applied sciences

MDPI

Article

Trust- and Energy-Aware Federated Learning for Robust and Sustainable IoT Edge Intelligence: Full-System Evaluation on PAMAP2 and Cross-Domain Robustness Assessment on CICIoT2023

Manuel J. C. S. Reis [1,2,*], Carlos Serôdio [1,3] and Frederico Branco [1,4]

[1] Engineering Department, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5000-801 Vila Real, Portugal; cserodio@utad.pt (C.S.); fbranco@utad.pt (F.B.)
[2] IEETA—Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
[3] ALGORITMI Research Centre, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
[4] INESC TEC, Campus da Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
* Correspondence: mcabral@utad.pt

Abstract

Federated learning (FL) supports collaborative edge intelligence without centralizing raw data, but non-IID data, Byzantine attacks, and limited client energy complicate IoT deployment. We study two complementary server-side mechanisms: TEA-FL, which couples consistency-based trust with residual energy in client selection and aggregation, and ARL-FL, which separates risk/lifetime-aware scheduling from pressure-adaptive robust aggregation. A matched two-domain protocol distinguishes full-system PAMAP2 evidence from aggregation-focused CICIoT2023 evidence and combines fold/seed-level inference with ablations, targeted sensitivities, and complete-round overhead measurements. Under the primary simulated resource model, TEA-FL achieved the highest overall PAMAP2 Macro-F1 and lowest normalized energy, but reduced participation fairness and did not complete every nominal-horizon run. ARL-FL preserved the largest active-client population and completed all PAMAP2 runs, but failed under the strongest sign-flip attack. Focused controls showed that trust drives TEA-FL's sign-flip robustness, energy-aware selection drives ARL-FL's sustainability benefit, and simpler fixed clipping-plus-trimming can be competitive under strong attack. The results support context-dependent robustness–sustainability trade-offs rather than universal superiority.

Keywords: federated learning; Byzantine robustness; trust-aware aggregation; energy-aware client selection; Internet of Things; non-IID data; human activity recognition; intrusion detection

Check for updates

Academic Editor: Gianluigi Ferrari

Received: 13 September 2026
Revised: 19 October 2026
Accepted: 3 November 2026
Published: 6 November 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) license.

1. Introduction

The Internet of Things (IoT) increasingly relies on distributed sensing and edge intelligence, but centralizing raw data can be impractical because of privacy, communication, ownership, and scale constraints. Federated learning (FL) addresses this by exchanging model updates rather than raw observations [1,2]. In resource-constrained IoT deployments, however, statistical heterogeneity, device/resource heterogeneity, and malicious updates interact: non-IID data can destabilize optimization, energy depletion can remove clients from the federation, and Byzantine updates can corrupt standard averaging [3–10].

Most prior studies emphasize one of these dimensions at a time. This work instead evaluates two complementary server-side strategies under explicit resource and threat models. TEA-FL directly couples consistency-based trust with residual energy in client selection and aggregation, whereas ARL-FL separates multi-signal risk/lifetime-aware scheduling from pressure-adaptive robust aggregation. PAMAP2 evaluates the complete selection–training–aggregation systems with simulated energy depletion and dropout; CICIoT2023 provides a deliberately narrower cross-domain assessment of the active trust/risk and aggregation components under frozen participation [11,12]. The evidence comprises 1020 core scientific runs plus 260 focused runs, for 1280 executions in total. These executions are not treated as 1280 independent inferential replicates: primary inference uses five held-out-subject blocks for PAMAP2 and five experimental-seed blocks conditional on one frozen CICIoT2023 split.

The study addresses five research questions:

- RQ1—Robustness: How do TEA-FL and ARL-FL compare with conventional and robust aggregation baselines under clean, label-flip, and sign-flip conditions?
- RQ2—Sustainability: What trade-offs arise among predictive performance, normalized energy consumption, residual energy, active-client retention, and participation fairness?
- RQ3—Mechanism attribution: Which trust, energy, and adaptive-pressure components are responsible for the observed predictive and operational effects?
- RQ4—Cross-domain robustness transfer: Under a common externally frozen participation schedule, do the active trust/risk updating and aggregation mechanisms exhibit qualitative robustness transfer from PAMAP2 to CICIoT2023?
- RQ5—Operational cost and scaling: What server-side computational and state overhead do the complete TEA-FL and ARL-FL round procedures introduce at the evaluated federation scale, and how does scheduling cost change as the eligible-client population grows?

The principal contributions are: (1) a two-mechanism server-side co-design in which TEA-FL couples peer-relative consistency trust and residual energy in both participation and aggregation, whereas ARL-FL separates persistent multi-signal risk, depletion pressure, and staleness during Pareto scheduling and uses a round-level pressure state to adapt robust clipping and trimming; (2) an explicit novelty boundary showing that the individual primitives—distance-based trust evidence, robust trimming/clipping, Pareto ranking, and resource-aware selection—are not claimed as new in isolation, while their coupling/separation into the two complete round procedures and their joint robustness–sustainability evaluation constitute the methodological contribution; (3) a matched two-domain evaluation that distinguishes full-system PAMAP2 evidence from frozen-participation CICIoT2023 component evidence; (4) multidimensional evaluation of predictive effectiveness, simulated energy, federation viability, retention, participation fairness, attack-time behavior, and controlled ablations/sensitivity analyses; and (5) a reproducible inferential and operational protocol based on fold-/seed-level matched units, exact paired tests, uncertainty intervals, explicit execution boundaries, and phase-resolved server-overhead measurements. Detailed run ledgers, hashes, dependency locks, and secondary audit artifacts are retained in the versioned archive rather than repeated in the main text.

2. Related Work

2.1. Federated Learning for Intelligent IoT Systems

Federated learning operates under forms of heterogeneity that are uncommon in conventional centralized machine learning. Statistical heterogeneity arises because each client

observes data generated by its own users, sensors, environments, or operational processes. Consequently, local datasets may differ in class proportions, feature distributions, sample sizes, noise levels, and temporal coverage. System heterogeneity arises simultaneously from differences in processing speed, available memory, communication quality, energy capacity, and client availability [2,3,6,13]. These sources of heterogeneity affect not only convergence but also which clients remain able to participate throughout training.

FedAvg reduces communication by allowing multiple local optimization steps before aggregation, but those steps can cause local models to move toward different local optima when client data are non-IID [1,3]. FedProx addresses this problem by adding a proximal term to each local objective, thereby restricting excessive deviation from the current global model [4]. This modification improves stability under heterogeneous data and variable amounts of local computation, although its performance remains dependent on the proximal coefficient and the severity of the client-distribution shift.

Several subsequent methods have addressed client drift and objective inconsistency more explicitly. SCAFFOLD employs server-side and client-side control variates to estimate and correct the drift between local and global optimization directions [14]. FedNova normalizes client updates according to the amount of local computation, preventing heterogeneous local workloads from causing convergence toward an inconsistent surrogate objective [15]. FedDyn instead introduces a dynamically updated regularization term that aligns local and global stationary solutions over successive communication rounds [16]. These methods provide important optimization mechanisms for benign statistical and computational heterogeneity. However, they do not by themselves determine whether a submitted update is malicious, nor do they directly manage the progressive depletion of energy-constrained clients.

Heterogeneity is also influenced by client selection. Selecting only a subset of devices per round reduces communication and computation, but the selected subset determines the direction and representativeness of the aggregated update. Random selection is simple and generally avoids explicit preference for particular clients, although it may repeatedly select slow, depleted, or low-utility devices. Conversely, performance-driven selection may accelerate convergence while creating systematic selection bias. Cho et al. showed that selection skew can improve short-term convergence when high-loss clients are prioritized but can also bias the optimized objective [17]. Thus, client-selection policies must be evaluated not only by final predictive performance but also by their effects on client representation, participation frequency, and long-term availability.

2.2. Byzantine-Robust Aggregation and Model-Poisoning Defenses

The aggregation server in conventional FL cannot directly inspect the private training data used to produce a client update. A malicious client may therefore submit an arbitrary update while remaining syntactically indistinguishable from an honest participant. Byzantine-robust aggregation attempts to limit the effect of such updates without requiring access to local data.

Krum and Multi-Krum select updates according to their distances from neighboring client updates, under the assumption that benign updates form a sufficiently coherent majority [7]. Coordinate-wise median and trimmed-mean aggregation reduce the influence of extreme parameter values independently in each model dimension [8]. Geometric-median-based robust federated aggregation instead searches for a central update in the full parameter space and can be implemented using smoothed iterative procedures [10]. These methods constitute important robust alternatives to arithmetic averaging and are frequently used as reference defenses.

Nevertheless, robust aggregation can become less reliable when benign clients are themselves highly heterogeneous. Under non-IID data, honest updates may be widely dispersed, while carefully designed malicious updates may remain within the apparent benign region. Model-poisoning attacks can exploit the aggregation rule, the expected fraction of malicious clients, and the geometry of benign updates to maximize global-model degradation [9]. This creates a fundamental ambiguity: update deviation may reflect either malicious behavior or legitimate statistical heterogeneity.

Trust-based defenses attempt to resolve this ambiguity by assigning reliability scores to clients or updates. FLTrust uses a small clean dataset at the server to compute a trusted reference update. Client updates receive lower trust scores when their directions deviate from the server reference, and their magnitudes are normalized before weighted aggregation [18]. This provides a concrete root of trust, but assumes that the server possesses a sufficiently representative clean dataset.

FoolsGold addresses coordinated poisoning by reducing the influence of clients whose historical update directions are unusually similar, based on the observation that Sybil clients controlled by the same adversary may exhibit correlated behavior [19]. Its detection signal is therefore particularly suited to coordinated attacks, but independent attackers or benign clients with similar data may not produce the same distinguishable pattern. FLAME combines clustering, clipping, and calibrated noise injection to mitigate federated backdoors while limiting deterioration of the main task [20]. Although highly relevant to secure FL, backdoor defenses target a different objective from untargeted label-flip or sign-flip attacks, for which the primary goal is often broad degradation of the global decision boundary.

These studies demonstrate that no single robust statistic or trust signal is sufficient under all threat models. Distance-based rules, coordinate-wise robust estimators, trusted server updates, temporal similarities, and clustering each rely on different assumptions about benign and malicious behavior. The present study therefore treats robustness as scenario-dependent and evaluates clean, label-flip, moderate sign-flip, and strong sign-flip conditions separately rather than assuming that a defense effective against one attack generalizes automatically to another.

The present experiments deliberately use deterministic label-flip and amplified sign-flip attacks as controlled untargeted stress tests. More adaptive adversaries may instead optimize poisoned updates against the aggregation rule, perform model replacement, coordinate across Sybil identities, or implant targeted backdoors while preserving benign-task performance. These attacks involve different adversarial objectives and success criteria and are therefore not assumed to be covered by the present empirical results. Evaluating TEA-FL and ARL-FL against such adaptive and targeted threat models constitutes an important extension of this work.

2.3. Trustworthiness Beyond Attack Detection

Trustworthy FL extends beyond the identification of poisoned updates. A trustworthy collaborative-learning process may also be expected to preserve privacy, fairness, transparency, reliability, accountability, and resilience to failures [21]. In resource-constrained IoT systems, trust is additionally linked to the probability that a selected client can complete local training and communication without exhausting its available resources.

A client may appear reliable from the perspective of model-update similarity while being close to dropout because of low residual energy. Conversely, a client with abundant energy may submit an inaccurate or malicious update. Trust and resource state therefore represent distinct dimensions. Combining them into a single unexamined score can obscure the reason why a client is selected, downweighted, or excluded.

Trust estimates may also be affected by non-IID data. Clients containing rare activities, uncommon attacks, or minority classes can generate updates that differ from the majority while remaining scientifically valuable. Aggressive anomaly filtering may consequently remove informative clients and improve apparent update consistency at the cost of data coverage. This risk motivates the joint reporting of predictive performance, active-client retention, and participation fairness, rather than assessing a trust mechanism only through final accuracy.

The proposed TEA-FL and ARL-FL mechanisms are positioned within this broader interpretation of trustworthiness. They do not assume that update reliability and energy availability are equivalent. Instead, the study evaluates how these signals interact and uses controlled ablations to determine whether observed benefits arise from trust, energy awareness, or adaptive participation control.

2.4. Resource-Aware and Energy-Aware Client Selection

Resource-aware FL research has proposed several strategies for selecting clients according to their expected training utility and execution capacity. FedCS estimates client-resource conditions and selects a feasible subset that can complete local computation and communication within a round deadline [22]. Oort prioritizes clients according to both statistical utility and system performance, improving time-to-accuracy by directing training toward informative clients that are also capable of completing the assigned work efficiently [23].

Tier-based approaches group devices according to their observed training performance. TiFL divides clients into resource tiers and selects participants from particular tiers, with adaptive tier selection intended to reduce the effect of stragglers while preserving model quality [24]. Such methods improve system efficiency, but tiering and utility-driven selection may repeatedly favor fast or high-resource clients unless participation balance is explicitly considered.

Energy-aware approaches focus more directly on battery-constrained IoT devices. FedLE clusters clients according to similarities in their partially trained model representations and reduces the probability of selecting clients with low residual energy, with the objective of extending the operational lifespan of the edge network [25]. This is particularly relevant when client dropout causes not only a reduction in the number of participants but also the permanent loss of locally held training distributions. The present work includes a FedLE-adapted baseline to represent this family of lifespan-oriented selection methods under the same data partitions, attack schedules, energy accounting, and predictive model used for the remaining methods.

Energy-efficient selection can nevertheless introduce a trade-off. Protecting low-energy clients by selecting them less frequently may extend their lifetime, but it may also reduce the contribution of clients holding rare data. Conversely, repeatedly selecting high-energy clients can create participation inequality and concentrate the learned model on a narrower portion of the federation. The q-Fair Federated Learning objective and its q-FedAvg optimization method were introduced to reduce disparities in client-level performance by placing greater emphasis on clients with higher losses [26]. Although this addresses model-performance fairness rather than battery fairness, it demonstrates that aggregate accuracy alone does not characterize how the benefits and burdens of FL are distributed across clients.

For this reason, energy-aware FL should be evaluated through multiple operational outcomes. Total energy consumption measures the aggregate resource burden, whereas residual energy and dropout indicate whether clients remain available. Active-client lifetime and final active-client count characterize federation continuity, while participation-frequency metrics reveal whether the training process depends disproportionately on a

small set of devices. Jain's fairness index provides a normalized measure of equality in accumulated client participation and is therefore used in this study alongside predictive and energy metrics.

2.5. Research Gap and Positioning of the Present Study

The reviewed literature provides strong individual solutions for heterogeneous optimization, Byzantine-robust aggregation, trust bootstrapping, client utility estimation, straggler mitigation, energy-aware selection, and fairness-aware objectives. However, these dimensions are often studied independently. Heterogeneity-oriented optimization methods generally assume benign clients; robust aggregation methods commonly abstract away battery depletion and long-term client availability; and energy-aware selection methods may not examine adversarial manipulation of the selected updates [2,6,21].

A second limitation concerns evaluation scope. Many studies emphasize final accuracy, convergence speed, or attack success rate, while fewer jointly examine balanced predictive performance, energy consumption, residual energy, client dropout, participation lifetime, and fairness. Final-round metrics also provide limited insight into the mechanism of failure or recovery. In attacks that begin after a benign warm-up period, it is important to examine the trajectory before and after attack onset, the lowest post-attack performance, and whether the global model subsequently recovers.

A third gap is the limited use of controlled mechanism attribution. When a method combines trust, energy, and adaptive participation criteria, an overall improvement does not establish which component produced it. Without ablation, a predictive gain may be incorrectly attributed to energy awareness when it is actually caused by trust-based rejection, or an energy saving may be attributed to trust when it results from reduced client participation.

Finally, conclusions obtained from one task may not generalize to another. Wearable activity recognition involves subject-specific sensor distributions and multiclass physical activities, whereas IoT intrusion detection involves network-flow features, severe class imbalance, and different attack semantics. Cross-domain evaluation is therefore necessary to distinguish mechanism-level regularities from dataset-specific effects.

The novelty boundary is therefore architectural and evaluative rather than a claim that the underlying primitives are individually new. Relative to FLTrust, TEA-FL does not assume a trusted server-side root dataset; its reliability signal is peer-relative consistency, and residual energy affects both participation and aggregation rather than trust-weighting alone [18]. Relative to Krum, coordinate-wise median/trimmed mean, and geometric-median aggregation, the proposed systems do not rely only on a fixed robust estimator: TEA-FL changes both who participates and how selected updates are weighted, while ARL-FL couples persistent risk/lifetime-aware scheduling to pressure-dependent clipping and trimming [7,8,10]. Relative to FedCS, Oort, TiFL, and FedLE, resource state is not used only for feasibility, utility, tiering, or lifespan extension; it is evaluated jointly with adversarial reliability, participation staleness, retention, and fairness [22–25]. Conversely, FoolsGold and FLAME address coordinated Sybil-like behavior or backdoor-specific defenses that are not reproduced here; the present threat model remains deliberately narrower [19,20].

Within that boundary, the present study contributes two contrasting complete round procedures rather than a single monolithic defense. TEA-FL deliberately scalarizes trust and residual energy for selection and reuses both states in aggregation. ARL-FL deliberately keeps client risk, predicted depletion pressure, and staleness as separate Pareto objectives, then converts the selected clients' risk distribution into a global pressure state that controls robust aggregation strength without direct risk or energy weighting. The experiments compare these designs with FedAvg, FedProx, Random Trimmed Mean, and FedLE-adapted

under matched non-IID partitions and attack conditions, and use ablations, fixed clipping-plus-trimming controls, mechanism diagnostics, and sensitivity analyses to determine which gains come from trust, resource-aware scheduling, or robust aggregation. PAMAP2 evaluates the complete selection–training–aggregation systems, whereas CICIoT2023 isolates the active trust/risk and aggregation components under frozen participation. The two campaigns are used to identify complementary, context-dependent strengths and explicit failure modes rather than to claim universal superiority for either method.

To consolidate the main research directions discussed above and clarify the positioning of the present study, Table 1 summarizes the objectives, representative approaches, and principal limitations of the most closely related federated-learning research streams.

Table 1. Summary of the main federated-learning research streams related to the present study.

Research Stream	Representative Approaches	Primary Objective	Main Limitations in the Context of Resource-Constrained and Adversarial IoT	Positioning of the Present Study
Federated optimization under statistical heterogeneity	FedProx [4], SCAFFOLD [14], FedNova [15], and FedDyn [16]	Reduce client drift and improve convergence under non-IID data and heterogeneous local computation	Generally, assumes benign client updates; does not directly address energy depletion, malicious participation, or long-term client retention	Evaluates statistical heterogeneity jointly with Byzantine attacks, energy consumption, client dropout, and participation fairness
Byzantine-robust aggregation	Krum [7], coordinate-wise median and trimmed mean [8], robust geometric-median aggregation [10]	Limit the influence of anomalous or malicious model updates	Robustness depends on assumptions regarding the malicious-client fraction and benign-update geometry; non-IID updates may be confused with malicious behavior	Compares conventional and robust aggregation under matched clean, label-flip, and sign-flip conditions and explicitly examines attack-dependent failure modes
Trust-aware and poisoning-resistant FL	FLTrust [18], FoolsGold [19], and FLAME [20]	Estimate update reliability or identify coordinated and backdoor attacks	May require a clean server dataset or attack-specific assumptions; typically, does not incorporate residual energy or client-lifetime constraints	Integrates trust signals with resource state and uses ablation to distinguish trust-driven robustness from energy-driven operational effects
Resource-aware client selection	FedCS [22], Oort [23], and TiFL [24]	Reduce latency and select clients capable of completing local training	May repeatedly favor faster or higher-resource devices, potentially reducing representativeness and participation equality	Assesses client selection through predictive performance, energy consumption, active-client lifetime, final client retention, and participation fairness
Energy- and lifespan-aware FL	FedLE [25]	Extend edge-network lifespan by reducing the participation of energy-constrained clients	Primarily focuses on resource preservation and does not explicitly address Byzantine manipulation or distinguish energy effects from trust effects	Uses a FedLE-adapted baseline and evaluates energy-aware participation under both benign and adversarial conditions
Fairness-aware federated optimization	q-FedAvg [26] and biased-client-selection analyses [17]	Reduce disparities in client-level performance and characterize selection bias	Predictive fairness and participation fairness are not equivalent; energy depletion and adversarial reliability are generally not modeled jointly	Reports Jain's participation fairness together with energy, retention, and predictive metrics in a unified multi-objective analysis
Present study	TEA-FL and ARL-FL	Co-design reliability, resource state, participation history, and aggregation strength within two contrasting complete round procedures	Individual trust, robust-estimation, Pareto-ranking, and resource-selection primitives are not claimed as new; scope remains limited by the stated threat and simulated-resource models	TEA-FL couples consistency trust and energy in selection and aggregation; ARL-FL separates risk/depletion/staleness Pareto scheduling from pressure-adaptive clipping/trimming, with ablation and focused controls used for mechanism attribution

3. Proposed Trust- and Energy-Aware Federated Learning Framework

This section presents the system model, threat model, resource model, and the two complementary federated-learning mechanisms investigated in this work: Trust- and Energy-Aware Federated Learning (TEA-FL) and Adaptive Risk- and Lifetime-Aware Fed-

erated Learning (ARL-FL). Both methods preserve the standard round-based FL workflow and do not require clients to transmit raw data. Their additional operations are performed primarily at the server and use information already available during conventional federated optimization, namely the submitted model updates, local sample counts, participation history, and simulated residual-energy states.

TEA-FL integrates consistency-based trust and residual energy directly into client selection and aggregation. ARL-FL follows a separation principle: risk, lifetime pressure, and participation staleness determine client scheduling, whereas aggregation uses global adversarial pressure to adapt norm clipping and coordinate-wise trimming without directly weighting updates by client risk or residual energy.

Figure 1 provides a conceptual overview of the proposed framework.

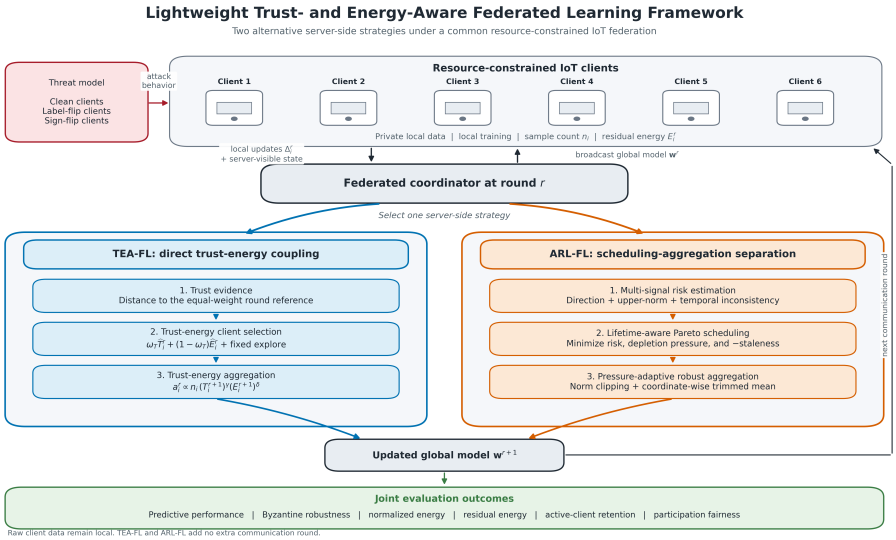


Figure 1. Overall architecture of the proposed trust- and energy-aware federated-learning framework. TEA-FL propagates trust and residual-energy information through client selection and aggregation, whereas ARL-FL separates multi-objective client scheduling from pressure-adaptive robust aggregation.

Notation and numerical safeguards. Throughout this section, bold lowercase symbols denote vectors, non-bold symbols denote scalars, and calligraphic uppercase symbols denote sets. The Euclidean norm and inner product are denoted by $\| \cdot \|_2$ and $\langle \cdot, \cdot \rangle$, respectively. The clipping operator is defined as

$$\mathrm{clip}(x, a, b) = \min\{b, \max\{a, x\}\}.$$

A numerical safeguard

$$\varepsilon = 10^{-12}$$

is used to prevent undefined logarithms and divisions by zero. This numerical constant is distinct from the stability parameter used by the Adam optimizer. For scalar arguments, median$(\cdot)$ denotes the conventional sample median. For vector arguments, the median is evaluated independently for each coordinate, as specified in Section 3.5.1. The notation $Q_{0.75}(\cdot)$ denotes the empirical 75th percentile computed using linear interpolation between adjacent order statistics.

To reproduce the numerical behavior of the frozen implementation, the stabilized cosine similarity is defined as

$$\cos_\varepsilon(a, b) = \begin{cases} \mathrm{clip}\left(\frac{\langle a,b\rangle}{\|a\|_2 \|b\|_2}, -1, 1\right), & \| a \|_2 > \varepsilon \; and \; \| b \|_2 > \varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

3.1. Federated System and Learning Model

Consider a federation containing N clients,

$$\mathcal{C} = \{1, \ldots, N\},$$

coordinated by a central server. Client i holds a private local dataset

$$\mathcal{D}_i = \left\{ \left(x_{i,j}, y_{i,j} \right) \right\}_{j=1}^{n_i},$$

where $n_i = |\mathcal{D}_i|$ is the number of local training examples. The local empirical objective of client i is

$$F_i(w) = \frac{1}{n_i} \sum_{j=1}^{n_i} l\left(w; x_{i,j}, y_{i,j} \right),$$

where w denotes the model parameters and $l(\cdot)$ is the task-specific loss function.

The conventional global objective is

$$\min_w F(w) = \sum_{i=1}^{N} \frac{n_i}{\sum_{j=1}^{N} n_j} F_i(w).$$

At communication round r, the server holds the current global model w^r and selects a subset of clients

$$\mathcal{S}_r \subseteq \mathcal{C}, \qquad |\mathcal{S}_r| = K.$$

The server broadcasts w^r to the selected clients. Each client performs local optimization on its private dataset and returns an updated local model w_i^{r+1}. The corresponding local update is defined as

$$\Delta_i^r = w_i^{r+1} - w^r.$$

In standard FedAvg, the server aggregates these updates according to the clients' local sample counts [1]:

$$w^{r+1} = w^r + \sum_{i \in \mathcal{S}_r} \frac{n_i}{\sum_{j \in \mathcal{S}_r} n_j} \Delta_i^r.$$

The proposed methods modify the client-selection, state-estimation, or aggregation stages while preserving this general communication structure.

The client data are allowed to be statistically heterogeneous. In particular, local class distributions may differ considerably across clients, and local sample counts need not be equal. No assumption is made that the local objectives F_i share the same optimum. This setting captures the subject-specific sensing patterns of wearable activity recognition and the heterogeneous traffic distributions of IoT intrusion detection.

3.2. Operational and Threat Model

The server is assumed to execute the prescribed protocol correctly and to maintain the client-level state required by TEA-FL or ARL-FL. The server does not access the clients' raw training data. The present work does not address a malicious or compromised central server, secure aggregation, inference attacks against submitted model updates, or cryptographic identity management.

Clients may exhibit three types of behavior:

1. Benign heterogeneous behavior: clients train normally, but their updates may differ because their local data are non-IID.

2. Label-flip behavior: malicious clients modify their local labels before training. For a task with C classes, the deterministic label-flip transformation is

$$y' = (y + 1) \bmod C.$$

3. Sign-flip behavior: malicious clients first compute a normal local update and then reverse and amplify it:

$$\widetilde{\Delta}_i^r = -\lambda \Delta_i^r, \quad (1)$$

where $\lambda > 1$ controls the attack magnitude. In the frozen primary experiments, $\lambda = 5$. A focused PAMAP2 sensitivity analysis evaluated $\lambda \in \{2.5, 5.0, 7.5\}$ while retaining the primary campaign configuration unchanged.

The set of malicious identities is fixed within each matched experimental condition. The attacks begin only after a clean warm-up period, allowing the temporal response of each method to be examined before and after attack onset. The evaluated threat model does not include Sybil identities, a malicious majority, adaptive adversaries with complete knowledge of the defense, stealth attacks optimized specifically against the proposed scores, or colluding clients that jointly solve an explicit defense-evasion problem.

TEA-FL should therefore be interpreted as a consistency-aware mechanism rather than as a formally Byzantine-resilient protocol. ARL-FL includes an explicit robust-aggregation stage, but it likewise does not provide a universal Byzantine-resilience guarantee. The experimental analysis consequently evaluates robustness empirically and reports the strong-attack conditions under which the methods fail.

3.3. Common Client-Energy Model

Each client i has a normalized residual-energy state E_i^r. These values represent comparable simulated resource units and must not be interpreted as measurements in physical joules.

At initialization,

$$E_i^0 \sim \mathcal{U}(E_{\min}, E_{\max}).$$

Each client also receives a fixed device-compute factor q_i, representing differences in computational efficiency:

$$q_i \sim \mathcal{U}(q_{\min}, q_{\max}).$$

An active client incurs a standby cost c_{st} during each communication round. A selected client additionally incurs a participation cost composed of communication and computation terms:

$$c_i = c_{\text{comm}} + c_{\text{comp}} \times q_i \frac{n_i}{\text{median}_{j \in \mathcal{C}}(n_j)}.$$

The residual-energy update is therefore

$$E_i^{r+1} = \max[0, E_i^r - c_{\text{st}} I(E_i^r \geq E_{\text{crit}}) - c_i I(i \in \mathcal{S}_r)],$$

where $I(\cdot)$ is the indicator function, and E_{crit} is the critical energy threshold.

A client is eligible for participation if

$$E_i^r \geq E_{\text{crit}}.$$

The eligible client set is

$$\mathcal{C}_r^{\text{elig}} = \{i \in \mathcal{C} : E_i^r \geq E_{\text{crit}}\}.$$

A client whose energy falls below the threshold is considered inactive and is not selected in subsequent rounds. The federation is considered operational while a predefined minimum fraction of clients remains active.

The same initial-energy profiles, device-compute factors, and cost equations are used across methods within each matched experimental condition. This prevents a method from receiving an artificial advantage through more favorable client-energy assignments.

In the frozen primary PAMAP2 configuration, $c_{\text{st}} = 0.0005$, $c_{\text{comm}} = 0.004$, $c_{\text{comp}} = 0.016$, $E_{\text{crit}} = 0.10$, and the federation-viability stop required at least 50% of the original clients to remain active. Because the computation term scales with local sample count and because residual energy enters TEA-FL and ARL-FL scheduling, these parameters can change the learning trajectory rather than merely rescale an accounting total. A matched resource-model sensitivity therefore reran representative conditions with a communication-dominant cost mix and alternative eligibility thresholds; the primary resource model itself remained unchanged.

3.4. Trust- and Energy-Aware Federated Learning

TEA-FL assigns two server-maintained state variables to each client:

- A trust score $T_i^r \in [T_{\min}, 1]$, representing geometric consistency with the current collaborative update pattern;
- A residual-energy state E_i^r, representing the client's continued capacity to participate.

Trust and energy are used both to determine participation and to modulate the influence of the selected local updates.

3.4.1. Trust Evidence

All clients are initialized with the same neutral trust value:

$$T_i^0 = T_0.$$

Following local training, each selected local model is flattened into a vector containing all floating-point model parameters. Let this vector be denoted by v_i^r. TEA-FL constructs an equal-weight round reference from the selected local models:

$$\overline{v}^r = \frac{1}{K} \sum_{j \in \mathcal{S}_r} v_j^r.$$

The distance between client i and the round reference is

$$d_i^r = ||\mathbf{v}_i^r - -\overline{\mathbf{v}}^r||_2.$$

The distances are normalized within the selected-client set. When the maximum and minimum distances differ, the current trust evidence is

$$e_i^r = \text{clip}\left(1 - \frac{d_i^r - d_{\min}^r}{d_{\max}^r - d_{\min}^r}, T_{\min}, 1\right),$$

where

$$d_{\min}^r = \min_{j \in \mathcal{S}_r} d_j^r,$$

and

$$d_{\max}^r = \max_{j \in \mathcal{S}_r} d_j^r.$$

When all selected clients have the same distance, $e_i^r = 1$ for every selected client.

The trust state is updated through an exponential moving average:

$$T_i^{r+1} = \mathrm{clip}[\beta_T T_i^r + (1-\beta_T) e_i^r, T_{\min}, 1], \qquad i \in \mathcal{S}_r.$$

The trust values of non-selected clients remain unchanged. The moving average prevents a single round from fully determining a client's long-term influence, while still allowing repeatedly inconsistent behavior to reduce its trust.

The trust score is a relative consistency proxy. It does not establish whether a client is malicious, and an honest client with an unusual data distribution may receive lower evidence if its update differs considerably from those of the remaining selected clients.

3.4.2. Joint Trust–Energy Client Selection

At the beginning of each round, trust and residual energy are independently min–max normalized over the eligible clients:

$$\widehat{T}_i^r = \frac{T_i^r - \min_{j \in \mathcal{C}_r^{\mathrm{elig}}} T_j^r}{\max_{j \in \mathcal{C}_r^{\mathrm{elig}}} T_j^r - \min_{j \in \mathcal{C}_r^{\mathrm{elig}}} T_j^r}$$

and

$$\widehat{E}_i^r = \frac{E_i^r - \min_{j \in \mathcal{C}_r^{\mathrm{elig}}} E_j^r}{\max_{j \in \mathcal{C}_r^{\mathrm{elig}}} E_j^r - \min_{j \in \mathcal{C}_r^{\mathrm{elig}}} E_j^r}$$

If all eligible values are identical, their normalized values are set to one.

The joint TEA-FL selection score is

$$s_i^r = \omega_T \widehat{T}_i^r + (1-\omega_T)\widehat{E}_i^r, \tag{2}$$

where $\omega_T \in [0,\ 1]$ controls the relative importance of trust and energy.

Purely deterministic selection could repeatedly favor the same clients. TEA-FL therefore uses an exploit–explore policy. Let

$$K_{\mathrm{exp}} = \max[1, \lceil \rho K \rceil],$$

be the number of exploration positions, where ρ is the exploration ratio. The remaining

$$K_{\mathrm{core}} = K - K_{\mathrm{exp}}$$

positions are assigned to the highest-ranked eligible clients according to s_i^r. The exploration clients are sampled uniformly without replacement from the remaining eligible pool.

The exploration component allows clients with temporarily lower trust or energy scores to remain observable and prevents complete lock-in to a fixed client subset.

3.4.3. Trust- and Energy-Aware Aggregation

After local training, trust updating, and deduction of the current participation cost, TEA-FL assigns the following unnormalized aggregation weight to client i:

$$\widetilde{a}_i^r = n_i \left(T_i^{r+1}\right)^{\gamma} \left(E_i^{r+1}\right)^{\delta},$$

where $\gamma \geq 0$ and $\delta \geq 0$ determine the influence of trust and residual energy.

The normalized weight is

$$a_i^r = \frac{\widetilde{a}_i^r}{\sum_{j \in \mathcal{S}_r} \widetilde{a}_j^r}. \tag{3}$$

The global model becomes

$$w^{r+1} = w^r + \sum_{i \in \mathcal{S}_r} a_i^r \Delta_i^r. \tag{4}$$

Setting $\gamma = \delta = 0$ reduces Equation (3) to the sample-size weighting used by FedAvg. Increasing γ concentrates aggregation weight on clients with greater consistency-based trust, whereas increasing δ favors clients with greater residual energy.

The frozen TEA-FL configuration uses

$$T_0 = 0.5, \quad T_{\min} = 0.05, \quad \beta_T = 0.8, \quad \omega_T = 0.8, \quad \rho = 0.1, \quad \gamma = 2.0, \quad \delta = 0.3.$$

With $K = 8$, the selection policy therefore contains seven exploitation positions and one exploration position per round.

3.4.4. TEA-FL Round Procedure

One TEA-FL communication round consists of the following operations:

1. Deduct the standby energy cost from all currently active clients.
2. Construct the eligible set using E_{crit}.
3. Normalize the eligible clients' trust and residual-energy values.
4. Compute the joint trust–energy selection score in Equation (2).
5. Select K_{core} exploitation clients and K_{exp} exploration clients.
6. Broadcast the current global model.
7. Perform local training on the selected clients.
8. Construct the equal-weight round reference.
9. Compute the current trust evidence and update the selected clients' trust states.
10. Deduct the communication and computation costs of participation.
11. Compute the trust- and energy-aware aggregation weights.
12. Update the global model using Equation (4).

3.5. Adaptive Risk- and Lifetime-Aware Federated Learning

ARL-FL was designed to avoid collapsing reliability, energy, and participation history into a single fixed scalar score. It separates the protocol into three stages:

1. Multi-signal client-risk estimation;
2. Lifetime-aware Pareto scheduling;
3. Pressure-adaptive robust aggregation.

Client risk and residual energy influence scheduling, but they are not used as direct aggregation weights. This separation prevents low-energy or high-risk scores from arbitrarily rescaling individual model updates during the robust-aggregation stage.

ARL-FL state initialization. At the beginning of each ARL-FL run, all clients were assigned a zero prior-risk state,

$$R_i^0 = 0, i \in C,$$

and no previous client-update vector was stored. The previous global adversarial pressure was initialized as

$$P^0 = 0,$$

while the last-participation state was initialized as

$$r_i^{\text{last}} = 0.$$

Since communication rounds are indexed from $r = 1$, all clients initially have a staleness value of one. The initial pressure $P^0 = 0$ also assigns three positions to random exploration during the first ARL-FL scheduling decision. Following a client's first participation, its most recent submitted update and last-participation round are stored and used in subsequent temporal-risk and staleness calculations.

The last-participation and staleness states affected client scheduling only in PAMAP2. In CICIoT2023, participation was externally frozen, although the risk, previous-update, and global-pressure states remained active for robust aggregation.

3.5.1. Multi-Signal Risk Estimation

For each selected client, ARL-FL computes the flattened model update Δ_i^r.

A coordinate-wise lower-median update is first constructed. For each model coordinate d, let

$$\Delta^r_{(1),d} \le \Delta^r_{(2),d} \le \cdots \le \Delta^r_{(K),d}$$

denote the ordered client-update values at that coordinate. The coordinate-wise reference is defined as

$$m_d^r = \Delta^r_{\left(\left\lfloor \frac{K+1}{2} \right\rfloor\right),d}, d = 1, \ldots, D,$$

and

$$m^r = (m_1^r, \ldots, m_D^r)^T.$$

Thus, when K is even, the lower of the two central coordinate values is used, consistently with the frozen implementation.

Direction anomaly. The directional alignment between client i's update and the coordinate-wise reference is:

$$c^r_{i,\mathrm{dir}} = \cos_\epsilon(\Delta_i^r, m^r).$$

The corresponding direction-anomaly score is

$$A_i^r = \mathrm{clip}\left[\frac{1 - c^r_{i,\mathrm{dir}}}{2}, 0, 1\right].$$

Upper-norm anomaly. The logarithm of the update norm is defined as

$$u_i^r = \log(\| \Delta_i^r \|_2 + \varepsilon).$$

The scalar median log norm and its median absolute deviation are respectively defined as

$$\widetilde{u}^r = \mathrm{median}_{\mathrm{j}\in \mathrm{S_r}} u_j^r$$

and

$$\mathrm{MAD}^{\mathrm{r}}_{\mathrm{u}} = \mathrm{median}_{\mathrm{j}\in \mathrm{S_r}} \left| u_j^r - \widetilde{u}^r \right|.$$

The factor 1.4826 gives the conventional normal-consistency scaling of the median absolute deviation [27]. Only unusually large update norms are penalized:

$$z^r_{i,\mathrm{norm}} = \max\left[0, \frac{u_i^r - \widetilde{u}^r}{1.4826\mathrm{MAD}^{\mathrm{r}}_{\mathrm{u}} + \varepsilon}\right].$$

This robust deviation is transformed to a bounded anomaly score:

$$N_i^r = \mathrm{clip}\left[1 - \exp\left(-\frac{z^r_{i,\mathrm{norm}}}{3}\right), 0, 1\right].$$

Temporal inconsistency. If client i has participated previously, the temporal alignment between its current update and its most recently submitted update is

$$c_{i,\text{temp}}^{r} = \cos_{\epsilon}\left(\Delta_i^r, \Delta_i^{\text{prev}}\right).$$

The temporal-inconsistency score is

$$H_i^r = \text{clip}\left[\frac{1 - c_{i,\text{temp}}^{r}}{2}, 0, 1\right].$$

For a client's first participation, $H_i^r = 0$, preventing the absence of a historical update from being treated as suspicious.

The instantaneous risk is the median of the three signals:

$$\zeta_i^r = \text{median}\ (A_i^r, N_i^r, H_i^r).$$

Median fusion ensures that one isolated high anomaly signal does not fully determine the client risk. At least two of the three signals must indicate substantial inconsistency before the fused value becomes large.

The persistent client-risk state is updated by

$$R_i^{r+1} = \text{clip}\ [\beta_R R_i^r + (1 - \beta_R)\zeta_i^r, 0, 1],$$

with $\beta_R = 0.8$.

The round-level global adversarial pressure is the 75th percentile of the updated risks among the selected clients:

$$P^r = Q_{0.75}\left(\left\{R_i^{r+1} : i \in \mathcal{S}_r\right\}\right).$$

Here, $Q_{0.75}$ is the empirical 75th percentile of the updated selected-client risks, computed using linear interpolation between adjacent order statistics. The resulting pressure value is clipped to the interval [0, 1].

3.5.2. Lifetime-Aware Pareto Scheduling

For each eligible client, ARL-FL estimates a lifetime-depletion pressure:

$$L_i^r = \frac{c_i}{\max\left(E_i^r, \varepsilon\right)},$$

where c_i is the predicted participation cost defined in Section 3.3. A client with low residual energy or a high expected participation cost therefore receives a higher depletion-pressure value.

ARL-FL also calculates participation staleness:

$$S_i^r = \min\left[S_{\max}, \max\left(0, r - r_i^{\text{last}}\right)\right],$$

where r_i^{last} is the last round in which client i was selected and $S_{\max}$ a staleness cap. In the frozen protocol, $S_{\max} = 10$.

Each eligible client is represented by the three-objective vector

$$o_i^r = (R_i^r, L_i^r, -S_i^r).$$

All three components are minimized. The negative sign on staleness means that clients absent for longer periods are preferred, all else being equal.

Clients are ordered using non-dominated sorting followed by crowding-distance ordering, following the core Pareto-ranking principles introduced in NSGA-II [28]. This preserves explicit trade-offs among risk, depletion pressure, and participation staleness without imposing a fixed weighted sum.

The number of random exploration positions depends on the pressure from the preceding completed round:

$$K_{\text{exp}}\left(P^{r-1}\right) = \begin{cases} 3, & P^{r-1} < 0.25, \\ 2, & 0.25 \le P^{r-1} < 0.50, \\ 1, & P^{r-1} \ge 0.50. \end{cases}$$

The number of Pareto-core clients is

$$K_{\text{core}} = K - K_{\text{exp}}.$$

When pressure is low, broader exploration improves participation coverage and allows the system to update stale risk estimates. As pressure increases, ARL-FL allocates more positions to the Pareto-ranked core and reduces random exposure.

The Pareto-core clients are selected according to non-dominated front rank, descending crowding distance, lower risk, lower depletion pressure, greater staleness, and finally a deterministic client-identifier tie-break. Exploration clients are sampled without replacement from the remaining eligible pool.

3.5.3. Pressure-Adaptive Robust Aggregation

ARL-FL does not assign direct client-specific aggregation weights based on risk or energy. Instead, the global pressure P^r controls the strength of two robust operations:

1. Update-norm clipping;
2. Coordinate-wise trimmed-mean aggregation.

Let

$$g_i^r = ||\Delta_i^r||_2.$$

The median and median absolute deviation of the selected update norms are

$$\widetilde{g}^r = \text{median}_{\text{i}\in\text{S}_\text{r}} g_i^r$$

and

$$\text{MAD}_\text{g}^\text{r} = \text{median}_{\text{i}\in\text{S}_\text{r}} \, | \, g_i^r - \widetilde{g}^r |.$$

The pressure-adaptive clipping coefficient is

$$\kappa(P^r) = \kappa_{\max} - (\kappa_{\max} - \kappa_{\min}) P^r.$$

Using the frozen values $\kappa_{\max} = 3$ and $\kappa_{\min} = 1$,

$$\kappa(P^r) = 3 - 2P^r.$$

The clipping threshold is

$$\tau^r = \widetilde{g}^r + 1.4826 \times \kappa(P^r) \text{MAD}_\text{g}^\text{r}.$$

To ensure a strictly positive clipping threshold, if its numerical evaluation produces a non-finite or non-positive value, it is replaced by

$$\tau^r \leftarrow \max(\tilde{g}^r, \varepsilon).$$

The client-specific clipping factor is

$$\chi_i^r = \min\left[1, \frac{\tau^r}{\max(g_i^r, \varepsilon)}\right],$$

and the clipped update is

$$\overline{\Delta}_i^r = \chi_i^r \Delta_i^r.$$

The coordinate-wise trimming level is

$$k(P^r) = \begin{cases} 1, & P^r < 0.5, \\ 2, & P^r \geq 0.5. \end{cases}$$

For every model coordinate d, the K clipped values are sorted:

$$\overline{\Delta}_{(1),d}^r \leq \cdots \leq \overline{\Delta}_{(K),d}^r.$$

The lowest and highest $k(P^r)$ values are removed, and the retained values are averaged:

$$\Delta_{\mathrm{ARL},d}^r = \frac{1}{K - 2k(P^r)} \sum_{j=k(P^r)+1}^{K-k(P^r)} \overline{\Delta}_{(j),d}^r.$$

The global update is

$$w^{r+1} = w^r + \Delta_{\mathrm{ARL}}^r.$$

This aggregation rule extends the coordinate-wise trimmed mean [8] with pressure-adaptive clipping and trimming. Under low pressure, clipping is relatively permissive, and one update is removed from each tail. Under high pressure, the clipping threshold becomes tighter, and two updates are removed from each tail.

The absence of direct risk or energy weighting is deliberate. Client-level risk and energy influence who participates, while the aggregation stage operates on the geometry of the selected updates. The global pressure changes the robustness level but does not determine individual aggregation coefficients.

To isolate the incremental effect of ARL-FL's adaptive pressure/risk control from the simpler combination of norm clipping and coordinate-wise trimming, a fixed norm-clipped trimmed-mean (FNCTM) control was evaluated separately on representative CICIoT2023 conditions. FNCTM first clipped each selected update to a fixed norm threshold $\kappa = 2.0$ and then removed k = 2 values from each coordinate tail, retaining 5/9 updates. It used the same frozen participation schedule and local-training protocol as the CICIoT2023 campaign, but no trust, risk, global-pressure, energy, or other adaptive parameter.

3.5.4. ARL-FL Round Procedure

One ARL-FL communication round consists of the following operations:

1. Deduct the standby cost from all active clients.
2. Construct the eligible-client set.
3. Compute each eligible client's risk, predicted depletion pressure, and staleness objectives.
4. Perform non-dominated sorting and crowding-distance ordering.
5. Determine the exploration count from the preceding round's global pressure.
6. Select the Pareto-core and exploration clients.
7. Broadcast the global model and perform local training.
8. Construct the selected clients' current model-update vectors.

9. Compute direction, upper-norm, and temporal anomaly signals.
10. Fuse the anomaly signals and update the persistent client risks.
11. Compute the current global pressure.
12. Deduct the selected clients' participation costs and update their selection histories.
13. Compute the pressure-adaptive clipping threshold and trimming level.
14. Clip the update norms and apply the coordinate-wise trimmed mean.
15. Update the global model.

3.6. Distinction Between TEA-FL and ARL-FL

Although both methods use reliability and resource information, their design philosophies differ substantially.

Their methodological contribution is the organization and interaction of established ingredients at the complete-round level. TEA-FL is not presented as a new robust statistic: it combines a server-maintained peer-consistency state with residual energy twice, first in exploit–explore client selection and then in sample-count/trust/energy aggregation. This differs from trust-root approaches such as FLTrust, which weight against a trusted server update, and from resource-only selection approaches, because the same reliability and resource states jointly shape both participation and update influence [18,22,25].

ARL-FL is likewise not presented as inventing trimmed mean, norm clipping, or Pareto ranking individually. Its contribution is to separate persistent client risk, expected depletion pressure, and participation staleness during scheduling and then translate the selected-client risk distribution into a global pressure variable that changes exploration, clipping tightness, and trimming strength. In contrast with fixed robust aggregation, robustness strength is therefore coupled to an online risk state; in contrast with scalar resource-aware schedulers, the scheduling objectives are kept separate by non-dominated sorting rather than collapsed into one utility score [8,23,28].

This distinction also limits the claims. The study does not argue that TEA-FL supersedes trusted-root or attack-specific defenses, nor that ARL-FL provides a new Byzantine-resilience guarantee. The contribution assessed experimentally is whether these two co-design patterns yield useful robustness–sustainability trade-offs under the declared threat, resource, and participation models.

TEA-FL uses a direct coupling strategy. Trust and residual energy jointly determine the selection score and the aggregation weights. This architecture is computationally simple and gives trusted, energy-sustainable clients greater influence throughout the round.

ARL-FL uses a separation strategy. Risk, depletion pressure, and staleness form separate Pareto objectives during scheduling. Aggregation subsequently uses global pressure to control robust clipping and trimming, but does not assign direct risk- or energy-based weights.

The main design differences are summarized in Table 2.

Table 2. Principal design differences between TEA-FL and ARL-FL.

Component	TEA-FL	ARL-FL
Reliability state	Distance-based consistency trust	Persistent multi-signal client risk
Reliability indicators	Distance to the equal-weight round reference	Direction anomaly, upper-norm anomaly, and temporal inconsistency
Resource indicator	Residual energy	Expected participation cost divided by residual energy

Table 2. *Cont.*

Component	TEA-FL	ARL-FL
Participation-history indicator	Indirectly preserved through random exploration	Explicit capped-staleness objective
Client selection	Weighted trust–energy score with fixed exploit–explore allocation	Non-dominated sorting and crowding distance with pressure-adaptive exploration
Aggregation	Sample count × trust × residual energy	Adaptive norm clipping followed by coordinate-wise trimmed mean
Direct reliability weighting during aggregation	Yes	No
Direct residual-energy weighting during aggregation	Yes	No
Global adversarial pressure	Not used	75th percentile of selected-client risk
Robustness adaptation	Trust changes client influence	Pressure changes clipping and trimming strength
Main design emphasis	Trust-driven robustness and energy-aware orchestration	Risk-aware scheduling, client retention, and adaptive robust aggregation

3.7. Dataset-Specific Execution Boundary

The complete TEA-FL and ARL-FL client-selection mechanisms were evaluated in the PAMAP2 campaign, where residual energy, client eligibility, dropout, participation lifetime, and fairness were explicit experimental outcomes.

For CICIoT2023, all six methods used the same externally frozen participation schedule. Consequently, method-specific selection functions did not determine the identities of participating clients in that campaign. This choice isolated local optimization, trust or risk updating, and server-side aggregation under exactly matched client participation. Therefore, the CICIoT2023 results provide cross-domain evidence regarding predictive and aggregation robustness, but they are not used to support claims concerning energy savings, dropout prevention, or method-specific client-selection fairness.

In CICIoT2023, the residual-energy values supplied to TEA-FL were fixed at the neutral constant

$$E_i^r = 1, \forall i, r.$$

Consequently, the energy term in the TEA-FL aggregation weight was neutral. Under the externally frozen participation schedule, TEA-FL retained trust-state updating and sample-count- and trust-weighted aggregation, but neither energy-aware selection nor energy-dependent aggregation was evaluated. ARL-FL retained risk-state updating, global-pressure estimation, norm clipping, and trimmed aggregation, whereas its Pareto scheduler did not determine client participation.

This execution boundary also applies to the operational-overhead analysis. Because method-specific scheduling was disabled in CICIoT2023, its campaign-level wall-time logs cannot quantify the cost of the complete TEA-FL or ARL-FL procedures. Complete-system overhead is therefore assessed in Section 5.8 using a dedicated PAMAP2 round benchmark in which the native client-selection mechanisms remain active. CICIoT2023 timing logs are retained only as descriptive execution records for the frozen-schedule campaign and are not used to estimate complete-system scheduling overhead.

3.8. Communication, Computation, and State Overhead

Neither TEA-FL nor ARL-FL requires an additional communication round. Each selected client receives one global model and returns one local model update, as in conventional synchronous FL.

For a model with D parameters and K selected clients, the TEA-FL trust calculation and aggregation require $O(KD)$ server operations. Selection over N eligible clients requires $O(N \log N)$ operations because of ranking. TEA-FL stores two scalar states per client: trust and residual energy.

ARL-FL risk estimation requires $O(KD)$ operations. With three scheduling objectives, non-dominated sorting has a worst-case cost of $O(N^2)$, while crowding-distance ordering adds $O(N \log N)$ operation [28]. Coordinate-wise sorting for robust aggregation requires $O(DK \log K)$ operations. ARL-FL stores scalar risk, residual energy, and last-participation information for each client, together with the most recent submitted update required for temporal consistency. Its worst-case server-state requirement is therefore $O(ND)$.

To empirically quantify these costs, a dedicated complete-round PAMAP2 benchmark used the canonical outer-fold-1, $\alpha = 1.0$, clean, seed-123 condition and reset each method to its canonical round-1 pre-state before every repetition. FedAvg, TEA-FL, and ARL-FL were each executed once as a warm-up and five additional times for measurement.

Client selection, local training, trust/risk processing, aggregation, and post-round evaluation were timed separately, and exact tensor storage was recorded. In addition, selection/scheduling was benchmarked without local training or evaluation for deterministic eligible-client populations of 28, 56, 112, and 224 clients, with 50 repetitions per population size. These measurements are descriptive engineering measurements; they do not alter the frozen scientific campaigns and do not constitute a new predictive-effectiveness endpoint.

4. Experimental Methodology

4.1. Experimental Design

Within each matched condition, methods shared the same initialization, data partition, malicious-client identities, and seed configuration. The core evidence comprises 600 primary PAMAP2 runs, 240 prespecified PAMAP2 ablation runs, and 180 CICIoT2023 runs (1020 core runs). The focused experimental component comprised 20 CICIoT2023 FNCTM runs, 20 CICIoT2023 coordinate-wise Median runs, 100 PAMAP2 mechanism-sensitivity runs, and 120 PAMAP2 resource-model-sensitivity runs, yielding 1280 scientific executions in total. These executions are not independent inferential replicates: PAMAP2 inference uses five held-out-subject fold blocks, and CICIoT2023 uses five seed blocks conditional on one frozen observation-level split. Detailed execution ledgers and hashes are retained in the versioned archive.

All TEA-FL and ARL-FL control parameters were frozen in the executable protocol before the corresponding primary scientific campaigns. The archived development record establishes the implemented values and technical validation sequence; it does not provide a prospective numerical tuning rule for every parameter. Appendix A reports the frozen values and roles. The ablation campaign evaluates complete mechanisms, while the sensitivity block varies only sign-flip magnitude, TEA-FL trust persistence, and the ARL-FL pressure switch over predeclared levels; it is a focused descriptive sensitivity analysis rather than retrospective outcome tuning or a factorial hyperparameter search.

Table 3 consolidates the principal design boundaries. PAMAP2 is the full-system campaign: method-specific selection, simulated energy, dropout, and federation viability are active. CICIoT2023 is an aggregation-focused cross-domain campaign: all methods use the same frozen participation schedule, so no method-specific energy, dropout, retention, or selection claim is inferred from that dataset.

Table 3. Overview of the PAMAP2 and CICIoT2023 experimental campaigns.

Characteristic	PAMAP2	CICIoT2023
Application domain	Wearable human activity recognition	IoT network intrusion detection
Prediction task	12-class classification	7-class classification
Input representation	Six inertial vector-magnitude channels	39 tabular network-flow features
Federated clients	28 per outer fold	30
Clients selected per round	8	9
Statistical heterogeneity	Dirichlet α = 0.1 and α = 1.0	Dirichlet α = 0.1 and α = 1.0
Communication rounds	100	20
Experimental seeds	123 and 456	42, 123, 456, 789, and 999
Main scenarios	Clean; label flip μ = 0.2; label flip μ = 0.4; sign flip μ = 0.2; sign flip μ = 0.4	Clean; sign flip μ = 0.2; sign flip μ = 0.4
Evaluated methods	FedAvg, FedProx, Random Trimmed Mean, FedLE-adapted, TEA-FL, and ARL-FL	Same six methods
Client selection	Method-specific	Externally frozen and identical across methods
Energy and dropout simulation	Yes	No
Main campaign size	600 runs	180 runs
Additional ablation campaign	240 runs	None
Focused experiments	220 runs: 100 mechanism-sensitivity + 120 resource-model-sensitivity	40 runs: 20 FNCTM + 20 Median matched-control runs
Primary predictive endpoint	Macro-F1	Test Macro-F1
Primary sustainability endpoint	Protocol-total normalized energy under the federation-viability stopping rule; post hoc energy per completed round and matched common-horizon energy	Not applicable
Primary inferential unit	Held-out subject fold (n = 5); two FL seeds averaged within fold	Experimental seed (n = 5); two α conditions averaged within seed
Evaluation split/population boundary	Leave-one-complete-subject-out outer evaluation	One frozen observation-level train/validation/test split; source-file overlap remains
Interpretation boundary	Complete method-specific selection–aggregation systems	Active trust/risk and aggregation components only under frozen participation

4.2. PAMAP2 Dataset and Preprocessing

4.2.1. Dataset

PAMAP2 is a publicly available physical activity monitoring dataset containing synchronized wearable measurements acquired from multiple subjects performing everyday, household, and sports activities [8]. The original recordings were sampled at 100 Hz using three Colibri wireless inertial measurement units (Trivisio Prototyping GmbH, Trier, Germany) positioned on the hand, chest, and ankle. The original dataset also included heart-rate measurements acquired with a BM-CS5SR monitor (BM innovations GmbH,

Hörgertshausen, Germany); however, the heart-rate stream was not used by the final federated model.

The raw protocol files contain 54 columns, including timestamp, activity label, heart rate, temperature, acceleration, gyroscope, magnetometer, and orientation measurements. The present study used the 12 activities consistently represented in the frozen evaluation protocol:

1. Lying;
2. Sitting;
3. Standing;
4. Walking;
5. Running;
6. Cycling;
7. Nordic walking;
8. Ascending stairs;
9. Descending stairs;
10. Vacuum cleaning;
11. Ironing;
12. Rope jumping.

Samples assigned to activity label 0, corresponding to transient or unlabeled periods, were excluded. Subject 109 was also excluded from the final federated evaluation because it did not satisfy the completeness requirements adopted for the cross-subject protocol.

4.2.2. Window Construction

Sensor observations assigned to activity label 0 were removed before temporal segmentation. For each subject, the retained sensor channels were ordered by timestamp and processed independently within activity-continuous segments. Missing internal sensor values were reconstructed using deterministic linear interpolation, subject to the frozen maximum-gap criterion

$$\text{max_gap} = 100,$$

where max_gap denotes the maximum number of consecutive missing samples permitted for internal linear interpolation.

Interpolation was not allowed to bridge subject boundaries, changes in activity label, or discontinuities exceeding this criterion. Such boundaries started a new continuous segment. Any values that remained unresolved after application of the interpolation and continuity rules were not propagated into the learning dataset.

Each valid continuous segment was divided into windows containing 256 samples, corresponding to 2.56 s at the original 100 Hz sampling rate. Consecutive windows used a stride of 128 samples, giving 50% overlap. Windows were generated entirely within one subject, one activity label, and one continuous segment; no window was permitted to cross a subject boundary, activity transition, or invalid temporal discontinuity. A candidate window was retained only when all 256 observations were available and finite for every channel in the frozen 36-channel representation.

The resulting processed dataset contained

$$14{,}972$$

windows, each initially represented as

$$256 \times 36$$

inertial-axis values. For each of the hand, chest, and ankle IMUs, the stored representation comprised the three axes of the ±16 g accelerometer, the three axes of the ±6 g accelerometer, the three gyroscope axes, and the three magnetometer axes (3 locations × 4 modalities × 3 axes = 36). The ±6 g channels were retained in Full36 storage, but were not used by the final six-channel federated representation.

4.2.3. Final Sensor Representation

A lightweight six-channel representation was selected before the main federated campaign through the frozen subject-wise model-selection protocol.

For each sensor location, the Euclidean vector magnitude was calculated for:

- The ±16 g accelerometer;
- The gyroscope.

The final input therefore contained:

3 sensor locations × 2 modalities = 6 channels.

The magnetometer axes and heart-rate stream were not used by the final federated model.

Feature standardization was performed independently within each outer fold. The mean and standard deviation were fitted using only the windows assigned to the training clients of that fold. The held-out test subject was never used to estimate normalization parameters.

4.2.4. Cross-Subject Evaluation Protocol

A five-fold leave-one-complete-subject-out design was used. The outer test subjects were:

$$101, 102, 105, 106, \text{and } 108.$$

Each selected test subject contained observations from all 12 activities. For a given outer fold, all remaining eligible core subjects were used to construct the federated training population.

Model representation and normalization choices were frozen through a nested subject-wise selection procedure. For each outer test subject, the other four complete subjects served in turn as inner validation subjects; the corresponding inner-training set contained all eight core subjects except the outer test subject and the current inner validation subject. Thus, 20 inner folds were evaluated in total. Each candidate was evaluated with inner seeds 123 and 456; selection used mean validation Macro-F1 across the eight inner runs per candidate, candidates within 0.005 Macro-F1 of the best were treated as equivalent, and ties were resolved by fewer input channels and then lower variability. Fold-specific z-score statistics were fitted on inner-training subjects only. All 20 inner-fold rows satisfied the exclusion criterion and confirmed that the outer test subject was excluded from both inner training and inner validation in every case. Table A4 lists the outer/inner subject structure. No decision made during the 600-run scientific campaign used the outer test results.

The global model was evaluated at round 0 and subsequently every five communication rounds. These evaluations supported temporal analysis and the post hoc common-horizon comparison. The prespecified protocol endpoint was the model state after round 100 or, when federation viability was lost, the state after the last completed training round. Consequently, protocol endpoints could occur at different operating horizons across methods. No checkpoint was selected by maximizing outer-test performance.

4.3. CICIoT2023 Dataset and Preprocessing

4.3.1. Dataset

CICIoT2023 was developed using a realistic large-scale IoT topology containing benign devices and multiple attack families [12]. It provides network-flow features for both normal traffic and a broad range of IoT attacks.

The original attack labels were consolidated into the following seven-class task:

1. Benign;
2. Brute Force;
3. DoS/DDoS;
4. Mirai;
5. Reconnaissance;
6. Spoofing;
7. Web-based attacks.

This class consolidation reduced excessive label fragmentation while preserving the principal operational attack families.

4.3.2. Exact-Unique Representation

Before duplicate removal, the raw CICIoT2023 records were mapped to the frozen seven-class task. Three structurally malformed physical records were excluded. In addition, every structurally valid row containing at least one non-finite numeric feature value—NaN, positive infinity, or negative infinity—was excluded before preprocessing. No missing or non-finite value was imputed, clipped, repaired, or replaced.

Exact feature vectors that remained associated with more than one seven-class label after the quality exclusions were treated as residually ambiguous. All records belonging to such conflicting exact-vector groups were removed rather than arbitrarily assigned to one class.

The remaining features were parsed in the canonical 39-feature order and represented in 64-bit floating-point format. Signed zero was canonicalized before comparison. Two records were considered duplicates only when all 39 canonical float64 feature values were bitwise identical. The 128-bit row signatures used during construction served only for bucket partitioning and indexing; final duplicate identity was established through direct exact comparison of the complete feature vectors.

One representative was retained from every exact canonical feature-vector group. The raw multiplicity and the corresponding source-file and original fine-class provenance counts were preserved as metadata. This procedure produced an effective dataset containing

$$20{,}687{,}272$$

unique observations.

A deterministic class-stratified split was subsequently constructed at the effective-observation level. The split contained 16,549,824 training observations, 2,068,724 validation observations, and 2,068,724 test observations, corresponding approximately to an 80%–10%–10% allocation.

After exact-vector deduplication, the deterministic split assigned each effective observation to one partition, preventing an identical retained effective vector from appearing in more than one partition. Source-file provenance was audited separately. All 309 source files contributed observations to the training, validation, and test partitions, and 100% of the effective observations were linked to at least one source file represented in all three partitions. The median within-file allocation was 79.98%/10.01%/10.01% (training/validation/test), with ranges of 78.04–80.90%, 9.33–11.31%, and 9.12–11.26%, respectively. Furthermore,

1,930,227 effective observations (9.33%) inherited provenance from more than one source file after exact-vector provenance consolidation. A targeted source-file feasibility audit found 4 Benign, 1 Brute Force, 214 DoS/DDoS, 76 Mirai, 5 Reconnaissance, 3 Spoofing, and 6 Web-based-attack source files. Because Brute Force occurs in only one source file, no nonempty source-file holdout can preserve all seven coarse classes simultaneously in both development and held-out sets. We therefore did not introduce a post hoc file-holdout effectiveness experiment.

4.3.3. Feature Transformation

The CICIoT2023 model used 39 numeric network-flow features. Each feature was transformed using a signed logarithmic transformation:

$$x' = \text{sign}(x) \log(1 + |x|).$$

The transformed values were subsequently scaled using the maximum absolute value estimated exclusively from the training partition:

$$\tilde{x}_j = \frac{x'_j}{\max_{\text{training}} | x'_j |}.$$

All preprocessed features were stored in 32-bit floating-point format. Validation and test observations were transformed using the statistics obtained from the training partition.

Across all 20,687,272 final 39-dimensional model-input vectors, an audit found zero remaining exact duplicate groups, zero cross-partition duplicate groups, zero conflicting-label duplicate groups, and zero cross-partition conflicting-label groups. Thus, the final float32 representation contained 20,687,272 unique model-input vectors. This audit concerns equality of the actual stored model inputs; it does not establish source-file, device, session, or attack-episode independence.

4.4. Federated Client Partitions and Participation

For clarity, the non-IID design can be summarized as follows: both datasets used class-conditional Dirichlet allocation with $\alpha \in \{0.1, 1.0\}$, where $\alpha = 0.1$ represents the strongly non-IID setting and $\alpha = 1.0$ the moderately heterogeneous setting. PAMAP2 used 28 virtual clients per outer fold, constructed as four subject-specific clients from each of seven training subjects; the partitions were fixed across the two federated seeds, and $K = 8$ clients were selected per round. CICIoT2023 used 30 virtual clients and 10 partition configurations formed by the two Dirichlet levels and five experimental seeds; $K = 9$ clients participated per round under the frozen common participation schedule. The detailed construction rules, partition audits, and client-level heterogeneity diagnostics are reported below and, for PAMAP2, further summarized in Appendix A.

4.4.1. Dirichlet Non-IID Partitions

For both datasets, statistical heterogeneity was controlled through class-conditional Dirichlet partitioning.

Two concentration values were evaluated:

$$\alpha = 1.0$$

and

$$\alpha = 0.1.$$

The $\alpha = 1.0$ setting produced moderately heterogeneous client-label distributions, whereas $\alpha = 0.1$ produced more concentrated and strongly non-IID local datasets.

The generated partitions were audited before training. The $\alpha = 0.1$ partitions exhibited lower normalized label entropy and greater divergence from the corresponding global fold distribution than the $\alpha = 1.0$ partitions.

4.4.2. PAMAP2 Client Construction

For each PAMAP2 outer fold, the seven training subjects were retained as distinct data sources, and each subject was partitioned into four subject-specific virtual clients, yielding 28 federated clients. No virtual client contained windows from more than one original subject; however, one physical subject corresponded to four federated clients, so the construction is not a one-to-one client–subject mapping. The construction imposed controlled within-subject class-distribution heterogeneity, whereas independence of the held-out population was enforced at the outer level through leave-one-complete-subject-out evaluation. Every virtual client contained at least 128 windows, and the client partitions were fixed across the two federated seeds.

At each communication round, eight eligible clients were selected:

$$K = 8.$$

FedAvg, FedProx, and Random Trimmed Mean used a common random client-selection schedule within each matched condition. FedLE-adapted, TEA-FL, and ARL-FL used their corresponding resource-, trust-, or risk-aware selection mechanisms.

The same initial client-energy profile, device-compute factor, malicious-client identities, model initialization, and experimental seed were shared across methods within a matched condition.

Table A1 in Appendix A characterizes the 28 virtual clients in each outer fold. For $\alpha = 0.1$, the median client class entropy ranged from 0.4654 to 0.5391 across folds; for $\alpha = 1.0$, it ranged from 0.8419 to 0.8531. Every virtual client contained windows from exactly one original subject, and the partitions did not depend on the federated seed.

4.4.3. CICIoT2023 Client Construction

The CICIoT2023 training observations were assigned to 30 virtual clients. Ten client-partition configurations were created from:

$$2\ \text{Dirichletlevels} \times 5\ \text{experimental seeds}.$$

The experimental seeds were

$$42, 123, 456, 789, \text{and } 999.$$

A common participation schedule was frozen for every seed. Nine clients participated per communication round, and each client participated exactly six times over the complete 20-round campaign.

Every client participation used a deterministic block of 51,200 local training observations. For each client/configuration, candidate observations were placed in a deterministic hash-based local order and the six scheduled participations used successive 51,200-example intervals from that order. The targeted audit covered all 10 partition configurations, 30 clients per configuration, and six blocks per client (1800 blocks total). No client had insufficient capacity, and no within-run block intervals overlapped; hence, examples were not reused across a client's six scheduled blocks within a run. Block-level class support ranged from 1 to 7 classes: 253 blocks supported 1 class, 223 supported 2, 144 supported 3, 161 supported 4, 74 supported 5, 119 supported 6, and 826 supported all 7 classes. With a batch size of 256, each participation contained exactly 200 local optimizer steps.

The resulting workload per complete CICIoT2023 run was

$$20 \times 9 \times 200 = 36{,}000$$

scientific optimizer steps.

Because the participation schedule was identical for all six methods, the CICIoT2023 campaign isolated differences in local optimization, trust or risk updating, and server-side aggregation. It did not evaluate method-specific client-selection energy savings or dropout prevention.

4.5. Prediction Models and Local Optimization

4.5.1. PAMAP2 Lightweight One-Dimensional CNN

PAMAP2 used a lightweight one-dimensional convolutional neural network with 77,004 trainable parameters.

The feature extractor contained six convolutional blocks:

1. 6→32, kernel size 7, stride 2;
2. 32→32, kernel size 5, stride 1;
3. 32→64, kernel size 5, stride 2;
4. 64→64, kernel size 3, stride 1;
5. 64→96, kernel size 3, stride 2;
6. 96→96, kernel size 3, stride 1.

Each convolution was followed by Group Normalization using eight groups and a rectified linear unit activation. Adaptive average pooling reduced the temporal dimension to one. The classifier consisted of dropout with probability 0.20 followed by a fully connected 96→12 output layer.

Each selected client performed one local epoch using:

- AdamW optimizer;
- Learning rate 10^{-3};
- Weight decay 10^{-4};
- Batch size 64;
- Cross-entropy loss.

FedProx used the same optimizer and added a proximal coefficient

$$\mu_{\text{prox}} = 0.01.$$

4.5.2. CICIoT2023 Multilayer Perceptron

CICIoT2023 used a multilayer perceptron with the following architecture:
39→256→128→7.

The two hidden layers used rectified linear unit activations and dropout with probability 0.20. The model contained 44,039 trainable parameters.

Local training used:

- Adam optimizer;
- Learning rate 10^{-3};
- Coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$;
- Numerical-stability term 10^{-8};
- Batch size 256;
- 200 optimizer steps per client participation;
- Weighted cross-entropy loss.

The class weights were calculated globally from the frozen training distribution and were kept identical across methods and experimental conditions. A provenance

audit reconstructed the exact count basis as final training effective rows, with class counts [874,262, 10,436, 12,773,255, 1,963,693, 545,764, 362,726, 19,688]. The frozen weights [2.361947655626154, 21.61842421005929, 0.6179314203594863, 1.5759929623559756, 2.989431643370104, 3.6669224591635454, 15.7394759911818] satisfy

$$w_c = \frac{2208.467791239651}{\sqrt{n_c}},$$

up to a maximum absolute reconstruction error of 3.55×10^{-15}. Equivalently, $w_c \propto n_c^{-1/2}$. Because the weighted cross-entropy used mean reduction, the common positive scale factor does not change the loss value or gradients relative to the proportional rule.

FedProx again used

$$\mu_{\text{prox}} = 0.01.$$

After every communication round, the global model was evaluated on the validation partition. The checkpoint with the highest validation Macro-F1 was retained; balanced accuracy was the first tie-breaker and the earliest round the second. The test partition was evaluated only after checkpoint selection. This original primary endpoint is denoted E0 (validation-selected over rounds 1–20). To distinguish checkpoint preservation from continued learning during an active attack, two sensitivity endpoints were defined before inspecting their test results: E1, the fixed final-round model at round 20 with no checkpoint selection, and E3, the best-validation checkpoint selected only among attack-active rounds 5–20 using the same validation-only lexicographic rule. Test metrics were never used to select E0 or E3.

All 180 E1 round-20 states were available from the preserved final checkpoints. For E3, the validation-selected attack-period state was already preserved for 50/120 attacked runs. For the remaining 70 runs, deterministic replay under the frozen scientific kernel was performed only to the validation-selected target round; the historical validation prefix reproduced exactly in 70/70 runs, and each reconstructed target state was evaluated on the test split once. The 50 preserved E3 states reused the original E0 test result because E0 and E3 referred to the same state. Supplemental outputs were isolated from the original campaign directories, which were not modified.

The main model architectures and local-training configurations used for the two datasets are summarized in Table 4.

Table 4. Main model and local-training configurations.

Parameter	PAMAP2	CICIoT2023
Model	Lightweight 1D CNN	Multilayer perceptron
Input	6×256	39 features
Output classes	12	7
Trainable parameters	77,004	44,039
Hidden channels/layers	32, 64, and 96 convolutional channels	256 and 128 hidden units
Normalization	Group Normalization	None
Dropout	0.20	0.20
Local optimizer	AdamW	Adam
Learning rate	10^{-3}	10^{-3}
Weight decay	10^{-4}	0
Local batch size	64	256
Local workload	One local epoch	200 local steps
Loss	Cross-entropy	Class-weighted cross-entropy
FedProx coefficient	0.01	0.01

4.6. Comparison Methods

Six federated methods comprised the frozen primary campaigns. Two robust controls—fixed norm-clipped trimmed mean (FNCTM) and coordinate-wise Median—were evaluated in focused CICIoT2023 campaigns using the same frozen participation and local-training protocol as the primary CICIoT2023 campaign.

4.6.1. FedAvg

FedAvg applied sample-count-weighted arithmetic averaging and served as the principal conventional baseline [1].

4.6.2. FedProx

FedProx used the same server aggregation as FedAvg but included the proximal term described in Section 2 to reduce local model drift under non-IID data [4].

4.6.3. Random Trimmed Mean

Random Trimmed Mean used the common random participation schedule in PAMAP2 and the externally frozen common participation schedule in CICIoT2023. For each floating-point model coordinate, the local model deltas relative to the current global model were sorted across the selected clients.

The two smallest and two largest values were removed:

$$k = 2.$$

at each coordinate and in both experimental campaigns. The retained coordinate values were averaged and added to the current global-model parameters. Consequently, four of the eight selected PAMAP2 updates and five of the nine selected CICIoT2023 updates contributed to each coordinate-wise aggregate.

This method separated the effect of fixed robust aggregation from the trust-, risk-, and energy-aware client-selection mechanisms used by the proposed methods.

4.6.4. FedLE-Adapted

FedLE-adapted implemented a lifespan-oriented client-selection strategy within the common PAMAP2 model, energy, and attack environment. At the beginning of each run, a one-time preflight was executed over all 28 federated clients. Each client performed one local epoch starting from the common initial global model.

For each client, the preflight representation was constructed from the normalized local-update deltas of three model tensors: the first convolutional-layer weight, the final linear-layer weight, and the final linear-layer bias. Using update deltas rather than complete local-model vectors removed the common global initialization from the clustering representation.

The resulting client vectors were L2-normalized and grouped once using deterministic K-means clustering with eight clusters and 20 initializations. The cluster assignments remained fixed throughout the corresponding run. The local computation and communication costs associated with the preflight were deducted from the clients' residual-energy states and were included in the scientific optimizer-step and energy ledgers.

At round r, each eligible client i received the selection score

$$S^r_{i,\mathrm{FedLE}} = \frac{E^r_i / E^0_i}{|G(i)|},$$

where E^r_i / E^0_i is the client's residual-energy fraction and $|G(i)|$ is the size of its frozen K-means cluster. The scores were normalized across the eligible population to form sampling probabilities, and eight clients were sampled without replacement using a deterministic

round-specific random seed. The submitted local updates were subsequently aggregated using sample-count-weighted FedAvg.

In CICIoT2023, the externally frozen common participation schedule disabled method-specific FedLE client selection. Under that boundary, FedLE-adapted and FedAvg are algorithmically equivalent: both use the same selected clients, local training, and sample-count-weighted averaging. The historical frozen campaign nevertheless used numerically different, but mathematically equivalent floating-point accumulation paths for the two labels; its separate FedLE-adapted row is therefore retained only as a provenance record and is not treated as independent algorithmic corroboration.

A controlled deterministic replay verified this equivalence over all 30 matched scenario–heterogeneity–seed pairs. With the same runner, data, schedules, RNG streams, and direct FedAvg aggregation implementation, the 20-round validation trajectories, BEST_STATE tensors, round-20 global states, and test metrics were exactly identical in 30/30 pairs. A one-round numerical canary traced the historical discrepancy to floating-point evaluation order: the alternative aggregation paths differed by at most 1.49×10^{-8} in tensor values, whereas the historical mean absolute FedLE–FedAvg Macro-F1 difference was 0.00273 (maximum 0.03852).

4.6.5. Fixed Norm-Clipped Trimmed Mean

FNCTM was a focused matched control for ARL-FL rather than a seventh primary-campaign baseline. It retained the CICIoT2023 frozen participation schedule and the same local training used by Random Trimmed Mean, but added a fixed update-norm clipping stage ($\kappa = 2.0$) before the same coordinate-wise trimming level k = 2. No parameter was selected from test outcomes. The focused design comprised clean and strong sign-flip ($\mu = 0.4$) conditions at $\alpha \in \{0.1, 1.0\}$, with five experimental seeds per cell (20 runs; 400 federated rounds). E0, E1, and E3 were defined exactly as in Section 4.5.2. This control was chosen because it isolates the missing fixed-clipping ingredient while holding the trimming rule constant. A second focused campaign evaluated coordinate-wise Median under the same frozen CICIoT2023 participation and local-training protocol, thereby changing only the aggregation functional.

4.6.6. Coordinate-Wise Median

Coordinate-wise Median was evaluated in a focused CICIoT2023 robust-aggregation campaign using the externally frozen participation schedule and the same local training as Random Trimmed Mean. For each floating-point model coordinate, the nine local model deltas relative to the current global model were stacked and their coordinate-wise median—the fifth order statistic for $K = 9$—was added to the current global parameters. The method used no clipping, trimming, trust, risk, pressure, energy weighting, or trusted server-side root data.

The focused design matched the FNCTM control: clean and strong sign-flip ($\mu = 0.4$) conditions at $\alpha \in \{0.1, 1.0\}$, with five experimental seeds per cell (20 runs; 400 federated rounds). The primary endpoint was E0, selected by validation Macro-F1 over rounds 1–20 exactly as in Section 4.5.2; test outcomes were not used for checkpoint selection. Because E0 may select one of the clean warm-up rounds 1–4 in attacked runs, the focused Median result is interpreted as validation-selected checkpoint preservation and not, by itself, as evidence of sustained post-attack robustness.

4.6.7. TEA-FL and ARL-FL

TEA-FL and ARL-FL were implemented according to the definitions in Section 3. Their principal differences concerned:

- Direct trust–energy coupling versus separated multi-objective scheduling;
- Weighted aggregation versus pressure-adaptive robust aggregation;
- Fixed versus pressure-dependent exploration;
- Consistency trust versus multi-signal persistent risk.

Except for the explicitly accounted one-time FedLE-adapted preflight, no method received additional round-level local epochs or communication rounds. TEA-FL and ARL-FL used the same local training workload and communication structure as their matched comparators.

The frozen primary campaigns used Random Trimmed Mean as the representative fixed robust-aggregation baseline because it could be applied under the same selected-client workloads without requiring a trusted server dataset, Sybil metadata, or additional communication. The focused CICIoT2023 controls additionally include FNCTM and coordinate-wise Median without modifying the frozen 180-run primary campaign. Krum and Multi-Krum require a predeclared Byzantine-count bound relative to the selected-client population. The realized number of selected malicious clients is schedule-dependent, so the admissibility assumptions required by Krum/Multi-Krum cannot be guaranteed uniformly across the evaluated settings without a separate protocol-specific certification. FLTrust requires a trusted server-side root dataset, which was outside the present threat and data model. FoolsGold primarily targets correlated Sybil behavior, whereas Sybil identities and coordinated similarity attacks were not part of the evaluated threat model. FLAME principally targets backdoor behavior and introduces additional clustering and noise assumptions. Accordingly, the comparison covers both trimmed-mean and median-style coordinate-wise robust estimators plus the fixed-clipping control, but is not intended as an exhaustive comparison of all Byzantine defenses.

4.7. Adversarial Scenarios

4.7.1. PAMAP2 Scenarios

Five scenarios were evaluated:

1. Clean;
2. Label flip with $\mu = 0.2$;
3. Label flip with $\mu = 0.4$;
4. Sign flip with $\mu = 0.2$;
5. Sign flip with $\mu = 0.4$.

With 28 clients, the two malicious fractions corresponded to:

$$6 \text{ malicious clients}$$

and

$$11 \text{ malicious clients},$$

respectively.

Attacks became active at communication round 20 and remained active for the rest of the run. The same malicious-client identities were used across methods, heterogeneity levels, and attack types for the corresponding fold, malicious fraction, and experimental seed.

The sign-flip operator defined in Equation (1) was used with $\lambda = 5$.

4.7.2. CICIoT2023 Scenarios

Three scenarios were included:

1. Clean;
2. Sign flip with $\mu = 0.2$;
3. Sign flip with $\mu = 0.4$.

The two malicious prefixes contained six and twelve of the 30 clients, respectively. Malicious identities were obtained from a deterministic ranking associated with each experimental seed and were identical across methods.

The clean scenario contained no active attack rounds. In both sign-flip scenarios, rounds 1–4 constituted a common clean warm-up period, and the attack was active from round 5 through round 20, inclusive. During each active round, every participating malicious client had its local update transformed by the sign-flip operator in Equation (1), with $\lambda = 5$. Because E0 selected from rounds 1–20, an attacked run could legitimately retain a clean-warm-up checkpoint; E1 and E3 were defined to distinguish checkpoint-protection effects from final-round and attack-period learning.

The CICIoT2023 campaign did not include label-flip scenarios because its purpose was focused on cross-domain validation of aggregation robustness under sign-flip attacks.

4.8. PAMAP2 Energy and Federation-Viability Configuration

The normalized energy environment followed the model defined in Section 3.3.

The frozen parameters were:

$$E_i^0 \sim \mathcal{U}(0.70, 1.00),$$

$$q_i \sim \mathcal{U}(0.80, 1.20),$$

$$c_{\text{st}} = 0.0005,$$

$$c_{\text{comm}} = 0.004,$$

$$c_{\text{comp}} = 0.016,$$

and

$$E_{\text{crit}} = 0.10.$$

A PAMAP2 run stopped early when the proportion of active clients fell below 50% of the original federation. These protocol-defined early stops were retained as valid scientific outcomes because they represented loss of federation viability rather than technical execution failure.

All methods faced identical initial resource profiles and cost coefficients. Differences in energy consumption or client retention therefore arose from their participation and aggregation behavior rather than from different resource budgets.

4.9. Campaign Structure and Component Ablation

4.9.1. Primary PAMAP2 Campaign

The primary PAMAP2 campaign comprised

$$5 \text{ folds } \times 2 \text{ values of } \alpha \times 5 \text{ scenarios } \times 2 \text{ seeds } \times 6 \text{ methods } = 600 \text{ runs}.$$

The two federated seeds were 123 and 456.

For every combination of fold, α, scenario, and seed, the six methods formed a matched comparison block.

4.9.2. PAMAP2 Component-Ablation Campaign

A prespecified component-ablation campaign evaluated four incomplete variants:

- tea_trust_only: trust retained and energy influence neutralized;
- tea_energy_only: energy retained and trust influence neutralized;
- arl_no_energy: energy-aware scheduling removed;
- arl_no_pressure: global adaptive pressure fixed to zero.

Six informative conditions were selected:

- $\alpha = 0.1$, clean;
- $\alpha = 1.0$, label flip $\mu = 0.4$;
- $\alpha = 0.1$, sign flip $\mu = 0.2$;
- $\alpha = 1.0$, sign flip $\mu = 0.2$;
- $\alpha = 0.1$, sign flip $\mu = 0.4$;
- $\alpha = 1.0$, sign flip $\mu = 0.4$.

The campaign comprised

$$6 \text{ conditions } \times 5 \text{ folds } \times 2 \text{ seeds } \times 4 \text{ ablation variants } = 240 \text{ runs.}$$

The complete TEA-FL and ARL-FL controls were not retrained. Their matched results were reused from the frozen primary campaign.

4.9.3. PAMAP2 Mechanism Diagnostics and Focused Sensitivity

Mechanism diagnostics were extracted as read-only from the frozen primary PAMAP2 telemetry for all 100 TEA-FL and all 100 ARL-FL runs. For TEA-FL, the analysis separated benign and malicious trust, selected-client participation, and aggregation mass, and related benign trust/weight to class coverage, normalized class entropy, and dominant-class concentration. For ARL-FL, it separated benign and malicious risk, clipping, and selected-client participation, and audited every round for the global-pressure switch at 0.5, the associated transition from one to two trimmed coordinates per tail, and the realized number of malicious clients selected in those rounds. No new training or test inference was required for these diagnostics.

A matched sensitivity analysis evaluated the three numerical controls. It used only the strong sign-flip condition ($\mu = 0.4$), both α values, all five outer folds, and federated seed 123. The predeclared levels were $\lambda \in \{2.5, 5.0, 7.5\}$, TEA-FL trust persistence $\beta_T \in \{0.65, 0.80, 0.95\}$, and the ARL-FL pressure switch $\tau_P \in \{0.35, 0.50, 0.65\}$. For TEA-FL, each fold–$\alpha$ block contained one contemporaneous central control plus two λ variants and two β_T variants; ARL-FL used the analogous contemporaneous central control, two λ variants, and two τ_P variants. The same τ_P controlled the transition to one exploration slot and the switch to two-value trimming, while the lower exploration threshold remained fixed at 0.25. This yielded 50 TEA-FL and 50 ARL-FL runs: 20 contemporaneous central controls and 80 non-central variants. Settings were fixed before their outcomes were inspected, test metrics were not used for parameter selection, and all comparisons are made against the contemporaneous central control within the same sensitivity block.

4.9.4. PAMAP2 Resource-Model Sensitivity

A predeclared resource-model sensitivity analysis tested whether the sustainability conclusions depended on the primary PAMAP2 resource configuration. The study used two representative conditions—clean $\alpha = 0.1$ and strong sign flip ($\mu = 0.4, \alpha = 1.0$)—all five held-out-subject folds, federated seed 123, and three methods (FedAvg, TEA-FL, and ARL-FL). Four resource regimes were evaluated: the frozen primary/computation-dominant setting $(c_{\text{comm}}, c_{\text{comp}}, E_{\text{crit}}) = (0.004, 0.016, 0.10)$; a communication-dominant setting $(0.016, 0.004, 0.10)$; a lower eligibility threshold $(0.004, 0.016, 0.05)$; and a stricter threshold $(0.004, 0.016, 0.20)$. The standby cost remained at 0.0005. The resulting design contained $4 \times 2 \times 5 \times 3 = 120$ matched scientific runs. The resource settings were fixed before execution, and the sensitivity was interpreted descriptively at the five-fold level rather than as parameter tuning.

4.9.5. CICIoT2023 Campaign

The CICIoT2023 campaign comprised

$$2 \text{ values of } \alpha \times 5 \text{ seeds } \times 6 \text{ methods } \times 3 \text{ scenarios } = 180 \text{ runs.}$$

Each run contained exactly 20 rounds and 36,000 accepted optimizer steps.

4.10. Evaluation Metrics

4.10.1. Predictive Metrics

Macro-F1 was the primary predictive endpoint because both datasets contained multi-class imbalance. It assigns equal weight to each class regardless of class frequency:

$$\text{MacroF1} = \frac{1}{\mathcal{C}} \sum_{c=1}^{\mathcal{C}} F1_c.$$

Balanced accuracy was calculated as the arithmetic mean of class recalls:

$$\text{BAcc} = \frac{1}{\mathcal{C}} \sum_{c=1}^{\mathcal{C}} \text{Recall}_c.$$

Accuracy and cross-entropy loss were also reported. For PAMAP2, per-class F1 values were retained for all 12 activities.

4.10.2. Energy and Lifetime Metrics

PAMAP2 additionally reported:

- Protocol-total normalized energy consumed under the federation-viability stopping rule;
- Normalized energy consumed per completed training round;
- Cumulative normalized energy at a matched common horizon;
- Completed training rounds and completion of the nominal 100-round horizon;
- Final mean residual energy;
- Final minimum residual energy;
- First client-dropout round;
- Active-client lifetime at 75% of the initial federation;
- Active-client lifetime at 50% of the initial federation;
- Final number of active clients;
- Participation fairness.

Let R_{bm} denote the number of completed training rounds and E_{bm}^{total} the protocol-total normalized energy for matched block b and method m. Energy per completed round was defined as $E_{bm}^{\text{rate}} E_{bm}^{\text{total}} / R_{bm}$. The matched common horizon was defined as $H_b = 5\lfloor \min_m R_{bm}/5 \rfloor$, that is, the latest scheduled evaluation round reached by all six methods in block b. Common-horizon cumulative energy was obtained from the decrease in total residual energy between round 0 and the matched common horizon, and common-horizon Macro-F1 was evaluated at the same round. Protocol-total energy was prespecified; the per-round and common-horizon outcomes were post hoc comparability analyses. Protocol-total energy was also reconstructed exactly as standby, communication, and computation components for every primary PAMAP2 run; the reconstruction matched the recorded totals to a maximum absolute error of 9.237×10^{-14}. Processed-example and accepted optimizer-step workloads were reconstructed from the same frozen run ledgers. FedLE-adapted's one-time preflight is part of its scientific communication/computation and workload totals. Its recorded round-0 residual-energy state occurs *after* that preflight; therefore, a round-0-to-

common-horizon difference alone omits the one-time preflight cost. Final common-horizon comparisons involving FedLE use the initial-to-horizon correction described in Section 5.2.

Let p_i denote the number of times client i participated. Jain's participation-fairness index was calculated as [29]

$$J = \frac{\left(\sum_{i=1}^{N} p_i\right)^2}{N \sum_{i=1}^{N} p_i^2}.$$

The index approaches one when participation is evenly distributed and decreases when a small number of clients dominate the federation.

4.10.3. Robustness Metrics

For each method and attack condition, absolute degradation relative to its matched clean run was defined as

$$D_{\text{abs}} = \text{MacroF1}_{\text{clean}} - \text{MacroF1}_{\text{attack}}.$$

Robustness retention was defined as

$$R_{\text{ret}} = \frac{\text{MacroF1}_{attack}}{\text{MacroF1}_{\text{clean}}}.$$

PAMAP2 temporal analyses additionally considered:

- Performance immediately before attack onset;
- Immediate post-attack change;
- Lowest post-attack Macro-F1;
- Recovery after the post-attack minimum;
- Evolution of residual energy and active clients;
- Evolution of ARL-FL global pressure.

4.11. Statistical Analysis

For explicit replication and uncertainty accounting, PAMAP2 used federated seeds 123 and 456 within each of the five held-out-subject folds; the two seed outcomes were averaged within fold before inferential testing. CICIoT2023 used experimental seeds 42, 123, 456, 789, and 999; for the primary scenario-level comparisons, the two α conditions were averaged within each seed, yielding five matched seed blocks. Summary tables report mean $\pm$ standard deviation across the stated fold/seed blocks, while effect-estimation analyses use deterministic 95% paired cluster-bootstrap confidence intervals. Statistical significance is assessed with the exact matched procedures defined below, with multiplicity control where applicable; bootstrap-interval exclusion of zero is not substituted for the declared exact-test criterion.

Inference used matched nonparametric procedures because only five primary blocks were available in each dataset. The significance level was 0.05. Friedman tests with Kendall's W summarized six-method omnibus differences, and planned paired contrasts used exact two-sided Wilcoxon signed-rank procedures [30–32]. Deterministic 95% percentile confidence intervals were obtained from 5000 paired cluster-bootstrap replicates and were interpreted as estimation-uncertainty summaries, not as substitutes for the exact tests [33]. Holm adjustment controlled the prespecified pairwise families [34]; exploratory false-discovery-rate summaries used Benjamini–Hochberg adjustment [35].

For PAMAP2, the held-out outer-subject fold was the inferential unit. The two federated seeds were averaged within fold before testing, producing five matched blocks for every scope. The outer-fold training sets overlap substantially because each leave-one-subject-out fold reuses most training subjects; therefore the folds are not described

as independent training datasets. They are matched held-out-subject evaluation blocks, and no procedure treats the two seeds or the overlapping training sets as additional population replicates. Primary endpoints were Macro-F1 and protocol-total normalized energy; per-completed-round energy, matched common-horizon energy, and completion fraction were comparability/sensitivity outcomes. With five non-zero paired fold differences, the smallest attainable exact two-sided pairwise p-value is 0.0625, so direction, magnitude, win–tie–loss consistency, and effect sizes are reported alongside p-values. Energy differences were sign-oriented so that positive benefit always favors the first method.

For CICIoT2023, the experimental seed was the stochastic replication unit, and the two α conditions were repeated conditions averaged within each seed for the primary scenario-level analysis. Thus, five seed-level matched blocks—not ten α-seed combinations—entered the confirmatory comparisons. Exact paired tests, omnibus randomization, and seed-cluster bootstrap intervals preserved that structure. Analyses stratified by α were exploratory. All five seeds share the same frozen train/validation/test split, so the resulting inference characterizes stochastic variability conditional on that split and must not be interpreted as population-level replication across independent source files, physical devices, sessions, or attack episodes. No inferential effect was pooled across PAMAP2 and CICIoT2023.

4.12. Computing Environment, Reproducibility, and Scientific-Provenance Controls

The overhead benchmark in Section 5.8 was executed separately from the frozen campaigns on the same Windows 11 host using Python 3.14.0 and PyTorch 2.9.0+cpu with 16 PyTorch CPU threads. Its outputs were isolated from the scientific campaign directories and created no new predictive-effectiveness endpoint.

Reproducibility controls included deterministic seed families, frozen manifests and partitions, fixed malicious identities, frozen participation schedules where applicable, run-level identities/contracts, result-binding hashes, completion-state audits, and an exact dependency lock. Detailed package inventories, optimizer-step ledgers, runner/build hashes, and post-run verification records are retained in the versioned archive.

The CICIoT2023 endpoint-sensitivity analysis was versioned separately. The E0 checkpoint rule was reconstructed exactly for 180/180 runs from the stored validation trajectories. The E1 dataset was sealed for all 180 final-round checkpoints. The E3 dataset was sealed for all 120 attacked runs from 50 preserved attack-period BEST_STATE checkpoints and 70 deterministic target-round replays; validation-prefix equality was exact for all 70 replayed runs. Each E1/E3 test evaluation followed the endpoint definition fixed before its result was inspected.

The PAMAP2 sensitivity block was also versioned separately and preserved the frozen primary campaign byte-for-byte. A preliminary exact-replay canary showed that fresh XPU execution was not trajectory-bitwise reproducible even when the source/configuration, round-1 selected-client identities, data windows, and local seeds matched: tiny round-1 local-training numerical differences preceded later trust/risk and client-selection divergence. Historical byte equality was therefore not used as a reproducibility criterion for the sensitivity analysis. Instead, every non-central sensitivity run was paired with a contemporaneous central control executed in the same matched sensitivity block and environment. All 100/100 sensitivity runs completed, the predeclared settings were unchanged, outcome tuning was not performed, and the primary outputs remained unchanged.

The resource-model sensitivity campaign used the same data, model, and attack protocol across all 120 runs. A two-run parity check verified 22/22 predefined consistency criteria before full execution. All 120 runs completed according to the specified design. A subsequent read-only reconstruction reproduced the sensitivity energy values for all runs

(maximum absolute error 9.237×10^{-14}), required no additional training or test inference, and confirmed the integrity of the primary 600-run PAMAP2 results.

Scientific campaigns were executed on 64-bit Windows 11 Pro using an Intel Core Ultra 9 185H processor (Intel Corporation, Santa Clara, CA, USA; 16 physical/22 logical processors), 32 GB RAM, and integrated Intel Arc Graphics. The workflow used Python 3.14.0. PAMAP2 training used PyTorch 2.12.1+xpu on the available Intel XPU backend; CICIoT2023 was deliberately frozen to deterministic CPU execution. CUDA was unavailable and was not used.

5. Results

5.1. Experimental Completion and Result Integrity

All planned campaigns were completed: 600 primary PAMAP2 runs, 240 PAMAP2 component-ablation runs, 180 CICIoT2023 runs, 20 CICIoT2023 FNCTM controls, 20 CICIoT2023 coordinate-wise Median controls, 100 PAMAP2 mechanism-sensitivity runs, and 120 PAMAP2 resource-model-sensitivity runs (1280 scientific executions total). The focused blocks did not modify any primary output. The execution count is a reproducibility/workload total, not an inferential sample size: the primary PAMAP2 and CICIoT2023 analyses each contain five matched inferential blocks as defined in Section 4.11.

Twenty-eight PAMAP2 runs ended before round 100 because the protocol-defined federation-viability threshold was crossed; these were retained as scientific outcomes rather than technical failures. Round 100 was reached in 87/100 TEA-FL runs, 95/100 runs for each common-schedule baseline, and 100/100 runs for FedLE-adapted and ARL-FL. TEA-FL early stops occurred between rounds 92 and 99 (median 96), while the early-stopped common-schedule baseline runs ended at round 97.

5.2. Overall PAMAP2 Effectiveness and Sustainability

Table 5a,b summarize the five held-out-subject fold blocks after averaging the two federated seeds within each fold and giving equal weight to the ten α–scenario conditions.

The decomposition shows that computation was the largest component for every method (70.83–74.10% of total primary-model energy), so part of TEA-FL's lower accounting total is associated with less local work. Across all primary runs, TEA-FL processed 297,620 examples and 5040.33 optimizer steps on average, compared with 351,056.90 examples and 5869.55 steps for FedAvg; ARL-FL lay between them at 326,140.67 examples and 5490.84 steps. This workload difference is an operational consequence of method-specific participation rather than an unmodeled adjustment, but it means that energy savings cannot be interpreted independently of selected-client workload.

The 120-run dynamic resource sensitivity then tested whether the ordering survived when the cost assumptions changed. In the clean $\alpha = 0.1$ reference cell, the primary/computation-dominant regime produced Macro-F1/energy pairs of 0.6322/17.4755 for FedAvg, 0.3586/14.4826 for TEA-FL, and 0.5428/16.7266 for ARL-FL. Reversing the communication/computation coefficients increased TEA-FL energy to 16.3863, while ARL-FL used 17.4477; the corresponding Macro-F1 values were 0.4276 and 0.5929. Under strong sign flip with $\alpha = 1.0$, TEA-FL changed from 0.6748 Macro-F1 and 15.5678 energy in the primary regime to 0.6153 and 16.4253 when communication dominated. Thus, TEA-FL's savings persisted but narrowed, showing that they are not solely a mechanical consequence of sample-count-scaled computation and are also not invariant to the resource model.

Table 5. (**a**) Overall PAMAP2 predictive, residual-energy, retention, and participation results. Values are mean ± standard deviation across five outer-fold blocks after averaging the two federated seeds within each fold and giving equal weight to the ten α–scenario conditions. (**b**) Overall PAMAP2 operating-horizon and energy-comparability results. Values are mean ± standard deviation across five outer-fold blocks after averaging the two federated seeds within each fold and giving equal weight to the ten α–scenario conditions, except the corrected FedLE common-horizon energy, reported as the exact corrected mean including the one-time preflight. Lower energy is preferable; higher common-horizon Macro-F1 and completed-round fraction are preferable. (**c**) Exact decomposition of primary PAMAP2 normalized energy and corresponding scientific workload. Energy values are method-level means over 100 primary runs per method. Processed examples and optimizer steps include the FedLE preflight where applicable; the FedLE preflight subset accounts for 0.586291 mean normalized energy, 12,810 processed examples, and 213.60 optimizer steps.

(**a**)

Method	Macro-F1	Balanced Accuracy	Final Mean Residual Energy	Final Active Clients	Jain Participation Fairness
FedAvg	0.3707 ± 0.0355	0.4156 ± 0.0291	0.2243 ± 0.0177	18.05 ± 2.09	0.9695 ± 0.0060
FedProx	0.3766 ± 0.0376	0.4196 ± 0.0308	0.2243 ± 0.0177	18.05 ± 2.09	0.9695 ± 0.0060
Random Trimmed Mean	0.4187 ± 0.0525	0.4667 ± 0.0434	0.2243 ± 0.0177	18.05 ± 2.09	0.9695 ± 0.0060
FedLE-adapted	0.3482 ± 0.0234	0.3969 ± 0.0206	0.2209 ± 0.0212	20.52 ± 1.95	0.9234 ± 0.0067
TEA-FL	**0.5023 ± 0.0355**	**0.5538 ± 0.0300**	**0.3010 ± 0.0192**	16.17 ± 0.72	0.7019 ± 0.0272
ARL-FL	0.4376 ± 0.0465	0.4847 ± 0.0397	0.2560 ± 0.0207	**23.30 ± 0.95**	0.8660 ± 0.0095

(**b**)

Method	Completed Training Rounds	Runs Reaching Round 100	Protocol-Total Energy	Energy per Completed Round	Common-Horizon Energy	Common-Horizon Macro-F1
FedAvg	99.85 ± 0.34	95/100	17.4113 ± 0.2508	0.17439 ± 0.00292	17.2941 ± 0.2127	0.3700 ± 0.0344
FedProx	99.85 ± 0.34	95/100	17.4113 ± 0.2508	0.17439 ± 0.00292	17.2941 ± 0.2127	0.3757 ± 0.0372
Random Trimmed Mean	99.85 ± 0.34	95/100	17.4113 ± 0.2508	0.17439 ± 0.00292	17.2941 ± 0.2127	0.4159 ± 0.0511
FedLE-adapted	100.00 ± 0.00	100/100	17.5076 ± 0.3300	0.17508 ± 0.00330	17.3681 †	0.3491 ± 0.0254
TEA-FL	99.44 ± 0.69	87/100	**15.2636 ± 0.2997**	**0.15356 ± 0.00354**	**15.2046 ± 0.2738**	**0.5153 ± 0.0372**
ARL-FL	**100.00 ± 0.00**	**100/100**	16.5246 ± 0.3108	0.16525 ± 0.00311	16.3753 ± 0.2207	0.4337 ± 0.0445

(**c**)

Method	Total	Standby	Comm.	Compute	Processed Examples	Optimizer Steps
FedAvg	17.4113	1.3128	3.1952	12.9033	351,056.90	5869.55
FedProx	17.4113	1.3128	3.1952	12.9033	351,056.90	5869.55
Random Trimmed Mean	17.4113	1.3128	3.1952	12.9033	351,056.90	5869.55
FedLE-adapted	17.5076	1.3487	3.3120	12.8470	350,139.36	5874.77
TEA-FL	15.2636	1.2560	3.1821	10.8256	297,620.00	5040.33
ARL-FL	16.5246	1.3707	3.2000	11.9539	326,140.67	5490.84

† For FedLE-adapted, the recorded round-0 residual-energy state is after its one-time preflight. The former round-0-to-horizon mean (16.7818) omitted 0.586291 mean preflight energy; 17.3681 is the corrected initial-to-common-horizon mean used for final comparative interpretation.

The dropout-threshold sensitivity further qualified the sustainability claim. TEA-FL strong-sign-flip Macro-F1 changed from 0.6748 at $E_{\text{crit}} = 0.10$ to 0.5307 at 0.05 and 0.3002 at 0.20. Under the stricter $E_{\text{crit}} = 0.20$ clean $\alpha = 0.1$ condition, ARL-FL mean energy was 16.4074 versus 16.2923 for FedAvg, so ARL-FL's lower-energy ordering is not universal. With five held-out-subject folds per sensitivity cell, the minimum exact two-sided sign-test p-value for five non-zero concordant differences is 0.0625; these resource-model results are therefore interpreted descriptively rather than as confirmatory tuning.

TEA-FL obtained the highest overall Macro-F1 (0.5023 ± 0.0355), whereas ARL-FL ranked second (0.4376 ± 0.0465) and retained the largest final active-client population (23.30 ± 0.95). Relative to FedAvg, the mean Macro-F1 benefits were 0.1315 for TEA-FL and 0.0669 for ARL-FL; both favored the proposed method in all five folds, but the exact two-sided Wilcoxon p-value was 0.0625 and therefore did not cross the 5% threshold.

The six-method Macro-F1 difference was pronounced (Friedman Q = 22.9429, Kendall's W = 0.9177). TEA-FL exceeded all four baselines in every fold, and ARL-FL did likewise; direction and effect magnitude are therefore emphasized rather than dichotomous pairwise significance.

Under the frozen primary resource model, TEA-FL used 15.2636 ± 0.2997 normalized energy units at the protocol endpoint, 0.15356 ± 0.00354 per completed round, and 15.2046 ± 0.2738 at the matched common horizon, corresponding to reductions of 12.3%, 11.9%, and 12.1% relative to FedAvg. ARL-FL achieved corresponding primary-model reductions of 5.1%, 5.2%, and 5.3%. Each primary-model ordering favored the proposed method over every baseline in all five folds, with exact two-sided p = 0.0625. The corrected FedLE initial-to-common-horizon mean is 17.3681 after including its one-time preflight. The dynamic resource sensitivity shows that these energy advantages are conditional on the simulated cost/eligibility regime rather than universal physical-device properties.

At the matched common horizon, TEA-FL retained the highest Macro-F1 (0.5153 ± 0.0372) with the lowest energy (15.2046 ± 0.2738), while ARL-FL combined the second-highest Macro-F1 (0.4337 ± 0.0445) with complete nominal-horizon viability. TEA-FL reached round 100 in 87/100 runs, whereas ARL-FL completed all 100 runs; both methods were non-dominated when effectiveness, energy, and completed-round fraction were considered jointly.

Operationally, ARL-FL retained 23.30 ± 0.95 active clients versus 16.17 ± 0.72 for TEA-FL and 18.05 ± 2.09 for the common-schedule baselines. Jain's participation-fairness index was 0.7019 ± 0.0272 for TEA-FL and 0.8660 ± 0.0095 for ARL-FL, versus 0.9695 ± 0.0060 for the common-schedule baselines, confirming that energy efficiency, retention, and equal participation are distinct objectives.

These overall effectiveness–energy trade-offs at the protocol endpoint and matched common horizon are visualized in Figure 2.

5.3. Effects of Statistical Heterogeneity and Attack Type

Table 6 presents the mean Macro-F1 for each combination of Dirichlet heterogeneity and experimental scenario.

Greater statistical heterogeneity (α = 0.1) generally reduced clean and label-flip performance. FedAvg and FedProx remained strongest in several clean and label-flip settings, showing that the proposed mechanisms were not universally superior under benign or semantic-label corruption.

The clearest separation occurred under sign flip. At μ = 0.2, TEA-FL ranked first for both α levels and ARL-FL ranked second; at α = 1.0, FedAvg, FedProx, and FedLE-adapted fell below 0.04. Under μ = 0.4, TEA-FL remained the only method with substantial performance (0.3319 at α = 0.1 and 0.5043 at α = 1.0), whereas ARL-FL fell to 0.0194 and 0.0253. TEA-FL's aggregate advantage therefore arose primarily from sign-flip robustness, while ARL-FL combined moderate-attack robustness with stronger retention but did not withstand the strongest sign-flip condition.

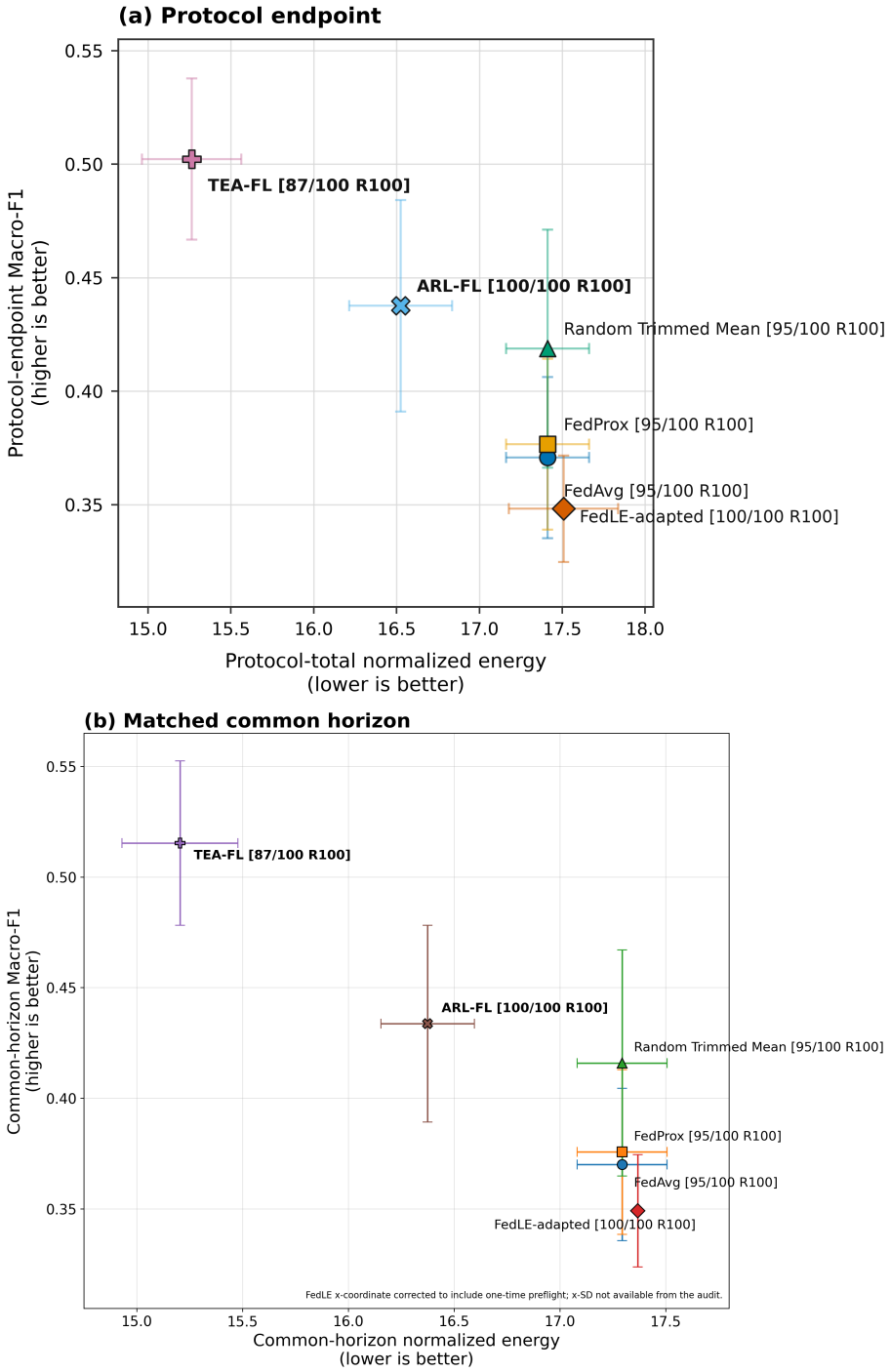


Figure 2. Overall PAMAP2 effectiveness–energy relationships. Points are method means after fold-level aggregation, and error bars are standard deviations across the five outer-fold blocks after averaging the two federated seeds within each fold. (**a**) Protocol-endpoint Macro-F1 versus protocol-total normalized energy under the federation-viability stopping rule. (**b**) Matched common-horizon Macro-F1 versus matched common-horizon cumulative energy. The FedLE x-coordinate in panel (**b**) uses the corrected initial-to-horizon mean (17.3681), including the one-time preflight; a corrected fold-level x-axis standard deviation was not available, so no horizontal FedLE error bar is shown. Labels report nominal-horizon completion. Colors and marker shapes distinguish the six methods, which are also labeled directly in each panel.

Table 6. (**a**) PAMAP2 Macro-F1 by heterogeneity level, attack scenario, and federated method. Values are means across five outer folds and two federated seeds. The highest value in each row is shown in bold. (**b**) Focused PAMAP2 sensitivity under strong sign flip ($\mu = 0.4$). Values are mean ± standard deviation of round-100 test Macro-F1 across five folds. Low/central/high levels are $\lambda = 2.5/5.0/7.5$, TEA-FL $\beta_T = 0.65/0.80/0.95$, and ARL-FL $\tau_P = 0.35/0.50/0.65$.

(**a**)

α	Scenario	FedAvg	FedProx	Random Trimmed Mean	FedLE-Adapted	TEA-FL	ARL-FL
0.1	Clean	0.5975	**0.6256**	0.4794	0.5729	0.4044	0.5427
0.1	Label flip, $\mu = 0.2$	0.5130	**0.5283**	0.4208	0.5275	0.3790	0.4647
0.1	Label flip, $\mu = 0.4$	0.3779	**0.3948**	0.2981	0.2886	0.2724	0.2592
0.1	Sign flip, $\mu = 0.2$	0.0260	0.0311	0.2874	0.0337	**0.4326**	0.3431
0.1	Sign flip, $\mu = 0.4$	0.0153	0.0148	0.0245	0.0129	**0.3319**	0.0194
1.0	Clean	**0.7877**	0.7871	0.7376	0.7523	0.6603	0.7228
1.0	Label flip, $\mu = 0.2$	0.7355	0.7326	**0.7414**	0.7276	0.7124	0.7277
1.0	Label flip, $\mu = 0.4$	0.6011	0.6011	0.5870	0.5329	**0.6149**	0.6033
1.0	Sign flip, $\mu = 0.2$	0.0392	0.0373	0.5972	0.0198	**0.7103**	0.6679
1.0	Sign flip, $\mu = 0.4$	0.0141	0.0133	0.0140	0.0140	**0.5043**	0.0253

(**b**)

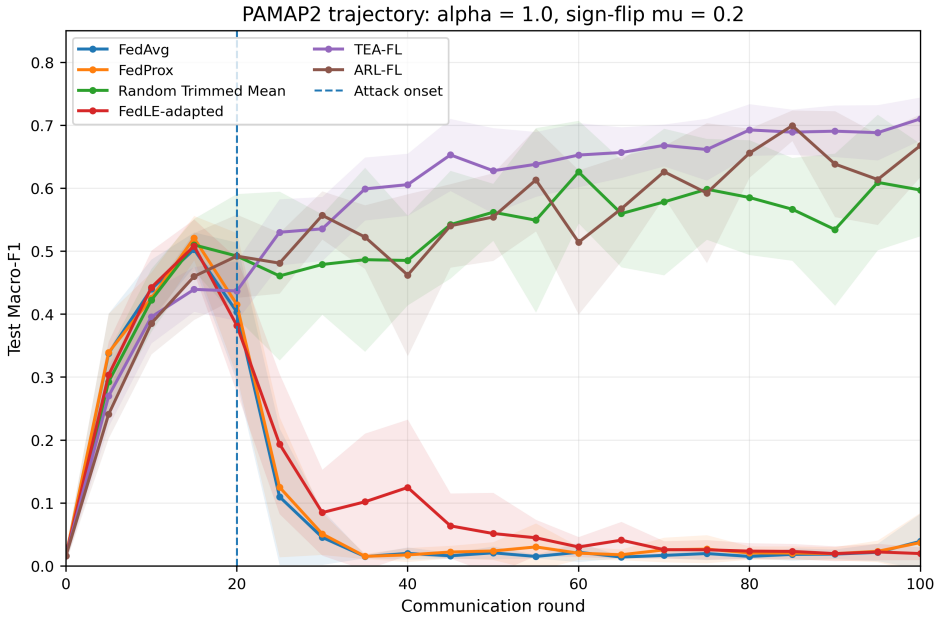
Method	Axis	α	Low	Central	High
TEA-FL	Attack λ	0.1	0.3072 ± 0.1090	0.4030 ± 0.1645	0.1878 ± 0.1890
TEA-FL	Attack λ	1.0	0.6564 ± 0.0998	0.5593 ± 0.2003	0.5697 ± 0.1517
TEA-FL	Trust β_T	0.1	0.3379 ± 0.1902	0.4030 ± 0.1645	0.0728 ± 0.0654
TEA-FL	Trust β_T	1.0	0.5914 ± 0.3299	0.5593 ± 0.2003	0.3456 ± 0.3054
ARL-FL	Attack λ	0.1	0.0706 ± 0.0400	0.0286 ± 0.0189	0.0153 ± 0.0022
ARL-FL	Attack λ	1.0	0.1183 ± 0.0851	0.0179 ± 0.0217	0.0142 ± 0.0049
ARL-FL	Pressure τ_P	0.1	0.0319 ± 0.0293	0.0286 ± 0.0189	0.0210 ± 0.0156
ARL-FL	Pressure τ_P	1.0	0.0355 ± 0.0528	0.0179 ± 0.0217	0.0147 ± 0.0052

5.4. Temporal Response to Sign-Flip Attacks

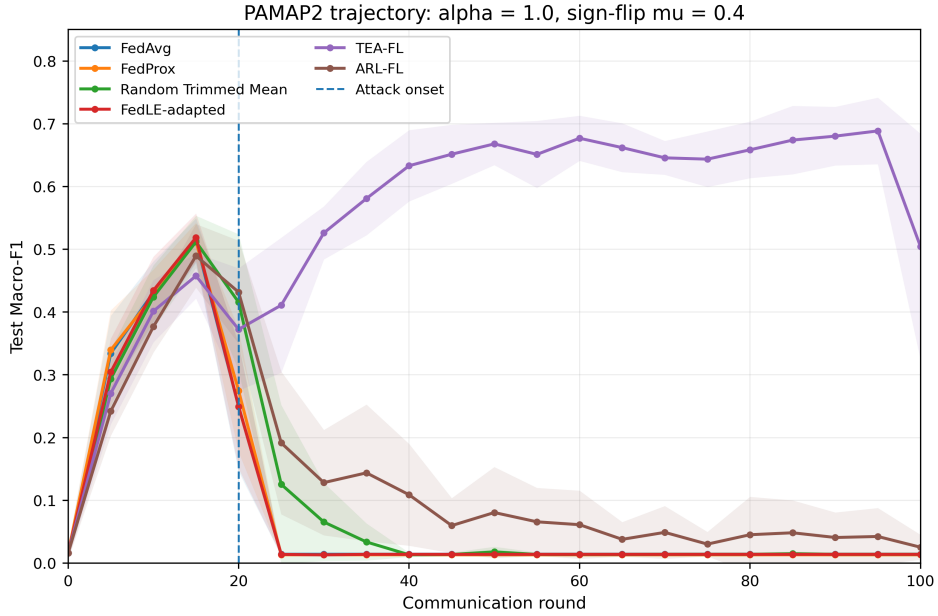
Figure 3 presents the PAMAP2 Macro-F1 trajectories under sign-flip attacks for α = 1.0. The attacks became active at round 20.

Under moderate sign flip, FedAvg fell from 0.5030 at round 15 to 0.1100 at round 25, whereas TEA-FL and ARL-FL reached 0.5302 and 0.4811 at round 25 and finished at 0.7103 and 0.6679. Their mean post-attack nadirs were 0.4210 and 0.3007, followed by recoveries of 0.2893 and 0.3672, respectively.

Under strong sign flip, TEA-FL retained 0.4109 at round 25 and finished at 0.5043, while FedAvg had already collapsed to 0.0141 and ARL-FL declined from 0.1915 at round 25 to 0.0253 at the endpoint. ARL-FL's global-pressure indicator rose from 0.2341 at round 20 to 0.3347 at round 100 (mean maximum 0.4570). The stronger controller regime was not merely theoretical: in the $\alpha = 1.0$, $\mu = 0.4$ condition, 5/10 primary ARL-FL runs crossed 0.5, producing 14 attack-period rounds with pressure above 0.5; all 14 exactly coincided with $k = 2$ trimming. Those rounds selected three malicious clients in 13/14 cases and four in 1/14 (mean 3.071 of eight selected clients), compared with a mean of 2.657 malicious clients across the remaining 796 matched low-pressure attack rounds. Thus, the stronger trimming regime was activated, but only in 1.73% of the 810 attack-period rounds, supporting insufficient activation frequency/capacity rather than complete controller inactivity as the explanation for strong-attack collapse.



(**a**)



(**b**)

Figure 3. Mean PAMAP2 Macro-F1 trajectories under sign-flip attacks with $\alpha = 1.0$. The vertical dashed line indicates attack activation at round 20. Shaded regions are deterministic 95% fold-cluster bootstrap intervals computed after averaging the two federated seeds within each outer fold. (**a**) $\mu = 0.2$; (**b**) $\mu = 0.4$.

5.5. Trust/Risk Diagnostics and Focused Sensitivity

The read-only primary-run diagnostics separated the behavior of known benign and malicious clients. For TEA-FL, the mean benign-minus-malicious trust gap during sign-flip attack periods was 0.4186 and 0.4586 for $\alpha = 0.1$ at $\mu = 0.2$ and 0.4, and 0.4522 and 0.5144 for $\alpha = 1.0$; the corresponding pre-attack gaps were close to zero. Malicious clients also received less aggregation mass than their selection frequency under sign flip: selection share/aggregation share was 0.0552/0.0140 and 0.1506/0.0321 for $\alpha = 0.1$, and 0.0624/0.0109 and 0.1595/0.0304 for $\alpha = 1.0$, for moderate and strong sign flip respectively. The trust signal is therefore discriminative for direct geometric reversal, but not equivalent to maliciousness.

The benign-client diagnostics quantify the latter limitation. At $\alpha = 0.1$, the mean per-run Spearman correlation between benign trust and number of represented classes ranged from -0.452 to -0.499 across the four attacked scenarios, while trust versus normalized class entropy ranged from -0.673 to -0.742. The corresponding relationships were weaker at $\alpha = 1.0$ (class-count correlation -0.200 to -0.262; entropy correlation -0.179 to -0.301). Thus, under stronger non-IID heterogeneity, benign clients with broader class support or higher class entropy could receive lower consistency trust, providing a direct mechanism-level explanation for benign downweighting and the clean-data penalty observed for TEA-FL.

ARL-FL showed a different diagnostic signature. During sign-flip attack periods, the mean malicious-minus-benign risk gap was 0.1738/0.1508 for $\alpha = 0.1$ and 0.1662/0.2391 for $\alpha = 1.0$, for $\mu = 0.2/0.4$ respectively, whereas the pre-attack gaps were near zero. Malicious updates were clipped much more frequently than benign updates: 95.1% versus 0.94% and 54.0% versus 0.38% at $\alpha = 0.1$, and 96.5% versus 0.87% and 66.2% versus 0.14% at $\alpha = 1.0$, for moderate and strong sign flip. These results show that ARL-FL detected and clipped sign-flipped updates, while the pressure audit above shows that escalation to two-value trimming was too infrequent to prevent collapse under the strongest condition.

Table 6b reports the fresh matched sensitivity. Values are round-100 test Macro-F1 across five held-out-subject folds; the sensitivity uses one federated seed and is therefore interpreted descriptively. The contemporaneous central values serve as within-block matched references; Table 6 reports the primary-campaign results.

The sensitivity supports three bounded conclusions. First, ARL-FL degraded as sign-flip magnitude increased: reducing λ from 5.0 to 2.5 improved Macro-F1 in all 5/5 folds at both α values (mean deltas +0.0421 and +0.1004 versus the contemporaneous central control), whereas $\lambda = 7.5$ remained near collapse. Second, TEA-FL was sensitive to excessive trust persistence: $\beta_T = 0.95$ reduced Macro-F1 in all 5/5 folds at both α values (mean deltas -0.3302 and -0.2136), while $\beta_T = 0.65$ had mixed fold-level effects. Third, varying τ_P produced much smaller and fold-inconsistent changes than varying attack magnitude; lowering the switch to 0.35 modestly increased the mean, but did not restore strong-attack robustness. TEA-FL's response to λ was heterogeneity-dependent rather than monotonic. Because each cell contains five independent folds, these results are descriptive: even a 5/5 same-direction effect has a minimum exact two-sided sign-test value of 0.0625 and is not presented as confirmatory parameter tuning.

5.6. Multi-Objective and Activity-Level PAMAP2 Results

The three-objective analysis maximized common-horizon Macro-F1, minimized common-horizon cumulative energy, and maximized completed-round fraction. TEA-FL and ARL-FL were both non-dominated: TEA-FL combined the highest common-horizon Macro-F1 (0.5153) with the lowest energy (15.2046), whereas ARL-FL combined the second-highest Macro-F1 (0.4337) and second-lowest energy (16.3753) with complete nominal-horizon viability.

At activity level, TEA-FL obtained the highest mean F1 for 10 of 12 activities; Random Trimmed Mean led sitting and FedProx led rope jumping. Across methods, rope jumping, standing, and sitting were the most difficult activities, while cycling, walking, and running had the highest mean F1 values. Figure 4 provides the complete activity–method heatmap.

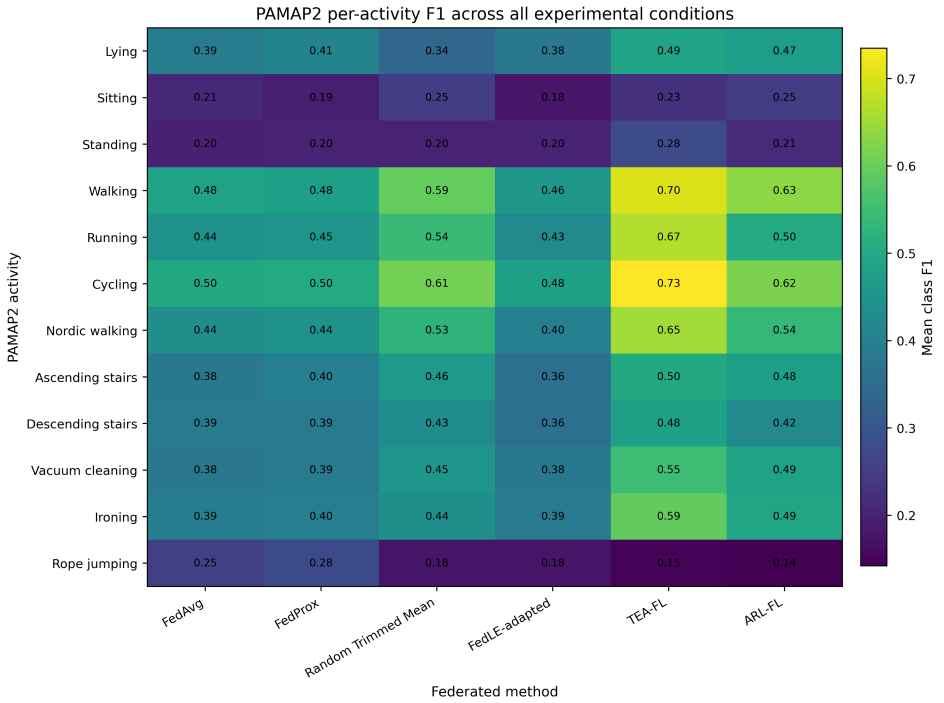


Figure 4. PAMAP2 mean per-activity F1 scores for the six federated methods across all heterogeneity and scenario conditions.

5.7. Component-Ablation Results

The component-ablation campaign isolated the trust and energy mechanisms of TEA-FL and the energy-aware selection and adaptive-pressure mechanisms of ARL-FL.

Table 7 summarizes the average effect of restoring each component to the corresponding incomplete method. A positive energy benefit indicates that the complete method consumed less normalized energy than its ablated counterpart.

Table 7. Summary of the PAMAP2 component-ablation results across the six prespecified ablation conditions. Effects are calculated as complete method minus ablated variant, with the sign reversed for energy so that a positive value denotes lower consumption by the complete method.

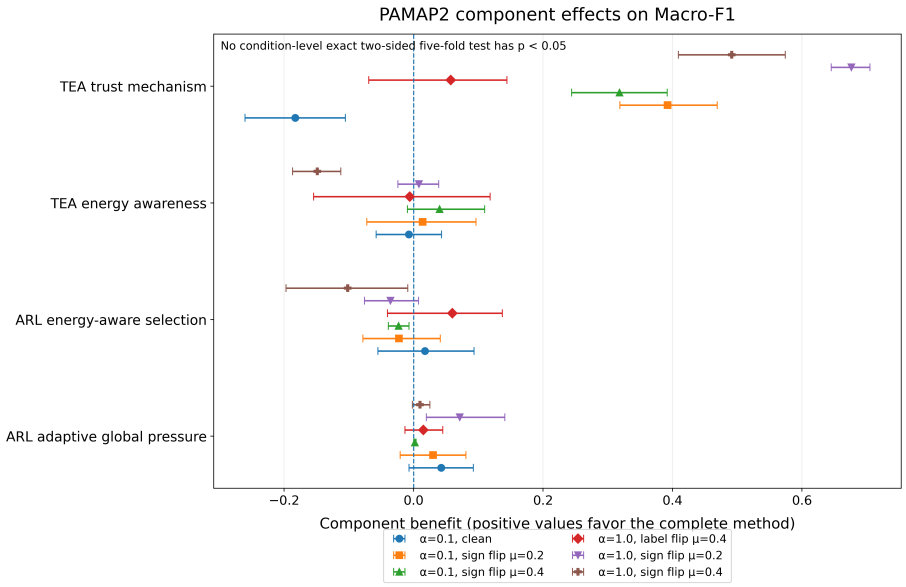
Restored Component	Mean Macro-F1 Effect	Conditions with Positive Mean Macro-F1 Effect	Mean Energy Benefit	Conditions with Positive Mean Energy Benefit	Mean Final Active-Client Effect	Main Observed Role
TEA-FL trust mechanism	+0.2921	5/6	+0.9958	6/6	−12.33	Sign-flip robustness and lower protocol-total energy, with reduced retention
TEA-FL energy awareness	−0.0166	3/6	−0.1466	2/6	+0.55	Modest participation and fairness adjustment
ARL-FL energy-aware selection	−0.0177	2/6	+0.5942	6/6	+4.12	Energy reduction and client retention
ARL-FL adaptive global pressure	+0.0285	6/6	+0.0801	6/6	+0.40	Modest adaptive operational control

The TEA-FL trust mechanism was the dominant predictive component: restoring trust produced large positive Macro-F1 effects in all four sign-flip conditions and favored the complete method in all five folds within each condition. Across the six prespecified conditions, trust increased mean Macro-F1 in five, reduced energy in all six, and reduced final active-client retention by 12.33 clients on average. The five-fold exact tests nevertheless could not attain $p < 0.05$; the minimum two-sided p-value for five non-zero same-direction differences is 0.0625.

TEA-FL energy awareness had a mean Macro-F1 effect of −0.0166, a mean energy benefit of −0.1466, and a +0.55 active-client effect, indicating a modest participation/fairness role rather than the primary predictive defense. ARL-FL energy-aware selection reduced energy in all six conditions (mean benefit 0.5942) and retained more clients in all six (mean +4.12), while its mean Macro-F1 effect was −0.0177 and positive in two conditions. ARL-FL adaptive pressure produced smaller positive mean effects on Macro-F1 (+0.0285), energy (+0.0801), and active-client count (+0.40).

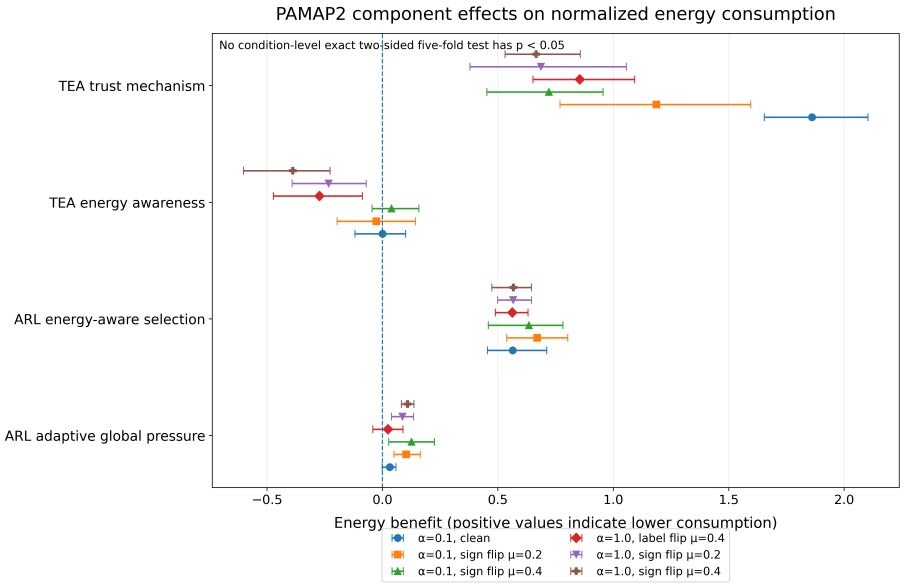
A matched-common-horizon audit showed that the main ablation energy conclusions were not artifacts of unequal retention horizons. Relative to the complete control, arl_no_energy consumed +0.6011 normalized units at the matched horizon and was higher in 60/60 pairs, while retaining 4.133 fewer active clients on average at that horizon. arl_no_pressure was +0.0801 energy units (46 higher, 14 lower pairs). tea_energy_only was +0.8777 units (56 higher, 4 lower), whereas tea_trust_only was −0.1019 units (40 lower, 20 higher). These horizon-matched results reinforce the attribution of ARL-FL's sustainability benefit to energy-aware selection and show that TEA-FL's component effects involve a genuine energy–retention trade-off rather than a simple early-stopping artifact.

No condition-level ablation contrast attained an exact two-sided p-value below 0.05. Figure 5 therefore presents fold-level paired effect estimates with deterministic 95% fold-cluster bootstrap intervals as estimation-uncertainty summaries; it does not use bootstrap-interval exclusion of zero as a substitute for the declared exact-test significance criterion.



(**a**)

Figure 5. *Cont.*



(**b**)

Figure 5. Paired PAMAP2 component effects for TEA-FL and ARL-FL across the six prespecified ablation conditions. Symbols distinguish the α–scenario conditions, and horizontal bars are deterministic 95% fold-cluster bootstrap intervals after averaging the two federated seeds within each outer fold. No condition-level exact two-sided five-fold test attained $p < 0.05$; bootstrap-interval exclusion of zero is not used as the significance criterion. The vertical dashed reference line at zero denotes no component effect relative to the corresponding ablated variant. Positive values favor the complete method; for energy, positive values indicate lower normalized consumption. (**a**) Macro-F1; (**b**) normalized energy.

5.8. CICIoT2023 Cross-Domain Robustness Assessment

Table 8a presents the original-primary E0 CICIoT2023 test Macro-F1 results. Each cell summarizes five experimental seeds.

Table 8. (**a**) CICIoT2023 original-primary (E0) test Macro-F1 by heterogeneity level, scenario, and federated method. E0 is validation-selected over rounds 1–20. Values are mean ± standard deviation across five experimental seeds. (**b**) CICIoT2023 checkpoint-policy sensitivity for attacked runs. E0 is the original validation-selected endpoint over rounds 1–20; E1 is the fixed final model at round 20; E3 is validation-selected only over attack-active rounds 5–20. Values are mean test Macro-F1 across ten runs per method/scenario (five seeds × two α conditions); no test metric was used for checkpoint selection. (**c**) Focused CICIoT2023 comparison of Random Trimmed Mean (RTM), fixed norm-clipped trimmed mean (FNCTM), and ARL-FL. Values are mean Macro-F1 across five experimental seeds for each scenario–α cell. E0 is validation-selected over rounds 1–20, E1 is fixed round 20, and E3 is validation-selected over attack-active rounds 5–20. The focused control was predefined with $\kappa = 2.0$ and k = 2; no test outcome was used to tune parameters or select checkpoints. (**d**) Focused CICIoT2023 E0 comparison of Random Trimmed Mean (RTM), coordinate-wise Median, and ARL-FL. Values are mean test Macro-F1 across five experimental seeds. E0 is selected by validation Macro-F1 over rounds 1–20; no test outcome was used for checkpoint selection.

(**a**)

α	Method	Clean	Sign Flip, μ = 0.2	Sign Flip, μ = 0.4
0.1	FedAvg	0.3669 ± 0.0358	0.1726 ± 0.0554	0.1430 ± 0.0410
0.1	FedProx	**0.4021 ± 0.0667**	0.1838 ± 0.0714	0.1398 ± 0.0344
0.1	Random Trimmed Mean	0.3450 ± 0.0359	**0.2764 ± 0.1001**	**0.1557 ± 0.0515**
0.1	FedLE-adapted †	0.3651 ± 0.0339	0.1676 ± 0.0591	0.1507 ± 0.0405
0.1	TEA-FL	0.2217 ± 0.0805	0.2760 ± 0.0982	0.1507 ± 0.0586
0.1	ARL-FL	0.3717 ± 0.0362	0.2573 ± 0.1058	0.1533 ± 0.0568

Table 8. *Cont.*

(**a**)

α	Method	Clean	Sign Flip, μ = 0.2	Sign Flip, μ = 0.4
1.0	FedAvg	0.6153 ± 0.0130	0.5495 ± 0.0100	0.5491 ± 0.0099
1.0	FedProx	**0.6190 ± 0.0059**	0.5315 ± 0.0100	0.5306 ± 0.0095
1.0	Random Trimmed Mean	0.6140 ± 0.0206	0.5807 ± 0.0174	0.5491 ± 0.0137
1.0	FedLE-adapted †	0.6156 ± 0.0131	0.5498 ± 0.0099	0.5492 ± 0.0097
1.0	TEA-FL	0.5988 ± 0.0245	0.5847 ± 0.0231	**0.5553 ± 0.0163**
1.0	ARL-FL	0.6158 ± 0.0158	**0.5857 ± 0.0227**	0.5499 ± 0.0128

(**b**)

Method	Attack	E0 Primary	E1 Final R20	E3 Post-Attack
FedAvg	μ = 0.2	0.3610	0.1285	0.3517
FedAvg	μ = 0.4	0.3460	0.0280	0.2858
FedProx	μ = 0.2	0.3577	0.1455	0.3472
FedProx	μ = 0.4	0.3352	0.0421	0.2710
Random Trimmed Mean	μ = 0.2	0.4285	0.3725	0.4285
Random Trimmed Mean	μ = 0.4	0.3524	0.0296	0.3393
FedLE-adapted	μ = 0.2	0.3587	0.1198	0.3488
FedLE-adapted	μ = 0.4	0.3499	0.0295	0.2892
TEA-FL	μ = 0.2	0.4303	0.3764	0.4303
TEA-FL	μ = 0.4	0.3530	0.1800	0.3168
ARL-FL	μ = 0.2	0.4215	0.3753	0.4215
ARL-FL	μ = 0.4	0.3516	0.1066	0.3420

(**c**)

Condition	Endpoint	RTM	FNCTM	ARL-FL
Clean, α = 0.1	E0	0.3450	0.3394	0.3717
Clean, α = 0.1	E1	0.2956	0.2903	0.3093
Clean, α = 1.0	E0	0.6140	0.6129	0.6158
Clean, α = 1.0	E1	0.6034	0.6046	0.6048
Sign flip μ = 0.4, α = 0.1	E0	0.1557	0.1662	0.1533
Sign flip μ = 0.4, α = 0.1	E1	0.0329	0.0899	0.0605
Sign flip μ = 0.4, α = 0.1	E3	0.1557	0.1662	0.1533
Sign flip μ = 0.4, α = 1.0	E0	0.5491	0.5497	0.5499
Sign flip μ = 0.4, α = 1.0	E1	0.0263	0.3648	0.1527
Sign flip μ = 0.4, α = 1.0	E3	0.5229	0.5417	0.5307

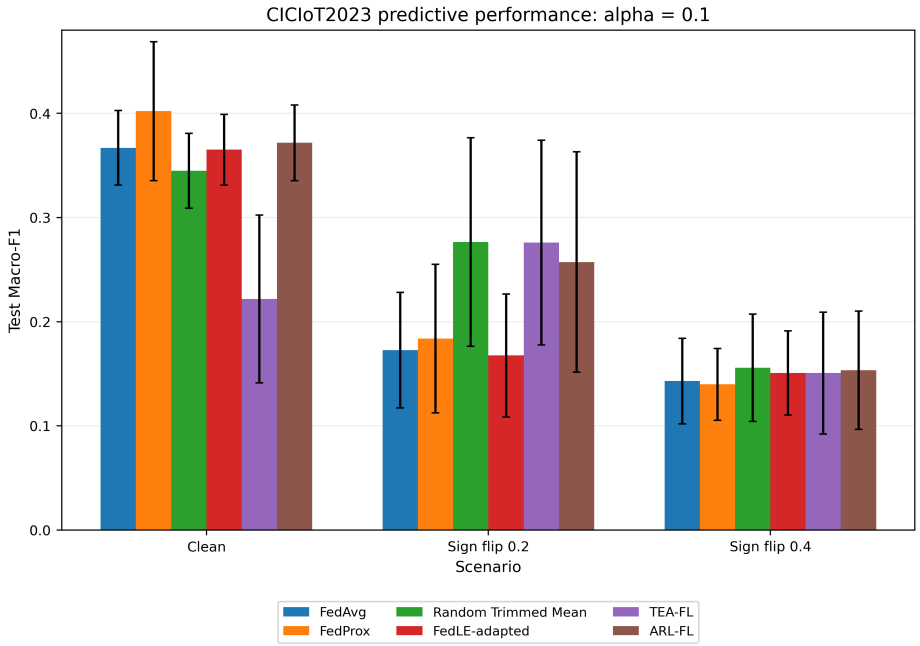
(**d**)

Condition	RTM	Median	ARL-FL
Clean, α = 0.1	0.3450	0.3042	0.3717
Clean, α = 1.0	0.6140	0.6121	0.6158
Sign flip μ = 0.4, α = 0.1	0.1557	0.1949	0.1533
Sign flip μ = 0.4, α = 1.0	0.5491	0.5605	0.5499

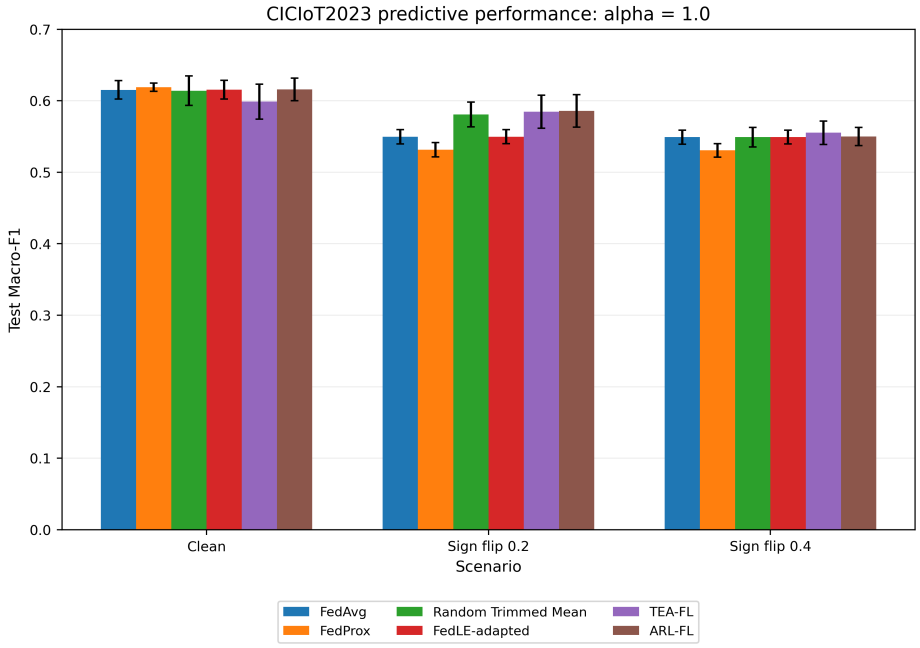
† The FedLE-adapted row is retained from the frozen historical campaign for provenance. Under the frozen CICIoT2023 participation schedule, it is algorithmically equivalent to FedAvg; a 30-pair deterministic replay using the same aggregation path produced exact equality of the full validation trajectories, selected states, round-20 states, and test metrics. Historical small numerical differences therefore do not constitute independent algorithmic evidence.

Figure 6 summarizes the E0 results graphically; Figure 7 reports matched-clean retention under the same E0 checkpoint policy.

For the E0 primary endpoint, inference used five experimental-seed blocks after equally averaging the two α-specific outcomes within each seed. In the clean scenario, ARL-FL exceeded TEA-FL by 0.0835 Macro-F1 (95% seed-cluster bootstrap CI 0.0456–0.1214) in all five seeds; the exact two-sided p-value was 0.0625 and the Holm-adjusted value was 0.3125.



(**a**)



(**b**)

Figure 6. CICIoT2023 original-primary E0 mean test Macro-F1 under clean, moderate sign-flip, and strong sign-flip conditions. E0 selects the best validation checkpoint over rounds 1–20. Error bars are standard deviations across five experimental seeds. (**a**) $\alpha = 0.1$; (**b**) $\alpha = 1.0$.

Under moderate sign flip ($\mu = 0.2$), the E0 endpoint showed ARL-FL exceeding the FedAvg/FedLE-equivalent control by approximately 0.06 Macro-F1 and FedProx by 0.0638, with the same direction in all five seeds (exact two-sided $p = 0.0625$ for each comparison). It did not show a consistent advantage over Random Trimmed Mean or TEA-FL. The historical FedLE-adapted row is not counted as a separate corroborating algorithm.

Under strong sign flip ($\mu = 0.4$), the E0 omnibus difference was not significant. ARL-FL exceeded FedProx in all five seeds by 0.0164, whereas its differences from the other methods were small and mixed. No planned E0 CICIoT2023 pairwise contrast crossed the 5% significance threshold.

These comparisons remain conditional on the single frozen observation-level split and therefore quantify stochastic variability under that design rather than population-level replication across independent source files, devices, sessions, or attack episodes.

For E0, heterogeneity remained influential. At $\alpha = 0.1$, FedProx led the clean setting and Random Trimmed Mean led both sign-flip settings; at $\alpha = 1.0$, ARL-FL and TEA-FL led moderate sign flip and TEA-FL led strong sign flip. At $\alpha = 1.0$, TEA-FL retained 97.7% and 92.8% of its matched-clean E0 Macro-F1 under moderate and strong sign flip, respectively; ARL-FL retained 95.1% and 89.3%. A recurrent Macro-F1 floor of approximately 0.1245 occurred in 30/180 technically complete runs and was retained in all analyses.

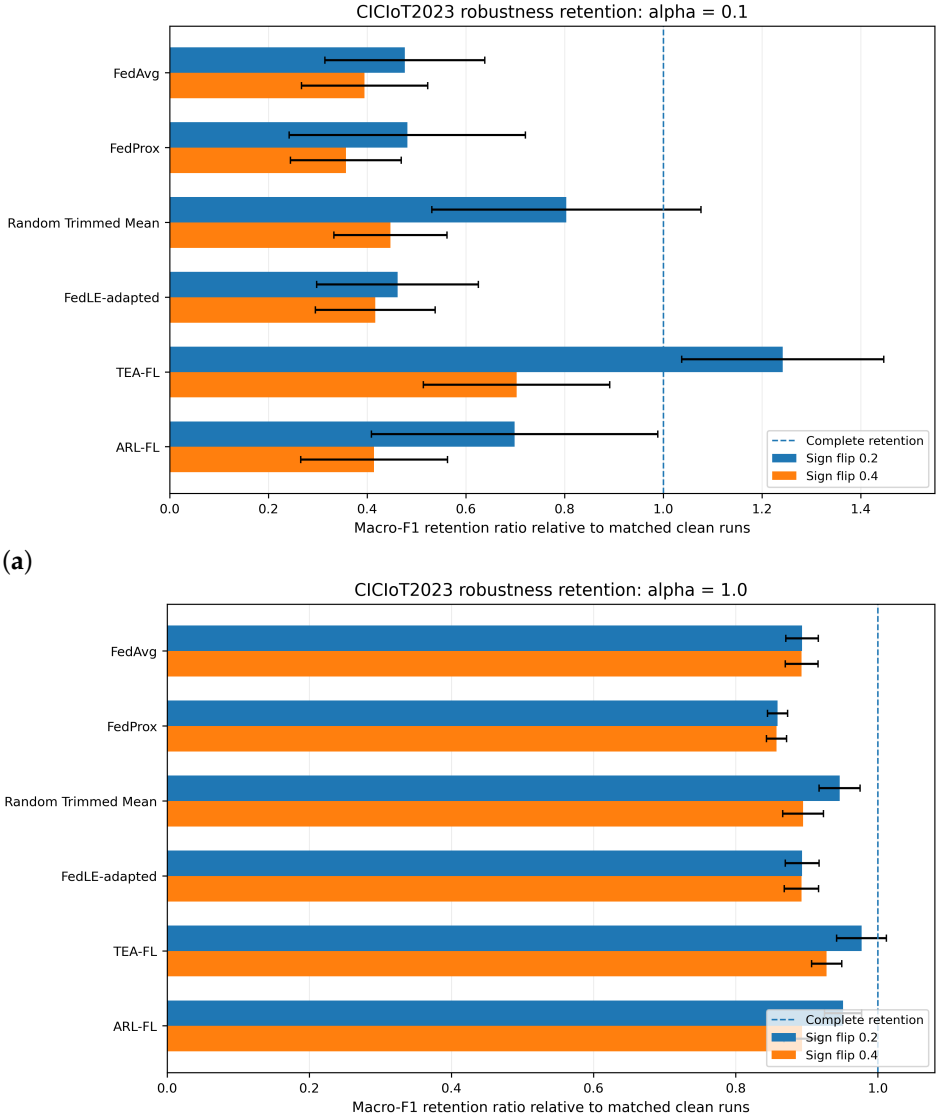


Figure 7. CICIoT2023 original-primary E0 Macro-F1 retention relative to matched clean E0 runs. Bars are means and error bars are standard deviations across five experimental seeds; one denotes complete matched-clean retention. (**a**) $\alpha = 0.1$; (**b**) $\alpha = 1.0$.

The checkpoint audit materially qualifies this E0 interpretation. Among the 120 attacked runs, 70 (58.3%) selected an E0 checkpoint from the clean warm-up rounds 1–4: 21/60 (35.0%) under moderate sign flip and 49/60 (81.7%) under strong sign flip. The exact selected-round distributions were R1:9, R3:2, R4:10, R5:3, R6:2, R7:3, R9:1, R10:1, R11:3, R14:3, R15:8, R16:1, R17:3, R18:5, R19:3, R20:3 for $\mu = 0.2$, and R1:20, R3:4, R4:25, R5:4, R6:2, R8:4, R20:1 for $\mu = 0.4$.

Table 8b therefore compares E0 with the fixed final-round E1 endpoint and the validation-selected post-attack E3 endpoint. Across all 120 attacked runs, mean Macro-F1 changed by −0.20935 from E0 to E1 and by −0.02281 from E0 to E3; E3 exceeded E1 by 0.18653, recovering 89.1% of the mean final-round penalty relative to E0. Method leadership changed in 2/4 scenario × α strata between E0 and E1 and in 1/4 between E0 and E3. Thus, the original primary endpoint provides substantial evidence of checkpoint protection, whereas sustained final-round learning is weaker; post-attack validation selection recovers most, but not all, of that loss.

The endpoint distinction is especially important under strong sign flip with α = 1.0: TEA-FL led E0 (0.5553) and E1 (0.3235), whereas ARL-FL led E3 (0.5307 versus 0.4829 for TEA-FL). Accordingly, cross-domain robustness claims below distinguish preservation of an earlier validation-selected state from the ability to maintain or recover performance after attack onset.

The focused FNCTM comparison further tests whether ARL-FL's adaptive risk/pressure controller adds predictive robustness beyond a simpler fixed clipping-plus-trimming rule. Clipping was active in the fixed control: the mean number of clipped client-events per run was 13.4 and 12.8 in the two clean cells, compared with 40.6 and 40.4 under strong sign flip; clipping occurred in 5.0, 10.0, 15.6, and 13.4 of 20 rounds, respectively.

Table 8c shows that fixed clipping explains a substantial part of the robustness available from a clipped trimmed mean. In clean conditions, FNCTM remained essentially aligned with Random Trimmed Mean and ARL-FL. Under strong sign flip, however, FNCTM was consistently stronger at the stricter E1/E3 endpoints: at α = 0.1 its E1 Macro-F1 was 0.0899 versus 0.0329 for Random Trimmed Mean and 0.0605 for ARL-FL; at α = 1.0 it reached 0.3648 versus 0.0263 and 0.1527. For E3 at α = 1.0, FNCTM reached 0.5417 versus 0.5229 for Random Trimmed Mean and 0.5307 for ARL-FL. With five seeds per cell these comparisons are descriptive, but they do not support claiming an incremental strong-attack predictive advantage for ARL-FL's adaptive pressure/risk control over the simpler fixed norm-clipped trimmed mean.

A focused CICIoT2023 campaign also evaluated coordinate-wise Median under the same frozen participation schedule, local training, clean/strong-sign-flip conditions, heterogeneity levels, and five experimental seeds used by the FNCTM control. Table 8d reports the validation-selected E0 endpoint. Median reached 0.3042 ± 0.0760 and 0.6121 ± 0.0242 in the two clean cells, and 0.1949 ± 0.1081 and 0.5605 ± 0.0227 under strong sign flip for $\alpha = 0.1$ and $\alpha = 1.0$, respectively.

The matched seed-level comparison is deliberately interpreted descriptively. At $\alpha = 0.1$, Median was lower than both RTM and ARL-FL in all five clean seeds (mean paired differences −0.0408 and −0.0676, respectively; exact two-sided sign $p = 0.0625$), but higher than both comparators in all five strong-sign-flip seeds (+0.0392 and +0.0416; $p = 0.0625$). At $\alpha = 1.0$, the differences were small and mixed. When the two α conditions were averaged within each experimental seed, the strong-sign-flip differences were +0.0253 versus RTM and +0.0261 versus ARL-FL, with 3–2 sign splits and $p = 1.0$. No exact paired sign comparison reached $p < 0.05$. Importantly, E0 selected a pre-attack warm-up checkpoint (rounds 1–4) in 3/5 seeds in each attacked α cell. The attacked E0 advantage therefore supports validation-selected checkpoint preservation, but it is not interpreted as evidence that Median sustained the same advantage after attack onset. The FNCTM focused result refines the role attributed to ARL-FL. Its principal demonstrated advantage remains system-level: energy-aware scheduling and risk/staleness control preserved client availability and federation continuity on PAMAP2, while its adaptive aggregation handled moderate adversarial pressure. For the representative strong CICIoT2023 attack, a fixed

clipping-plus-trimming rule was at least as competitive predictively and was markedly stronger at the final-round endpoint.

5.9. Complete-Round Operational Overhead Benchmark

The CICIoT2023 campaign used an externally frozen participation schedule and therefore could not measure the scheduling cost of the complete TEA-FL and ARL-FL systems. To address this boundary directly, a dedicated phase-resolved benchmark was performed on the complete PAMAP2 round procedure under the canonical outer-fold-1, $\alpha = 1.0$, clean, seed-123 condition. Each method was reset to its canonical round-1 pre-state, followed by one warm-up repetition and five measured repetitions. The benchmark separately timed client selection, local training, trust/risk processing, aggregation, and the post-round evaluation path used by the frozen PAMAP2 engine.

Table 9, Panel A reports the measured phase medians. For each measured repetition, the server-side selection, trust/risk, and aggregation times were also summed before summary statistics were calculated. The resulting median server-side totals were 5.184 ms for FedAvg, 13.639 ms for TEA-FL, and 55.953 ms for ARL-FL; the corresponding means $\pm$ standard deviations across the five measured repetitions were 5.137 $\pm$ 0.414 ms, 13.825 $\pm$ 0.570 ms, and 56.106 $\pm$ 1.760 ms, respectively. Relative to the FedAvg median, the additional measured server-side processing was 8.455 ms per round for TEA-FL and 50.769 ms per round for ARL-FL at the evaluated 28-client scale.

Table 9. Phase-resolved complete-round PAMAP2 overhead benchmark. Panel A reports medians over five measured repetitions after one warm-up repetition under outer fold 1, $\alpha = 1.0$, clean condition, seed 123, round 1. Panel B reports exact accounted tensor storage, and Panel C reports median server-only selection/scheduling time over 50 repetitions per eligible-client population. End-to-end times are descriptive and are not interpreted as algorithmic speed rankings because method-specific scheduling changes the selected local-training workload.

Panel A. Complete-Round Phase Timing

Method	Selection (ms)	Local Training (s)	Trust/Risk (ms)	Aggregation (ms)	Post-Round Eval. (ms)	Complete Round (s)
FedAvg	0.228	1.635	–	4.955	203.688	1.839
TEA-FL	1.590	1.664	7.421	4.528	194.679	1.895
ARL-FL	9.721	1.141	23.162	22.917	171.890	1.373

Panel B. Accounted Server Tensor Storage

Method	Global Model State (MiB)	Current Local States (MiB)	Additional Persistent Tensor State (MiB)
FedAvg	0.294	2.350	0
TEA-FL	0.294	2.350	0
ARL-FL	0.294	2.350	4.700

Panel C. Server-Only Client-Selection/Scheduling Scaling

Eligible Clients	TEA-FL Median (ms)	ARL-FL Median (ms)
28	1.263	8.885
56	0.788	22.642
112	0.845	88.273
224	0.963	352.522

Note: tensor accounting excludes Python objects, allocator overhead, and process working-set memory. TEA-FL additionally retains scalar per-client trust and energy states. The post-round evaluation phase follows the frozen PAMAP2 round engine and is reported as an execution-time component only; it does not create a new scientific effectiveness endpoint.

End-to-end round times were 1.839 s for FedAvg, 1.895 s for TEA-FL, and 1.373 s for ARL-FL. These values are not interpreted as algorithmic speed rankings. Method-specific scheduling selects different clients, and the corresponding median local-training times were 1.635 s, 1.664 s, and 1.141 s, respectively. The lower ARL-FL end-to-end time therefore primarily reflects the workload of the clients selected in this benchmark rather than evidence that ARL-FL is intrinsically faster. The phase-separated server timings provide the appropriate overhead comparison.

Exact tensor accounting, excluding Python/runtime allocator overhead, gave 0.294 MiB for the global model state and 2.350 MiB for the eight returned local model states for all three methods. TEA-FL required no additional persistent tensor state beyond these common tensors, although it retained its per-client scalar trust and energy states. ARL-FL additionally retained 4.700 MiB of previous-update tensors for temporal risk estimation. The corresponding accounted tensor totals were therefore approximately 2.644 MiB for FedAvg and TEA-FL and 7.344 MiB for ARL-FL.

Table 9, Panel C reports a server-only scheduling scaling check. TEA-FL selection remained in the approximately 0.8–1.3 ms median range from 28 to 224 eligible clients, with no monotonic growth resolvable above the low-millisecond timing variability at these sizes. ARL-FL scheduling increased from 8.89 ms at 28 clients to 22.64 ms at 56, 88.27 ms at 112, and 352.52 ms at 224 clients. The approximately 15.6-fold increase from 56 to 224 clients for a fourfold increase in population is consistent with the expected adverse scaling of the non-dominated-sorting stage. Consequently, the measured ARL-FL overhead is modest at the federation size evaluated scientifically, but it should not be extrapolated as negligible for substantially larger eligible-client populations.

The benchmark is descriptive rather than inferential. It was performed after completion of the frozen scientific campaigns and created no new scientific effectiveness endpoint. The original CICIoT2023 wall-time logs remain useful as execution records for the matched frozen-schedule campaign, but they are no longer used as evidence for the overhead of the complete TEA-FL and ARL-FL systems.

6. Discussion

The combined evidence supports complementary, deployment-dependent roles for TEA-FL and ARL-FL rather than a universal method ranking. On PAMAP2, TEA-FL provided the strongest resistance to direct sign-flip update reversal and the lowest simulated energy under the frozen primary resource model, whereas ARL-FL prioritized client retention and federation continuity. The mechanism audits explain this difference: TEA-FL trust strongly separated benign from malicious sign-flip clients and suppressed malicious aggregation influence, while ARL-FL risk and clipping reacted strongly but its higher-pressure two-value trimming regime was activated in only 14 attack-period rounds. Under the strongest malicious exposure, escalation therefore occurred but was too sparse to prevent collapse. The focused FNCTM control further shows that fixed norm clipping plus trimming can outperform ARL-FL at the strong-sign-flip final-round endpoint, so adaptive pressure is not claimed to provide a unique predictive advantage in that regime.

The sustainability findings are similarly conditional. TEA-FL's lower primary-model energy persisted after per-completed-round and matched-common-horizon normalization and remained lower when communication and computation coefficients were reversed, although the margin narrowed. ARL-FL retained the largest active-client population and completed all primary PAMAP2 runs, but its lower-energy ordering did not survive every stricter dropout setting. Energy decomposition and processed-example/optimizer-step accounting show that method-specific client scheduling changes local workload and therefore

contributes directly to the simulated totals. These results support relative comparisons under the declared normalized resource model; they are not physical battery-life measurements.

The cross-domain CICIoT2023 study provides a narrower form of evidence. Frozen participation isolates the active trust/risk and aggregation components, but it cannot validate method-specific scheduling, energy, dropout, retention, or fairness. The E0/E1/E3 analysis also separates checkpoint protection from attack-period learning: many attacked E0 runs selected a clean-warm-up checkpoint, final-round E1 performance was substantially weaker, and post-attack validation selection E3 recovered most but not all of the E1 penalty. FedLE-adapted was verified to be algorithmically equivalent to FedAvg under the frozen-participation protocol. In addition, the source-file audit and final float32 collision audit justify only observation-level generalization conditional on the frozen split; they do not establish unseen-device, unseen-source, session, or attack-episode independence.

Operationally, both proposed methods add server-side processing without adding communication rounds or local-training epochs. At 28 clients, the measured server-side selection/trust-risk/aggregation overhead was modest relative to local training, but ARL-FL was materially more expensive than TEA-FL and retained 4.700 MiB of additional persistent update tensors. The server-only scaling check is therefore important: TEA-FL scheduling remained near the millisecond scale through 224 eligible clients, whereas ARL-FL scheduling increased to hundreds of milliseconds. End-to-end round time is not used as an algorithmic speed ranking because method-specific schedulers select different local workloads.

6.1. Practical Implications

TEA-FL is most attractive when direct update reversal and low aggregate simulated energy are central objectives and concentrated participation is acceptable; ARL-FL is more appropriate when retention and federation continuity under moderate adversarial pressure are prioritized. Conventional FedAvg/FedProx remain competitive in trusted clean or several label-flip settings, and simple fixed robust aggregation can be preferable when resource-aware orchestration is unnecessary. Parameter and resource sensitivities caution against treating the frozen trust persistence, pressure threshold, or energy coefficients as universally optimal. Deployment should therefore begin from the threat model and device/network resource model rather than from a single global ranking.

6.2. Limitations

The main limitations are consolidated here. PAMAP2 used 28 subject-specific virtual clients and normalized simulated energy rather than physical-device power measurements. The threat model covered deterministic label flip and amplified sign flip with fixed malicious identities, not adaptive poisoning, Sybil/collusion, backdoors, model replacement, privacy attacks, or a compromised server. The focused mechanism and resource sensitivities were deliberately narrow and descriptive, and the complete-round overhead benchmark used one canonical PAMAP2 condition with five measured repetitions; larger models, network latency, and populations beyond 224 eligible clients were not benchmarked. Statistical resolution is limited to five held-out-subject blocks on PAMAP2 and five seed blocks on CICIoT2023; PAMAP2 outer-fold training sets overlap, while all CICIoT2023 seeds share one frozen observation-level split. The CICIoT2023 source-file distribution cannot support a seven-class source-file holdout because Brute Force occurs in only one source file, and device identifiers are unavailable. Finally, the two datasets use different models, clients, workloads, endpoints, and replication units, so their evidence is complementary rather than poolable into a common effect estimate.

6.3. Research-Question Summary

Table 10 consolidates the evidence-based answers and principal interpretation boundaries, avoiding repetition of the detailed numerical results above.

Table 10. Summary of the research questions, principal findings, and boundaries of interpretation.

Research Question	Evidence-Based Answer	Main Qualification
RQ1—Robustness	TEA-FL provided the strongest PAMAP2 sign-flip robustness; ARL-FL recovered under moderate sign flip; conventional methods remained competitive in clean and label-flip conditions.	No method was universally robust; ARL-FL failed under strong PAMAP2 sign flip; with five folds, directionally consistent pairwise effects had a minimum exact two-sided p-value of 0.0625.
RQ2—Sustainability	Under the frozen primary resource model, TEA-FL had the lowest protocol-total, per-completed-round, and common-horizon energy; ARL-FL retained the most active clients and completed every run. TEA-FL savings persisted but narrowed in the communication-dominant sensitivity.	ARL-FLś lower-energy ordering was not universal at strict dropout eligibility; all energy conclusions are simulation-based and resource-model conditional. TEA-FL reached round 100 in 87/100 primary runs.
RQ3—Mechanism attribution	TEA-FL trust separated benign from malicious sign-flip clients and suppressed malicious aggregation influence; ARL-FL risk/clipping reacted strongly, while the $p > 0.5$ regime activated in 14 attack rounds. Focused sensitivity showed strong dependence on attack magnitude and excessive TEA-FL trust persistence, but smaller pressure-threshold effects.	The sensitivity used five folds and one FL seed per cell, so it is descriptive rather than confirmatory tuning; a 5/5 same-direction fold effect still has minimum exact two-sided $p = 0.0625$.
RQ4—Cross-domain robustness transfer	Under frozen CICIoT2023 participation, E0 supported checkpoint protection, while E1 showed substantially weaker final-round performance; E3 recovered 89.1% of the mean E1 penalty and remained within 0.02281 Macro-F1 of E0 across attacked runs.	70/120 attacked E0 runs selected rounds 1–4; E0→E3 changed the leading method in 1/4 scenario × α strata. The evidence concerns active trust/risk and aggregation components conditional on one observation-level split.
RQ5—Operational cost and scaling	At 28 clients, the median per-repetition server-side selection/trust-risk/aggregation total was 13.639 ms for TEA-FL and 55.953 ms for ARL-FL, versus 5.184 ms for FedAvg; ARL-FL additionally retained 4.700 MiB of persistent update tensors.	End-to-end round times are confounded by method-specific selected-client workloads; ARL-FL scheduling increased from 8.89 ms at 28 clients to 352.52 ms at 224 clients, so the measured low-scale overhead should not be extrapolated as negligible in large federations.

Together, these results answer the research questions within explicit experimental boundaries: trust is the principal defense against direct update reversal, energy-aware scheduling primarily controls resource use and retention, adaptive pressure is an operational regulator rather than a universal Byzantine defense, cross-domain transfer is component- and endpoint-specific, and complete-system overhead is modest only at the evaluated small federation scale.

7. Conclusions

TEA-FL and ARL-FL represent two complementary server-side strategies for resource-constrained federated IoT learning. TEA-FL couples trust and residual energy and was strongest against direct sign-flip reversal while achieving the lowest simulated energy under the primary PAMAP2 resource model. ARL-FL separates risk/lifetime-aware schedul-

ing from robust aggregation and provided the strongest client-retention and federation-continuity profile, but it did not withstand the strongest PAMAP2 sign-flip condition and its server scheduling cost grew substantially with the eligible-client population. CICIoT2023 confirmed only conditional transfer of the active trust/risk and aggregation components; checkpoint-policy and fixed-clipping controls prevent a claim of universal adaptive-control superiority.

The central conclusion is therefore a trade-off rather than a ranking: robustness, energy, retention, fairness, and operational overhead must be evaluated jointly but interpreted separately. Future work should validate the resource model on physical heterogeneous devices and networks, broaden the adversarial and robust-baseline set, test provenance-grouped intrusion-detection generalization, extend multi-parameter sensitivity and large-scale overhead studies, and evaluate an independently designed TEA-FL/ARL-FL hybrid.

Author Contributions: Conceptualization, M.J.C.S.R.; methodology, M.J.C.S.R. and C.S.; software, F.B.; validation, M.J.C.S.R., C.S. and F.B.; formal analysis, M.J.C.S.R.; investigation, M.J.C.S.R. and F.B.; resources, M.J.C.S.R., C.S. and F.B.; data curation, C.S.; writing—original draft preparation, M.J.C.S.R.; writing—review and editing, M.J.C.S.R., C.S. and F.B.; visualization, M.J.C.S.R.; project administration, C.S. and 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. This study performed secondary analyses of publicly available, de-identified datasets and involved no new data collection from human participants.

Informed Consent Statement: Not applicable. No new participants were recruited, and no identifiable participant information was used.

Data Availability Statement: The PAMAP2 Physical Activity Monitoring dataset is publicly available through the UCI Machine Learning Repository (https://doi.org/10.24432/C5NW2H accessed on 19 October 2026). The CICIoT2023 dataset is publicly available through the Canadian Institute for Cybersecurity at the University of New Brunswick; its accompanying dataset article is available at https://doi.org/10.3390/s23135941, accessed on 19 October 2026. The source datasets and record-level derived data are not redistributed in this study. The frozen source code, experimental configurations, partition-generation scripts, malicious-client identities, participation schedules, statistical-analysis scripts, environment specifications, run-level evidence, and publication figures used to reproduce and audit the reported results are publicly archived on Zenodo (version 1.0.0; https://doi.org/10.5281/zenodo.21923445 accessed on 19 October 2026). The corresponding development repository is available at https://github.com/mcabralreis/tea-arl-fl-iot-reproducibility accessed on 19 October 2026.

Acknowledgments: During the final revision of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.6) for language refinement and to assist with the drafting of reproducibility and audit scripts and the organization of revision documentation. All AI-assisted outputs were independently reviewed, verified, and edited by the authors, and all scripts were executed and validated by the authors against the underlying experimental evidence. The authors take full responsibility for the content of this publication.

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

Appendix A

This appendix documents the frozen TEA-FL and ARL-FL control parameters and decision rules implemented in the primary campaigns. Table A1 summarizes their values and operational roles. These values define the primary configuration. The sensitivity analyses vary only the mechanism controls in Table A2 and the resource regimes in Table A3 and are interpreted as sensitivity analyses rather than parameter optimization.

Table A1. Frozen TEA-FL and ARL-FL control parameters and decision rules used in the reported experimental campaigns. The values are reproduced from the executable protocol, together with their operational roles; the table documents the implemented configuration and should not be interpreted as a factorial sensitivity or independent hyperparameter-optimization analysis.

Method	Parameter	Frozen Value/Rule	Operational Role
TEA-FL	Initial trust	0.5	Initial client reliability state
TEA-FL	Trust EMA β	0.8	Persistence of trust updates
TEA-FL	Trust floor	0.05	Minimum client trust
TEA-FL	Trust ceiling	1.0	Maximum client trust
TEA-FL	Selection trust weight	0.8	Trust contribution to the selection score; energy weight is 0.2
TEA-FL	Exploration ratio	0.1	Random exploration share
TEA-FL	Aggregation trust exponent	2.0	Trust weighting in aggregation
TEA-FL	Aggregation energy exponent	0.3	Residual-energy weighting in aggregation
ARL-FL	Risk EMA β	0.8	Persistence of the fused risk state
ARL-FL	Global-pressure quantile	0.75	Pressure from the selected-client risk distribution
ARL-FL	Staleness cap	10 rounds	Cap in the Pareto scheduling objective
ARL-FL	Exploration slots	3/2/1	Three if pressure < 0.25; two if $0.25 \leq$ pressure < 0.5; one otherwise
ARL-FL	Clipping κ	$3 - 2 \times$ pressure, bounded to $[1,3]$	Pressure-adaptive norm clipping
ARL-FL	Coordinate trim count	1 or 2 per tail	One if pressure < 0.5; two otherwise

Detailed outer-fold virtual-client window-count, entropy, and Jensen–Shannon summaries previously tabulated in the Appendix A are retained in the versioned archive. The main-text design table and Section 4.3 retain the information required to interpret the client construction and heterogeneity protocol.

Table A2. Focused PAMAP2 sensitivity design. The central values are the frozen primary settings; low and high levels were predeclared before the sensitivity-analysis outcomes were inspected.

Applies to	Sensitivity Axis	Low	Central	High
TEA-FL and ARL-FL	Sign-flip magnitude λ	2.5	5.0	7.5
TEA-FL	Trust EMA persistence β_T	0.65	0.80	0.95
ARL-FL	Pressure switch τ_P	0.35	0.50	0.65

The sensitivity block used strong sign flip only ($\mu = 0.4$), $\alpha \in \{0.1, 1.0\}$, all five outer folds, and federated seed 123. Twenty contemporaneous central controls and 80 non-central variants were evaluated within one matched sensitivity block.

Table A3. PAMAP2 resource-model sensitivity. All regimes retained standby cost 0.0005, the same data/model/attack protocol, five held-out-subject folds, and federated seed 123.

Regime	c_{comm}	c_{comp}	E_{crit}	Role
Primary/computation-dominant	0.004	0.016	0.10	Frozen primary setting
Communication-dominant	0.016	0.004	0.10	Reversed cost mix
Low dropout threshold	0.004	0.016	0.05	More permissive eligibility
High dropout threshold	0.004	0.016	0.20	Stricter eligibility

The resource block used two representative conditions (clean $\alpha = 0.1$ and strong sign flip $\mu = 0.4, \alpha = 1.0$) and three methods (FedAvg, TEA-FL, and ARL-FL), yielding 120 matched runs.

Table A4. PAMAP2 nested subject-wise model-selection structure. For each row, each listed complete subject served once as the inner validation subject; inner training used all eight core subjects except the outer test subject and that inner validation subject.

Outer Fold	Outer Test Subject	Inner Validation Subjects (One per Inner Fold)
1	101	102, 105, 106, 108
2	102	101, 105, 106, 108
3	105	101, 102, 106, 108
4	106	101, 102, 105, 108
5	108	101, 102, 105, 106

All 20 inner-fold rows satisfied the exclusion criterion: the outer test subject was absent from both inner training and inner validation. The six candidate configurations crossed three representations (27 raw axes, six acceleration/gyroscope magnitudes, and nine acceleration/gyroscope/magnetometer magnitudes) with BatchNorm or GroupNorm.

References

1. 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, Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, pp. 1273–1282.
2. Kairouz, P.; McMahan, H.B. Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
3. Zhu, H.; Xu, J.; Liu, S.; Jin, Y. Federated Learning on Non-IID Data: A Survey. *Neurocomputing* **2021**, *465*, 371–390. [CrossRef]
4. 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.
5. Wang, S.; Tuor, T.; Salonidis, T.; Leung, K.K.; Makaya, C.; He, T.; Chan, K. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 1205–1221. [CrossRef]
6. Imteaj, A.; Thakker, U.; Wang, S.; Li, J.; Amini, M.H. A Survey on Federated Learning for Resource-Constrained IoT Devices. *IEEE Internet Things J.* **2022**, *9*, 1–24. [CrossRef]
7. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
8. 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, Stockholm, Sweden, 10–15 July 2018; pp. 5650–5659.
9. Fang, M.; Cao, X.; Jia, J.; Gong, N. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In Proceedings of the 29th USENIX Security Symposium, Boston, MA, USA, 12–14 August 2020; pp. 1605–1622.
10. Pillutla, K.; Kakade, S.M.; Harchaoui, Z. Robust Aggregation for Federated Learning. *IEEE Trans. Signal Process.* **2022**, *70*, 1142–1154. [CrossRef]
11. Reiss, A. *PAMAP2 Physical Activity Monitoring*; UCI Machine Learning Repository: Irvine, CA, USA, 2012.
12. Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors* **2023**, *23*, 5941. [CrossRef] [PubMed]
13. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Seneviratne, A.; Li, J.; Poor, H.V. Federated Learning for Internet of Things: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1622–1658. [CrossRef]
14. 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, Virtual Event, 12–18 July 2020; pp. 5132–5143.
15. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 7611–7623.
16. Acar, D.A.E.; Zhao, Y.; Matas, R.; Mattina, M.; Whatmough, P.; Saligrama, V. Federated Learning Based on Dynamic Regularization. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, 3–7 May 2021.

17. Cho, Y.J.; Wang, J.; Joshi, G. Towards Understanding Biased Client Selection in Federated Learning. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Virtual Event, 28–30 March 2022; pp. 10351–10375.
18. Cao, X.; Fang, M.; Liu, J.; Gong, N.Z. FLTrust: Byzantine-Robust Federated Learning via Trust Bootstrapping. In Proceedings of the Network and Distributed System Security Symposium (NDSS), Reston, VA, USA, 21–24 February 2021; pp. 1–18. [CrossRef]
19. Fung, C.; Yoon, C.J.M.; Beschastnikh, I. The Limitations of Federated Learning in Sybil Settings. In Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses, San Sebastián, Spain, 14–16 October 2020; pp. 301–316.
20. Nguyen, T.D.; Rieger, P.; Chen, H.; Yalame, H.; Möllering, H.; Fereidooni, H.; Marchal, S.; Miettinen, M.; Mirhoseini, A.; Zeitouni, S.; et al. FLAME: Taming Backdoors in Federated Learning. In Proceedings of the 31st USENIX Security Symposium, Boston, MA, USA, 10–12 August 2022; pp. 1415–1432.
21. 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. [CrossRef]
22. Nishio, T.; Yonetani, R. Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge. In Proceedings of the ICC 2019–2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019; pp. 1–7.
23. Lai, F.; Zhu, X.; Madhyastha, H.V.; Chowdhury, M. Oort: Efficient Federated Learning via Guided Participant Selection, 2021. In Proceedings of the 18th USENIX Symposium on Networked Systems Design and Implementation, Virtual Event, 12–14 April 2021; pp. 19–35.
24. Chai, Z.; Ali, A.; Zawad, S.; Truex, S.; Anwar, A.; Baracaldo, N.; Zhou, Y.; Ludwig, H.; Yan, F.; Cheng, Y. TiFL: A Tier-Based Federated Learning System. In Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing, New York, NY, USA, 23–26 June 2020; pp. 125–136.
25. Wu, J.; Drew, S.; Zhou, J. FedLE: Federated Learning Client Selection with Lifespan Extension for Edge IoT Networks. In Proceedings of the ICC 2023—IEEE International Conference on Communications, Rome, Italy, 28 May–1 June 2023; pp. 985–990.
26. Li, T.; Sanjabi, M.; Beirami, A.; Smith, V. Fair Resource Allocation in Federated Learning. *arXiv* **2019**, arXiv:1905.10497.
27. Rousseeuw, P.J.; Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [CrossRef]
28. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
29. Jain, R.; Chiu, D.; Hawe, W. *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems*; Technical Report DEC-TR-301; Digital Equipment Corporation: Maynard, MA, USA, 1984.
30. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
31. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
32. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
33. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
34. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
35. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]

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.