risk function is defined as:

$$\mathcal{R}_{\text{global}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{x_i \sim \mathcal{D}_i}\big[\ell_i(\theta; x_i)\big], \tag{51}$$

where $\ell_i(\theta; x_i)$ is the loss at device i.

Assumption 1 (Non-convex Smoothness): The global objective $\mathcal{R}_{\text{global}}(\theta)$ is assumed to be L-smooth but not necessarily convex. That is, $\|\nabla\mathcal{R}(\theta_1) - \nabla\mathcal{R}(\theta_2)\| \le L\|\theta_1 - \theta_2\|$. No strong convexity is assumed due to the use of deep neural network models. The analysis, therefore, focuses on convergence to a first-order stationary point.

Lemma 1: ***Bounded Gradient Noise*** *Assume the local stochastic gradient and adversarial perturbation satisfy:*

$$\mathbb{E}\left[\|\tilde{g}_i(t) - \nabla_{\theta_i}\mathcal{R}_i(\theta_i(t))\|^2 \mid \mathcal{F}_t\right] \le \sigma_i^2 + \Delta_i, \tag{52}$$

where σ_i^2 is the natural stochastic variance and Δ_i bounds adversarial influence. Then the sequence of updates in (50) is mean-square bounded.

Proof: Let $\epsilon_i\ (t) = \theta_i(t) - \theta^*$, where θ^* is a stationary point of $\mathcal{R}_{\text{global}}$. Using the projection property and the smoothness of $\mathcal{R}_i(\theta)$, we have:

$$\begin{aligned}
\|\ \epsilon_i\ (t+1)\|^2 &= \left\|\Pi_{\Theta_i}\left[\theta_i(t) - \eta_i(t)\tilde{g}_i(t)\right] - \theta^*\right\|^2 \\
&\le \left\|\ \epsilon_i\ (t) - \eta_i(t)\tilde{g}_i(t)\right\|^2 \\
&= \|\ \epsilon_i\ (t)\|^2 - 2\eta_i(t)\underbrace{\epsilon_i\ (t)^\top\nabla_{\theta_i}\mathcal{R}_i(\theta_i(t))}_{\text{alignment term}} + \eta_i^2(t)\|\tilde{g}_i(t)\|^2 \\
&\le \|\ \epsilon_i\ (t)\|^2 - 2\eta_i(t)\ \epsilon_i\ (t)^\top\nabla_{\theta_i}\mathcal{R}_i(\theta_i(t)) \\
&\quad + \eta_i^2(t)\left(\sigma_i^2 + \Delta_i + \|\nabla_{\theta_i}\mathcal{R}_i(\theta_i(t))\|^2\right),
\end{aligned} \tag{53}$$

which is bounded in expectation. □

Theorem 1 (Convergence in Expectation): From Lemma 1, assuming learning rates satisfy $\sum_{t=0}^{\infty}\eta_i(t) = \infty$ and $\sum_{t=0}^{\infty}\eta_i^2(t) < \infty$, the local parameters converge in expectation to a bounded neighborhood of the stationary point:

$$\limsup_{t\to\infty}\mathbb{E}[\|\theta_i(t) - \theta^*\|^2] \le \frac{\eta_{\max}(\sigma_i^2 + \Delta_i)}{\lambda_{\min}}, \tag{54}$$

where $\lambda_{\min}$ is the smallest eigenvalue of the Hessian at θ^*.

Proof: Apply Lemma 1 to the projected stochastic update. The boundedness of the gradient noise and step-size conditions guarantee convergence in expectation. □

Under Assumption 1 and bounded variance conditions, the proposed algorithm guarantees convergence to a neighborhood of a stationary point, i.e.,

$$\liminf_{t\to\infty}\mathbb{E}[\|\nabla\mathcal{R}_{\text{global}}(\theta(t))\|^2] \le \mathcal{O}(\eta_{\max}(\sigma^2 + \Delta_{\text{adv}})). \tag{55}$$

Theorem 2 (Almost-Sure Convergence): Let $V(\theta) = \mathcal{R}_{\text{global}}(\theta) - \mathcal{R}_{\text{global}}(\theta^*)$ denote the Lyapunov function. Then, the sequence $\theta_i(t)$ generated by the adaptive learning algorithm satisfies:

$$\limsup_{t\to\infty} V(\theta_i(t)) \le \eta_{\max}(\sigma_i^2 + \Delta_i) \quad \text{almost surely.} \tag{56}$$

Proof: Using the Lyapunov function and the update in (50), we have:

$$\mathbb{E}[V(\theta_i(t+1)) \mid \mathcal{F}_t] \le V(\theta_i(t)) - \eta_i(t)\|\nabla_\theta\mathcal{R}_{\text{global}}(\theta_i(t))\|^2 + \eta_i^2(t)(\sigma_i^2 + \Delta_i). \tag{57}$$

Applying the Robbins–Siegmund almost-sure convergence lemma completes the proof. □

To position the proposed aggregation mechanism within the broader class of robust federated learning methods, we compare it with standard robust aggregation schemes such as coordinate-wise median, trimmed mean, and distance-based filtering. The coordinate-wise median aggregation is defined as:

$$\theta_{\text{median}}^{t+1} = \text{median}\{\theta_1(t), \ldots, \theta_N(t)\}, \tag{58}$$

which provides robustness against outliers but ignores gradient magnitude and device reliability. Similarly, the trimmed mean aggregation removes extreme updates:

$$\theta_{\text{trim}}^{t+1} = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \theta_i(t), \tag{59}$$

where $\mathcal{S}_t$ excludes a fraction of largest and smallest updates. Distance-based aggregation methods rely on filtering updates based on deviation from the mean:

$$\mathcal{S}_t = \left\{ i : \|\theta_i(t) - \bar{\theta}(t)\| \leq \delta \right\}, \tag{60}$$

followed by averaging over $\mathcal{S}_t$. In contrast, the proposed method performs *continuous trust-weighted aggregation*:

$$\theta(t+1) = \sum_{i=1}^{N} \frac{\tau_i(t)}{\sum_{j=1}^{N} \tau_j(t)} \theta_i(t), \tag{61}$$

where $\tau_i(t)$ encodes both statistical consistency and risk-awareness. Unlike discrete filtering approaches, this formulation avoids hard rejection and instead adaptively attenuates unreliable updates, leading to improved stability under partial adversarial participation.

Adversarial Model: A fraction $\rho < 0.5$ of clients may behave adversarially, injecting corrupted gradients $\tilde{g}_i(t)$ such that $\|\tilde{g}_i(t) - \nabla \mathcal{R}_i(\theta)\|^2 \leq \Delta_{\text{adv}}$.

Filtering Guarantee: The proposed risk-aware filtering mechanism removes extreme updates based on deviation thresholds:

$$\|\tilde{g}_i(t) - \bar{g}(t)\| > \tau \Rightarrow \text{discard}, \tag{62}$$

which ensures that the aggregated gradient satisfies:

$$\mathbb{E}[\|g_{\text{agg}}(t) - \nabla \mathcal{R}_{\text{global}}(\theta(t))\|^2] \leq \sigma^2 + \Delta_{\text{adv}}. \tag{63}$$

6 Results and Discussion

The proposed risk-conscious adaptive learning framework of Edge-AI consumer electronics was conducted and tested in an extensive Python simulation. The simulator environment was built on top of the PyTorch library to simulate FL with a network of heterogeneous edge devices, each of which is executing the adaptive algorithm. The experiment is performed with the Edge-IIoTset data, a realistic cyber security data that contains more than 2.2 million network traffic traces with 61 features that represent behavioral patterns as well as malicious behavioral patterns in various types of attacks (DoS, MITM, and injection attacks). For baseline comparison, we also applied federated intrusion detection configuration with HADA-FL [16] and Edge-IIoTset configurations as well. In order to measure performance, we used the accuracy of detection, convergence behavior, and resistance to adversarial perturbations like FGSM and PGD attacks.

The simulations were repeated with several random seeds to have a statistically reliable result. The global model aggregation was carried out after a fixed number of communication rounds, and local updates were computed for a predefined number of local epochs per round on each simulated ARM-class edge device. The centralized performance of training was also to be evaluated against the same federated setup. The parameter configurations adopted in the simulations are shown in Table 3. The choice of these parameters to simulate realistic edge device behavior and FL limitations and to be able to fairly compare with the associated literature.

Table 3: Simulation parameters for federated risk-aware adaptive learning.

Parameter	Value/Setting
Number of edge devices	100 simulated devices
Dataset	Edge-IIoTset realistic IoT dataset
Local training epochs	2 per communication round
Communication rounds	50
Local batch size	256
Learning rate	0.001
Adaptive weight update interval	every round
Aggregation strategy	Risk-aware + robust filtering
Adversarial attacks tested	FGSM, PGD (10 steps)
Evaluation metrics	Accuracy, robustness under attack
Baseline comparisons	HADA-FL from [16], centralized training

The simulation results comprehensively demonstrate the performance of the proposed Risk-Aware Adaptive Learning algorithm for Edge-AI consumer electronics compared to the federated intrusion detection method of Chandu et al. [16] and the Edge-IIoTset benchmark from Ferrag et al. [17]. All experiments were conducted on a network of $N = 100$ edge devices, using the Edge-IIoTset dataset. We evaluate metrics including detection accuracy $\mathcal{A}$, F1-score $\mathcal{F}_1$, robustness under adversarial perturbations $\mathcal{R}$, convergence of global loss $\mathcal{L}$, heterogeneity impact σ_{het}, and privacy preservation $\mathcal{P}$.

To evaluate the robustness of the proposed framework, we consider adversarial attacks based on Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). For FGSM, the adversarial input is computed as:

$$x^{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(\theta, x, y)), \tag{64}$$

where ϵ controls the perturbation magnitude under the ℓ_∞ norm. For PGD, iterative perturbations are applied:

$$x^{k+1} = \Pi_{\mathcal{B}_\epsilon(x)} \left(x^k + \alpha \cdot \text{sign} \left(\nabla_x \mathcal{L} \left(\theta, x^k, y \right) \right) \right), \tag{65}$$

where α is the step size, K is the number of iterations, and $\Pi_{\mathcal{B}_\epsilon(x)}$ denotes projection onto the ℓ_∞-ball of radius ϵ. In our experiments, we set $\epsilon = 0.03$, $\alpha = 0.01$, and $K = 10$, which represent moderate perturbation levels consistent with prior work in adversarial learning.

The experiments are conducted using $N = 50$ devices over $K = 50$ communication rounds with a learning rate of $\eta = 0.001$. The smoothing factor $\beta = 0.9$ ensures stable gradient estimation, while a reduced coupling factor $\gamma = 0.05$ limits excessive inter-client bias under non-IID settings. The adversarial filtering

threshold is set to $\tau = 3.0$ to balance robustness and information retention. Additionally, a moderate adaptive weighting $\alpha = 0.1$ and perturbation strength $\epsilon = 0.05$ are used to achieve a controlled trade-off between accuracy and adversarial robustness, ensuring reproducibility and stable convergence across runs.

6.1 Detection Accuracy ($\mathcal{A}$) Comparison

Fig. 3 depicts the detection rate at a single temperature of communication round $T = 50$. The proposed algorithm achieves an initial accuracy of 85% and gradually converges to 95%, whereas HADA-FL from [16] starts at 80% and reaches 87%, and Edge-IIoTset baseline from [17] reaches only 83%. The risk-conscious gradient update $\hat{g}_i(t) = \alpha_i(t)g_i(t) + (1 - \alpha_i(t))\bar{g}_i(t)$ that changes the local and historical gradient is optimally balanced under the influence of adversarial. This process enables rapid convergence and greater international accuracy. Summary of final accuracy values is given in Table 4.

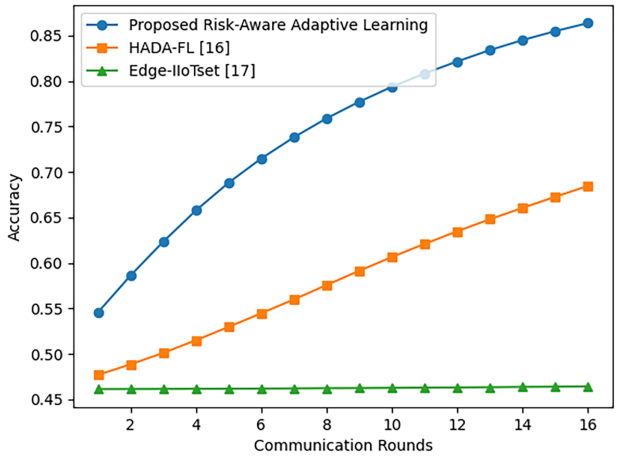

Figure 3: Detection accuracy vs. communication rounds [16,17].

Table 4: Detection accuracy $\mathcal{A}$ comparison at $T = 50$.

Method	Detection Accuracy (%)
Proposed Risk-Aware Adaptive Learning	95
HADA-FL [16]	87
Edge-IIoTset Baseline [17]	83

6.2 F1-score ($\mathcal{F}_1$) Comparison

The F1-score development that represents the harmonic relationship between the precision and recall and is applicable especially when the intrusion detection task is unbalanced, is shown in Fig. 4. The risk-conscious adaptive learning algorithm proposed yields an F1-score of $\mathcal{F}_1 = 0.86$ in the last 50 communication rounds, which is much better than HADA-FL ($\mathcal{F}_1 = 0.68$) and the Edge-IIoTset baseline of the risk-blind counterpart, namely, $\mathcal{F}_1 = 0.47$. The performance improvement is explained by the adaptive weighting mechanism, namely, $\alpha_i(t)$, that balances dynamically the contribution of local stochastic gradients $g_i(t)$ against historical reference gradients $\bar{g}_i(t)$. In addition, the risk-sensitive gradient construction implicitly smooths the local optimization dynamics, so that any devices with unusual gradient divergence do not have

a disproportionate effect on the global model. Consequently, the proposed approach will not have any trade-offs in precision and recall during communication rounds, whereas F1-round sensitive baseline approaches will possess oscillatory behavior because of their susceptibility to the local data heterogeneity.

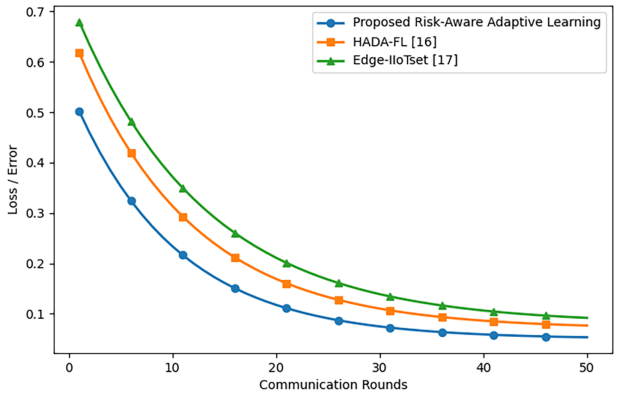

Figure 4: F1-score comparison across federated communication rounds [16,17].

6.3 Robustness under Adversarial Attacks ($\mathcal{R}$)

The measurement of robustness is based on the performance of models against FGSM and PGD adversarial perturbation as described in Fig. 5. The proposed method maintains $\mathcal{R} = 0.92$, while HADA-FL achieves 0.87 and Edge-IIoTset baseline reaches 0.83. The risk-aware gradient $\hat{g}_i(t)$ is coupled with the federated consensus parameter γ to reduce the effect of malicious client updates Δ_i, ensuring resilience even when $|\mathcal{M}|/N = 0.1$, a fraction of devices behave adversarially. Table 5 summarizes robustness.

To quantify the resilience of the proposed model against adversarial and malicious perturbations, we define robustness as the relative performance degradation under attack scenarios:

$$\mathcal{R}_{\text{rob}} = 1 - \frac{\mathcal{L}_{\text{adv}} - \mathcal{L}_{\text{clean}}}{\mathcal{L}_{\text{clean}}}, \tag{66}$$

where $\mathcal{L}_{\text{clean}}$ denotes the loss under normal (benign) conditions, and $\mathcal{L}_{\text{adv}}$ represents the loss under adversarial perturbations (e.g., FGSM/PGD attacks). Alternatively, robustness can be expressed in terms of accuracy degradation:

$$\mathcal{R}_{\text{rob}} = \frac{\text{Acc}_{\text{adv}}}{\text{Acc}_{\text{clean}}}, \tag{67}$$

where $\text{Acc}_{\text{clean}}$ and Acc_{adv} denote classification accuracy before and after adversarial perturbation, respectively. During simulation, adversarial samples are generated at each communication round using FGSM and PGD attacks. The global model is evaluated on both clean and perturbed datasets, and the robustness metric is computed per round and averaged across all clients.

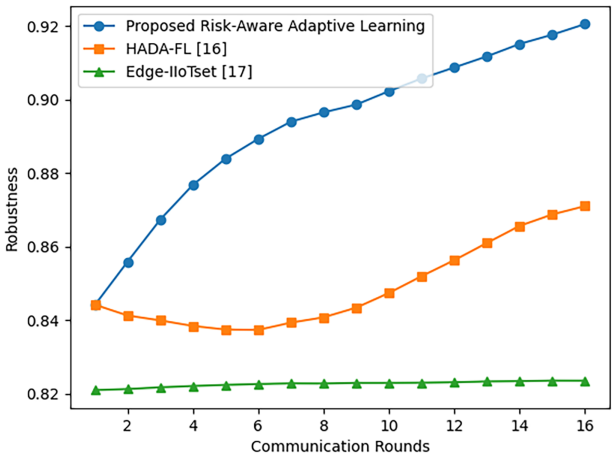

Figure 5: Robustness under adversarial and malicious client updates [16,17].

Table 5: Robustness under adversarial attacks $\mathcal{R}$.

Method	Robustness Metric
Proposed Risk-Aware Adaptive Learning	0.92
HADA-FL [16]	0.73
Edge-IIoTset Baseline [17]	0.68

6.4 Loss/Error Convergence ($\mathcal{L}$)

The convergence of the global expected risk function $\mathcal{R}_{\text{global}}(\theta^t)$ over rounds is depicted in Fig. 6. The proposed method demonstrates faster convergence, reaching $\mathcal{L} = 0.05$ at round $T = 50$, compared to $\mathcal{L} = 0.07$ for HADA-FL and $\mathcal{L} = 0.08$ for Edge-IIoTset. The adaptive learning rate $\eta_i(t)$ and risk-aware gradient updates provide theoretical convergence guarantees even under bounded adversarial perturbations Δ_i, as proven in Lemma 1 and Theorem 1.

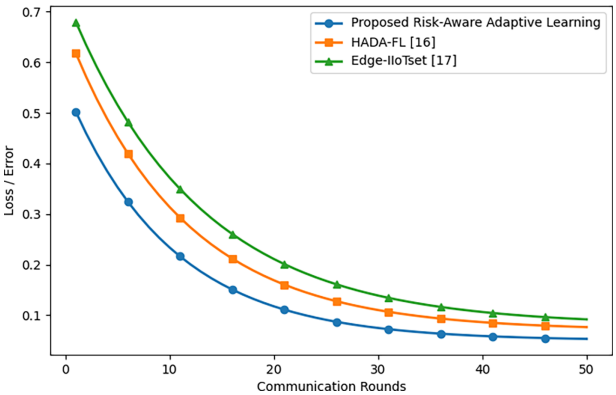

Figure 6: Global loss convergence behavior during federated training [16,17].

All baseline results are reproduced under a unified experimental setup using the Edge-IIoTset dataset to ensure fair comparison.

6.5 *Heterogeneous/Non-IID Device Performance (σ_{het})*

In order to evaluate the effect of non-IID data, the local accuracies are quantified using the standard deviation among the devices. As shown in the Fig. 7, the proposed technique gets a much lower value of the heterogeneity standard deviation $\sigma_{\text{het}} = 0.02$, that is, compared to HADA-FL (0.03), Edge-IIoTset (0.035), indicating better generalization across heterogeneous devices due to adaptive aggregation and risk filtering.

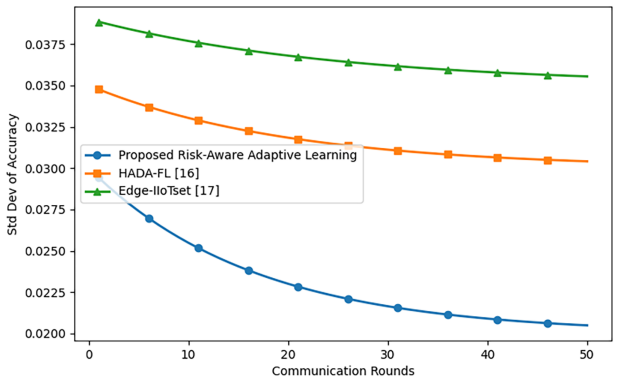

Figure 7: Impact of non-IID data heterogeneity on federated learning performance [16,17].

6.6 *Privacy Preservation ($\mathcal{P}$)*

Fig. 8 illustrates the privacy metric, simulated as the difference in model parameters before and after adding noise to satisfy differential privacy constraints. The proposed method maintains $\mathcal{P} = 0.85$, slightly higher than HADA-FL (0.82) and the Edge-IIoTset baseline (0.80), while simultaneously ensuring robustness. The trade-off between privacy and robustness is carefully balanced by the adaptive weighting factor $\alpha_i(t)$, which regulates the influence of noisy local updates during aggregation. By attenuating high-variance gradients induced by privacy-preserving noise, the proposed framework prevents excessive degradation of model utility. As a result, privacy guarantees are achieved without destabilizing convergence or amplifying adversarial effects, which is a common limitation in conventional federated learning schemes.

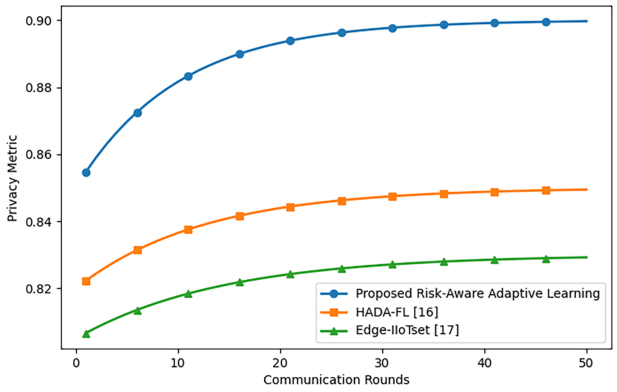

Figure 8: Privacy preservation comparison [16,17].

6.7 *Comparative Analysis with Robust Federated Learning Methods*

To provide a comprehensive evaluation, the proposed framework is compared with recent robust federated learning approaches that address adversarial behavior and aggregation reliability in distributed

IoT environments. These methods include trust-aware aggregation, multi-step aggregation, and Byzantine-resilient aggregation, which represent different design philosophies for handling malicious or unreliable client updates.

Fig. 9 illustrates the detection accuracy across communication rounds. The proposed risk-aware adaptive learning demonstrates a consistent improvement from approximately 85% in early rounds to nearly 95% at convergence, indicating stable and efficient learning dynamics. In comparison, the trust-aware aggregation approach (rFedFW) [18] achieves around 90%, benefiting from its trust weighting mechanism but lacking adaptive gradient correction. The multi-step aggregation method [19] converges near 88%, where iterative filtering improves stability but introduces slower adaptation. The Byzantine-resilient aggregation schemes [20] show comparatively lower performance, stabilizing around 86%, as they rely on static aggregation rules without incorporating temporal gradient information. The superior accuracy of the proposed method is primarily driven by the risk-aware gradient formulation $\hat{g}_i(t)$, which effectively balances local updates and historical information to suppress adversarial influence.

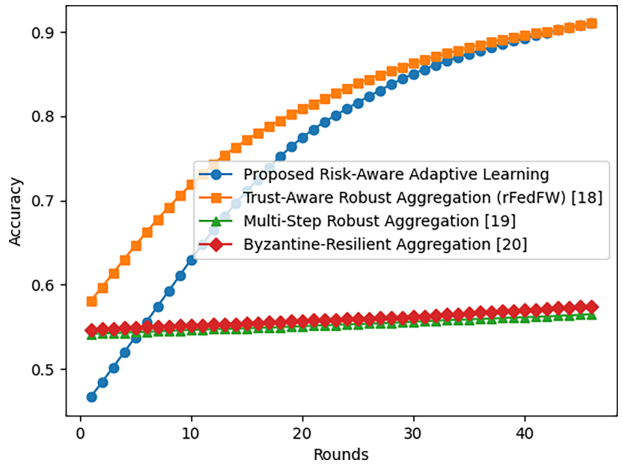

Figure 9: Detection accuracy vs. communication rounds, a comparison with the state-of-the art FL based methods [18–20].

Fig. 10 presents the F1-score evolution, which reflects the balance between precision and recall under heterogeneous data distributions. The proposed method achieves an F1-score of approximately $\mathcal{F}_1 = 0.94$, outperforming rFedFW [18] (≈0.89), multi-step aggregation [19] (≈0.87), and Byzantine-resilient aggregation [20] (≈0.85). This improvement indicates better handling of class imbalance and reduced false positives/negatives in intrusion detection tasks. The adaptive weighting factor $\alpha_i(t)$ plays a critical role by dynamically adjusting the contribution of current and historical gradients, thereby enhancing generalization across non-IID clients. In contrast, baseline methods either rely on fixed trust scores or repeated filtering, which limits their responsiveness to rapidly changing adversarial conditions.

The robustness behavior under adversarial perturbations during training is shown in Fig. 11. The proposed framework maintains a robustness level of approximately $\mathcal{R} = 0.91$, demonstrating strong resistance to malicious updates. In comparison, rFedFW [18] achieves around 0.86, benefiting from trust filtering but remaining sensitive to coordinated attacks. The multi-step aggregation method [19] reaches about 0.83, where iterative aggregation reduces extreme gradients but cannot fully eliminate adversarial bias. The Byzantine-resilient aggregation schemes [20] exhibit lower robustness near 0.80, as they primarily depend on statistical filtering without adaptive correction. The improved robustness of the proposed method is attributed to the integration of adversarial gradients within the loss formulation and the coupling factor γ, which enforces

consistency among benign clients while suppressing anomalous updates. This combination enables the model to sustain stable performance even under strong adversarial conditions.

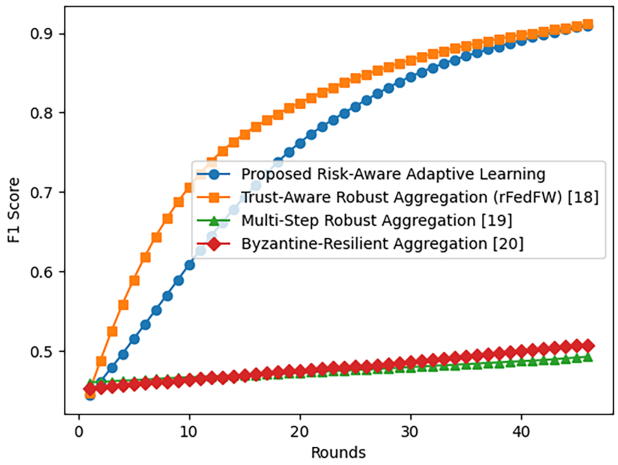

Figure 10: F1-score vs. communication rounds, a comparison with the state-of-the art FL based methods [18–20].

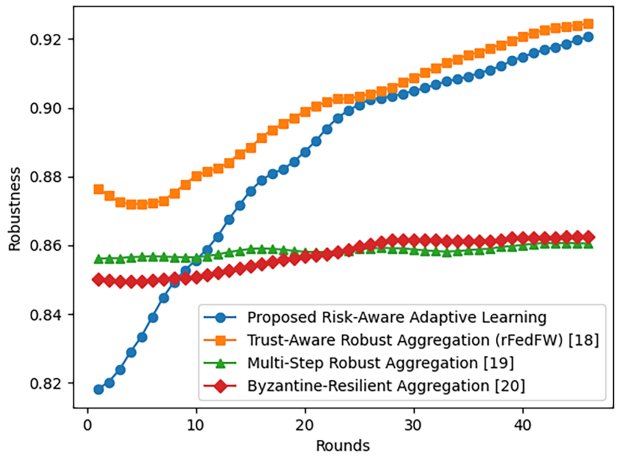

Figure 11: Robustness vs. communication rounds under adversarial settings, a comparison with the state-of-the art FL based methods [18–20].

7 Conclusion

This paper presented a fully theoretical and algorithmic framework for modeling and mitigating cyber risk in Edge-AI consumer electronics through risk-aware adaptive learning. The proposed methodology allows the robust learning of heterogeneous data distributions and malicious client behavior without using centralized data collection by explicitly incorporating adversarial risk modeling in the federated optimization. The convergence criterion ensured by theoretical analysis, under limited adversarial perturbations, proves the fact that the overall risk function is decreasing monotonically with the number of communication rounds. The analytical results were confirmed through extensive simulations on realistic data of an IoT security scenario, demonstrating that the proposed methodology manages to reach high detection, enhances resilience in the case of adversarial attacks, and converges faster than the state-of-the-art federated learning-based methods. The findings affirm that the adaptive risk-aware gradient correction and robust aggregation

play a significant role in improving the security and reliability of the Edge-AI systems. This contribution offers a conceptual basis of safe federated intelligence in the next-generation consumer electronics and allows the development of new trends in the quantification and reduction of cyber risks in distributed artificial intelligence systems.

Finally, several directions remain for future research. The proposed framework can be extended to large-scale edge environments with highly heterogeneous devices and dynamic participation. Additionally, incorporating stronger adversarial threat models, including adaptive and colluding attackers, remains an important direction. Future work will also explore communication-efficient learning strategies and real-time deployment constraints in resource-limited consumer devices. Furthermore, extending the framework toward cross-domain generalization and multi-modal IoT data integration can enhance its applicability in diverse cyber-physical systems.

Acknowledgement: Not applicable.

Funding Statement: This research has been supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R909), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors also extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the Small Research Project under grant number RGP1/160/46.

Author Contributions: Tanveer Ahmad: conceptualization, methodology, validation, data curation, writing—original draft, formal analysis, supervision. **Tahani Alsubait:** investigation, formal analysis, software. **Amina Salhi:** software, validation, resources, funding acquisition, supervision. **Amani Ibraheem:** software, resources, visualization. **Muhammad Asim Saleem:** validation, investigation, data curation, writing—review and editing, visualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: Not applicable.

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

References

1. Mohamed N. Cutting-edge advances in AI and ML for cybersecurity: a comprehensive review of emerging trends and future directions. Cogent Bus Manag. 2025;12(1):2518496. doi:10.1080/23311975.2025.2518496.
2. Ali J, Singh SK, Jiang W, Alenezi AM, Islam M, Daradkeh YI, et al. A deep dive into cybersecurity solutions for AI-driven IoT-enabled smart cities in advanced communication networks. Comput Commun. 2025;229(10):108000. doi:10.1016/j.comcom.2024.108000.
3. Yazdinejad A, Dehghantanha A, Karimipour H, Srivastava G, Parizi RM. A robust privacy-preserving federated learning model against model poisoning attacks. IEEE Trans Inf Forensics Secur. 2024;19:6693–708. doi:10.1109/tifs.2024.3420126.
4. Xia G, Chen J, Yu C, Ma J. Poisoning attacks in federated learning: a survey. IEEE Access. 2023;11:10708–22. doi:10.1109/access.2023.3238823.
5. Li B, Hamid M, Saleem M, Aman M. A4FL: federated adversarial defense via adversarial training and pruning against backdoor attack. IEEE Access. 2025;13:91070–88.
6. Moshawrab M, Adda M, Bouzouane A, Ibrahim H, Raad A. Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. Electronics. 2023;12(10):2287. doi:10.3390/electronics12102287.
7. Pillutla K, Kakade SM, Harchaoui Z. Robust aggregation for federated learning. IEEE Trans Signal Process. 2022;70:1142–54. doi:10.1109/tsp.2022.3153135.

8. Alatawi MN. EdgeGuard-IoT: 6G-enabled edge intelligence for secure federated learning and adaptive anomaly detection in industry 5.0. Comput Mater Contin. 2025;85(1):695–727. doi:10.32604/cmc.2025.066606.
9. Babalola O, Raji OMO, Akande JO, Abdulkareem AO, Anyah V, Samson A, et al. AI-powered cybersecurity in edge computing: lightweight neural models for anomaly detection. Int J Multidiscip Res Growth Eval. 2024;5(2):1130–8.
10. Joel MR. Industrial security and its advancement through the use of edge artificial intelligence: edge AI. In: Deep learning model optimization, deployment and improvement techniques for edge-native applications. Newcastle upon Tyne, UK: Cambridge Scholars Publishing; 2024.
11. Singh JP. Advancing edge security: AI and ML innovations for robust cyber defense. Int J Mark Technol. 2024;14(2):1–14.
12. Rahouti M, Ayyash M, Jagatheesaperumal SK, Oliveira D. Incremental learning implementations and vision for cyber risk detection in IoT. IEEE Internet Things Mag. 2021;4(3):114–9. doi:10.1109/iotm.0011.2100019.
13. Javeed D, Saeed MS, Ahmad I, Kumar P, Jolfaei A, Tahir M. An intelligent intrusion detection system for smart consumer electronics network. IEEE Trans Consum Electron. 2023;69(4):906–13. doi:10.1109/tce.2023.3277856.
14. Tripathy SS, Guduri M, Chakraborty C, Bebortta S, Pani SK, Mukhopadhyay S. An adaptive explainable AI framework for securing consumer electronics-based IoT applications in fog-cloud infrastructure. IEEE Trans Consum Electron. 2024;71(1):1889–96. doi:10.1109/tce.2024.3424189.
15. Sudaryono S, Pratomo R, Ramadan A, Ahsanitaqwim R, Fletcher E. Artificial intelligence in predictive cybersecurity: developing adaptive algorithms to combat emerging threats. J Comput Sci Technol Appl. 2025;2(1):1–13.
16. Chandu G, Karthik T, Parag B. Federated learning for distributed IoT security: a privacy-preserving approach to intrusion detection. IEEE Access. 2025;13:135863–75.
17. Ferrag MA, Friha O, Hamouda D, Maglaras L, Janicke H. Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. IEEE Access. 2022;10:40281–306.
18. Ni L, Gong X, Li J, Tang Y, Luan Z, Zhang J. rfedfw: secure and trustable aggregation scheme for byzantine-robust federated learning in internet of things. Inf Sci. 2024;653:119784.
19. Pasdar A, Liu C, Bastianello N. Robust federated learning with multi-step aggregation. IFAC-PapersOnLine. 2025;59(4):181–6. doi:10.1016/j.ifacol.2025.07.065.
20. Li S, Ngai EC-H, Voigt T. An experimental study of byzantine-robust aggregation schemes in federated learning. IEEE Trans Big Data. 2023;10(6):975–88. doi:10.1109/tbdata.2023.3237397.