

Energy-aware Adaptive Federated Optimization for Sustainable IoT: A Unified Energy-communication-accuracy Trade-off Framework

Almamoon Alauthman ^{1,*} and Feras Alfaqih ²

¹ Electrical Engineering Department, Al-Balqa Applied University, As Salt, Jordan

² Department of Mathematics, Al-Hussein Bin Talal University, Ma'an, Jordan

Email: almamoon.athanneh1@bau.edu.jo (A.A.); falfaqih@ahu.edu.jo (F.A.)

Manuscript received May 5, 2026; revised July 29, 2026; accepted September 16, 2026; published October 26, 2026

*Corresponding author

Abstract—Federated learning lets Internet of Things devices train a shared model without transmitting raw data, but its energy cost on battery-powered hardware remains a barrier to deployment. That cost has two components, local computation and parameter exchange, and the number of local epochs performed between synchronizations trades one against the other. This paper formulates the trade-off as a normalized multi-objective problem and derives a closed-form expression for the epoch count that minimizes it. The expression depends on one measurable quantity, the fraction of total energy attributable to communication, and is embedded in a Lyapunov drift-plus-penalty controller that adapts the epoch count online under a long-run energy budget, with proven mean-rate stability and an $O(1/V)$ optimality gap. Experiments on the CICIoT2023 intrusion-detection dataset support the prediction: the optimal epoch count rises with the communication energy fraction, confirmed through two independent mechanisms, varying the uplink rate and the quantization width. Against federated averaging at a fixed two local epochs, the controller improves accuracy significantly at three levels of statistical heterogeneity. Against federated averaging tuned over a grid of fixed epoch counts it is less energy-efficient in a single run; its advantage is reaching a comparable operating point without that search, an 81.4% reduction in total energy once the search is counted. The framework yields negative savings in high-bandwidth settings, which the theory predicts. All energy figures reported are outputs of an analytical energy model rather than hardware measurements.

Index Terms—adaptive optimization, communication-energy trade-off, energy-efficient computing, federated learning, Internet of Things, Lyapunov optimization

I. INTRODUCTION

The Internet of Things (IoT) has placed sensing and computation at the network edge across smart cities, healthcare and industrial monitoring [1, 2]. These devices generate continuous data streams whose analysis supports intelligent decision-making [3, 4], but transmitting raw data to a central server raises privacy exposure, latency and bandwidth cost [5, 6]. Federated Learning (FL)

addresses this by keeping data local: clients train on their own data and exchange only model parameters [3, 6].

The gains are not free. IoT nodes operate under strict energy limits, often on batteries or harvested power, and FL imposes two recurring costs on every client: local gradient computation, and uplink transmission of model parameters once per communication round [7, 8]. Both scale with the number of rounds. Reducing rounds by performing more local work per round lowers communication cost but raises computation cost and, under Non-Independent and Identically Distributed (Non-IID) data, allows client models to drift apart between synchronizations [9, 10].

A. Prior Work in Brief

Three lines of work address parts of this problem. Communication-efficient methods reduce the volume or frequency of exchange through compression, quantization and client selection [6, 10, 11]. Energy-efficient methods model device power explicitly and allocate computation or bandwidth to minimize consumption [7, 8, 12, 13]. Adaptive methods adjust learning parameters during training in response to observed system state [9, 14–16]. Section II reviews each in detail and Table I positions the present work among them.

B. Problem Statement

Consider a federation of N clients coordinated over R communication rounds, each client performing τ local epochs per round. Three quantities depend on τ and pull in different directions.

Communication cost. Each round requires every participating client to upload its parameter vector. For a fixed number of local updates, increasing τ reduces the number of rounds and hence the total volume exchanged.

Computation cost. Each local epoch processes the client's full local dataset. At a fixed number of rounds, computation energy grows linearly with τ .

Model accuracy. Additional local work reduces optimization error, but with diminishing returns, and

under non-IID partitions excessive local work induces client drift that degrades the aggregated model [9, 17, 18].

These three cannot be optimized independently: any change to τ improves at least one and degrades at least one other. The problem addressed here is to characterize the value of τ that balances them, to express that value in closed form so it can be computed rather than searched, and to adapt it online when the operating conditions change. Section III-F states the objective formally and Section III-G derives its minimizer.

C. Research Gaps

The literature reviewed in Section II leaves three gaps.

Gap 1: absence of a joint formulation. Energy, communication and accuracy are typically optimized in isolation or combined heuristically. Where a combined objective is used, the terms are frequently summed without normalization, so a quantity in joules is added to a dimensionless loss and the weighting is not interpretable [16–18].

Gap 2: absence of closed-form parameter selection. The number of local epochs is usually fixed by grid search or by convention. Analytical characterizations that identify the optimum from measurable system quantities are uncommon [7, 14], and where a closed form is claimed the derivation is often not reproducible from the stated assumptions.

Gap 3: limited treatment of environmental cost. Carbon accounting in FL, where present, applies a fixed conversion factor to total energy. Whether energy minimization and carbon minimization coincide, and under what conditions they diverge, is rarely stated [19, 20].

D. Contributions

The contributions of this work are as follows. Where a component builds on prior work, the relationship is described in the section that develops it.

A normalized multi-objective formulation. Energy and loss are referred to a fixed reference configuration before combination, so the objective is dimensionless and the weight α is an interpretable preference parameter rather than a scaling constant (Section III-F).

A closed-form optimal local epoch count. Under a stated loss model, the minimizer of the objective is derived with first- and second-order conditions and shown to be the unique global optimum. The result depends on the communication energy fraction ρ , a quantity computable from device and channel parameters (Section III-G, supplementary Appendix A).

A Lyapunov controller with proven guarantees. A virtual energy-deficit queue converts a long-run energy budget into a per-round decision. The resulting policy inherits the closed form of contribution 2, satisfies the budget in the time average, and attains an $O(1/V)$ optimality gap with $O(V)$ backlog (Section III-G, supplementary Appendix A).

An empirical characterization, including negative results. The predicted dependence of τ^* on ρ is tested through two independent mechanisms. The method is compared against baselines given their own tuned epoch

count, not a fixed one. Where a component shows no measurable effect, this is reported (Section IV).

E. Organization

Section II reviews related work and positions this study. Section III develops the system, energy and carbon models and states the objective. It further derives the optimal epoch count, presents the Lyapunov controller and the aggregation rule, and describes the algorithm. Section III-H details the experimental setup. Section IV reports and discusses the results, together with threats to validity and limitations. Section V concludes. Complete derivations and the implementation listing are provided in two supplementary documents, Appendix A and Appendix B, submitted alongside this manuscript.

II. LITERATURE REVIEW

A. Communication-efficient Federated Learning

Federated averaging established the pattern of multiple local steps between synchronizations, trading communication for computation [21]. Subsequent work reduced the exchanged volume through gradient compression and sparsification [6, 11, 22] and reduced the number of participants per round through client selection [10]. These methods lower communication cost at fixed local work; they do not select the amount of local work itself.

B. Energy-efficient Federated Learning

A second line models device energy explicitly. Computation energy is derived from the CMOS dynamic power relation and communication energy from transmit power and achievable rate, and the resulting problem is solved for bandwidth or frequency allocation [7, 8, 12, 23]. These formulations treat the local epoch count as given. The energy model adopted in Section III-C follows this literature, with the correction noted there.

C. Adaptive Federated Optimization

Adaptive methods adjust parameters during training. Server-side adaptive optimizers modify the aggregation step [24]; proximal regularization limits client drift [25]; normalized averaging corrects for unequal local work across clients [26, 27]. Within the federated learning literature on edge resources, [28] applies Lyapunov optimization to energy-efficient adaptive federated learning and is the closest prior work to the formulation developed here; the distinction is drawn in Section II-E. These target the optimization dynamics rather than the energy budget, and none derives the local epoch count from an energy-aware objective.

D. Sustainable and Carbon-aware Learning

Reporting of the environmental cost of training has become established practice [19, 29, 30]. Most treatments convert measured energy to emissions with a fixed regional factor. The consequence of that assumption for the optimization problem, namely that under a constant factor the two objectives share a minimizer, is stated as Proposition 3 in Section III-E.

TABLE I: POSITIONING OF THIS WORK

Ref.	Approach	Objectives	Adaptive parameter	Closed form	Guarantee
[21]	Federated averaging	A	None	No	Convergence
[24]	Proximal regularization	A	None	No	Convergence
[17]	Normalized averaging	A	None	No	Convergence
[25]	Server-side adaptive optimizer	A	Server step size	No	Convergence
[10]	Client selection by resource	C, A	Participating set	No	Empirical
[7]	Energy-aware communication design	E, C	Bandwidth	No	Optimality
[8]	Algorithm-hardware co-design	E, A	Model compression	No	Empirical
[11]	NOMA and power allocation	E, C	Transmit power	No	Optimality
[14]	Adaptive split federated learning	E, C, A	Split point	No	Convergence
[21]	Cell-free network optimization	E, C	Resource allocation	No	Optimality
[13]	Carbon-aware model sizing	CO ₂ , A	Model size	No	Empirical
[15]	Lyapunov edge-resource optimization	E, A	Resource allocation	No	Stability
This work	E2A-FO	E, C, A, CO ₂	Local epoch count	Yes	Convexity; mean-rate stability; $O(1/V)$ gap

E = energy, C = communication, A = accuracy.

E. Positioning

This work differs from the entries above in the quantity adapted and in how that quantity is obtained. Prior work adapts bandwidth, transmit power, the participating set, the split point or the model size, and obtains the setting numerically. Here the adapted quantity is the local epoch count and the setting is available in closed form from measurable system parameters.

The closest prior work is [15], which also applies Lyapunov optimization to energy-efficient adaptive federated learning at the edge. Two differences are material. That formulation allocates edge resources under an energy constraint and obtains the allocation by solving the per-slot problem numerically; the present formulation adapts the amount of local computation and solves the per-round problem analytically through Theorem 1, so the per-round cost is $O(1)$. The limitations of this contribution are established empirically in Section IV-E and stated in Section IV-N.

III. MATERIALS AND METHODS

A. Dataset and Feature Representation

Experiments use CICIOT2023 [31], a network intrusion dataset collected from a testbed of commercial IoT devices. The distributed file contains 47 columns, of which one is the class label. The feature vector is therefore

$$x_i \in \mathbb{R}^{46}, \quad y_i \in \{1, \dots, 34\},$$

and the network input layer has 46 units.

The features fall into four groups: flow-level statistics (duration, rate, inter-arrival time), packet-level statistics (header length, total size, and the mean, standard deviation, variance, magnitude, radius and covariance of packet sizes), protocol identifiers, and TCP control-flag counts. No feature selection is applied, so that results remain comparable with published work using the same schema.

The class distribution is severely imbalanced: the largest class holds 1,210,546 instances and the smallest 208, a ratio of approximately 5,800 to 1. This governs the choice of evaluation metric in Section V-C.

B. Notation

Symbols are defined at first use. Table II collects every symbol used in this work together with its unit; dimensionless quantities are marked “n/a”.

TABLE II: NOTATION AND UNITS

Symbol	Meaning	Unit
N	Number of clients in the federation	n/a
$\mathcal{S}_r$	Set of clients participating in round r	n/a
R	Number of global communication rounds	n/a
τ	Local epochs per client per round	n/a
n_k	Number of training samples at client k ; $n = \sum_k n_k$	n/a
p_k	Sample proportion n_k/n	n/a
α_k	Aggregation weight of client k	n/a
w	Model parameter vector	n/a
d	Number of model parameters	n/a
b	Quantization width per parameter	bit
S	Uplink payload, $S = bd$	bit
κ	Effective switched-capacitance coefficient	J·s ² /cycle ³
C_k	CPU cycles required per training sample	cycles/sample
f_k	CPU clock frequency of client k	Hz
P_{tx}, P_{rx}	Uplink transmit and downlink receive power	W
B	Channel bandwidth	Hz
g_k	Channel power gain of client k	n/a
N_0	Noise power spectral density	W/Hz
r_k	Achievable uplink rate of client k	bit/s
$E_{\text{comp}}^{\text{unit}}(k)$	Computation energy, one local epoch	J
$E_{\text{comm}}(k)$	Communication energy, one round	J
E_{total}	Total energy over the run	J
$A_{\text{comp}}, A_{\text{comm}}$	Per-round federation-wide computation, communication energy	J
ρ	Communication energy fraction, Eq. (7)	n/a
$F_k(w), F(w)$	Local and aggregate empirical loss	nat
$L(\tau)$	Attainable loss at τ local epochs	nat
L^*, A, D	Loss floor, optimization coefficient, drift coefficient	nat, nat, nat
α	Energy-accuracy weight in the objective	n/a
η	Energy-performance weight in aggregation	n/a
V	Lyapunov penalty parameter	n/a
$Q(r)$	Virtual energy-deficit queue	J
$\bar{E}$	Per-round energy budget	J
$\gamma(r)$	Grid carbon intensity at round r	gCO ₂ e/kWh

C. Energy Model

Computation. Under the CMOS dynamic power relation, a processor operating at clock frequency f dissipates

$$P = \kappa f^3, \quad (1)$$

where κ aggregates effective switched capacitance and voltage-frequency scaling. Executing C CPU cycles requires time C/f , so the energy consumed is

$$E = P \cdot \frac{C}{f} = \kappa C f^2. \quad (2)$$

One local epoch at client k processes n_k samples at C_k cycles per sample, giving $C = C_k n_k$ and

$$E_{\text{comp}}^{\text{unit}}(k) = \kappa C_k f_k^2 n_k \quad [\text{J}], \quad (3)$$

so that τ local epochs cost $\tau E_{\text{comp}}^{\text{unit}}(k)$.

In Eq. (1) the coefficient κ is a per-cycle quantity and the cycle count C does not appear in the power law. A formulation in which C entered the power relation would make energy quadratic in the cycle count, and therefore quadratic in n_k , whereas the energy of one local epoch is linear in the number of samples processed.

Communication. Each participating client uploads a quantized parameter vector of $S = bd$ bits over a channel whose achievable rate follows the Shannon relation

$$r_k = B \log_2 \left(1 + \frac{P_{tx} g_k}{N_0 B} \right) \quad [\text{bit/s}], \quad (4)$$

giving a per-round communication energy

$$E_{\text{comm}}(k) = P_{tx} \frac{S}{r_k} + P_{rx} \frac{S}{r_{dl}} \quad [\text{J}]. \quad (5)$$

The transmission time is obtained from the achievable rate rather than from the bandwidth: B is measured in hertz and S in bits, so S/B is not a time. Eq. (4) supplies the rate, and the payload is expressed in bits throughout.

Aggregate. With $A_{\text{comp}} = \sum_{k \in \mathcal{S}} E_{\text{comp}}^{\text{unit}}(k)$ and $A_{\text{comm}} = \sum_{k \in \mathcal{S}} E_{\text{comm}}(k)$ denoting per-round federation-wide quantities, the total energy over R rounds is

$$E_{\text{total}}(\tau, R) = R(\tau A_{\text{comp}} + A_{\text{comm}}). \quad (6)$$

The ratio governing everything that follows is the **communication energy fraction**

$$\rho = \frac{A_{\text{comm}}}{A_{\text{comp}} + A_{\text{comm}}} \in (0, 1), \quad (7)$$

evaluated at $\tau = 1$. It measures how communication-dominated the federation is, and is computable in advance from device and channel parameters.

Energy in Eq. (6) depends on the algorithm and not only on the schedule (N, R, τ) . Methods performing additional local work, such as the proximal term of FedProx, consume correspondingly more energy. An energy model that depends only on the schedule cannot distinguish between algorithms and would report identical consumption for methods that perform different amounts of work.

D. Energy Model Parameters

Table III presents the parameters of the energy model. Client heterogeneity in f_k and g_k is required for the aggregation rule of Section III-G to be non-degenerate: Proposition 2 shows that identical clients reduce it exactly to federated averaging.

TABLE III: ENERGY MODEL PARAMETERS

Symbol	Description	Value	Unit	Basis
κ	Switched-capacitance coefficient	1×10^{-28}	$\text{J} \cdot \text{s}^2 / \text{cycle}^3$	DVFS literature [7, 32]
C_k	CPU cycles per sample	5×10^5	cycles/sample	Edge-device profiling [33]
f_k	Clock frequency (heterogeneous)	$\mathcal{U}(0.5, 2.0) \times 10^9$	Hz	Cortex-A class devices
P_{tx}	Uplink transmit power	0.2	W	3GPP NB-IoT power class [34]
P_{rx}	Downlink receive power	0.1	W	Device datasheet
N_0	Noise power spectral density	-174	dBm/Hz	Thermal noise
b	Quantisation width	8, 16, 32	bit	Design variable
γ	Grid carbon intensity	30 to 800 (swept)	$\text{gCO}_2\text{e/kWh}$	IEA regional range

E. Carbon Model

Total energy converts to equivalent emissions as

$$\text{CO}_2 = \sum_{r=1}^R \frac{E^{(r)}}{3.6 \times 10^6} \gamma(r) \quad [\text{kgCO}_2\text{e}], \quad (8)$$

with $E^{(r)}$ in joules and $\gamma(r)$ in gCO_2e per kWh; the constant converts joules to kilowatt-hours.

Proposition 3 (energy-carbon equivalence). *If $\gamma(r) \equiv \gamma$ is constant, then CO_2 is a strictly increasing affine transform of E_{total} , and for any decision variable x ,*

$$\arg\min_x E_{\text{total}}(x) = \arg\min_x \text{CO}_2(x).$$

Proof. Strictly increasing affine transforms preserve minimizers.

Under a constant emission factor, including carbon in the objective introduces redundancy without altering the solution, which is why carbon is reported here as a derived

quantity rather than optimized. The equivalence fails when $\gamma(r)$ varies: two schedules consuming identical energy on grids at 50 and 800 $\text{gCO}_2\text{e/kWh}$ differ sixteen-fold in emissions. Carbon-aware scheduling under time-varying intensity is therefore a distinct problem, identified as future work in Section V.

F. Normalized Multi-objective

Energy is measured in joules and loss in nats. Summing them directly is not meaningful, and it produces a further difficulty: with energy of order 10^2 to 10^3 and loss of order 10^{-1} , the loss term is numerically invisible at any moderate weight. An objective formed by direct summation would therefore be numerically indistinguishable from the energy term alone.

Both terms are therefore referred to the configuration $\tau = 1$ at fixed R :

$$\begin{aligned}\hat{E}(\tau) &= \frac{E_{\text{total}}(\tau)}{E_{\text{total}}(1)} = \tau(1 - \rho) + \rho, \\ \hat{L}(\tau) &= \frac{L(\tau)}{L(1)}\end{aligned}\quad (9)$$

both dimensionless and equal to unity at the reference point. The objective is

$$J(\tau) = \alpha \hat{E}(\tau) + (1 - \alpha) \hat{L}(\tau), \quad \alpha \in (0,1). \quad (10)$$

J is dimensionless and α is a dimensionless preference weight: $\alpha \rightarrow 1$ recovers pure energy minimisation, $\alpha \rightarrow 0$ pure loss minimisation. The reference point is fixed and external to the search, rather than derived from the minimum and maximum over a search grid, which would make the optimum depend on the grid.

G. Proposed Method

1) Loss model

Deriving the optimal local epoch count requires a model of how attainable loss varies with τ . That model is stated here, with its theoretical motivation and its empirical status made explicit.

Under L -smoothness, bounded stochastic gradient variance σ^2 and bounded gradient dissimilarity ζ^2 (Assumptions 1 to 3, supplementary Appendix A), local-SGD analysis of federated averaging admits a bound of the form

$$\begin{aligned}\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} \quad \|\nabla F(w^{(r)})\|^2 &\leq \frac{C_1}{\zeta R} + C_2 \zeta L \sigma^2 \\ &+ C_3 \zeta^2 L^2 \tau^2 (\sigma^2 + \zeta^2)\end{aligned}\quad (11)$$

in which two terms depend on τ and act in opposition. The optimisation term decreases as $1/\tau$, since more local work per round means more gradient steps at fixed R ; the client-drift term increases with τ , since clients diverge further between synchronizations, an effect amplified by heterogeneity.

Eq. (11) bounds the squared gradient norm, not the loss value. Converting between them requires a Polyak-Lojasiewicz or strong-convexity condition and is not available in general non-convex settings. The following is therefore adopted as an empirical model whose functional form is motivated by Eq. (11), with coefficients estimated from data, and not as a derived identity:

$$L(\tau) = L^* + \frac{A}{\tau} + D\tau, \quad A > 0, D \geq 0, \quad (12)$$

where L^* is the attainable loss floor, A measures the diminishing-returns benefit of additional local computation, and D measures client drift. The distinction between a motivated model and a derived identity is stated explicitly, since the coefficients are estimated rather than obtained from the bound.

Eq. (12) admits two regimes. When $D \approx 0$ the loss decreases monotonically and flattens; when $D > 0$ it attains an interior minimum at $\tau = \sqrt{A/D}$. Which regime applies is an empirical question, resolved for the present data in Section IV-A.

2) Closed-form optimal local epochs

Substituting Eq. (9) and Eq. (12) into Eq. (10), and writing $\Delta_L = L(1) = L^* + A + D$, gives

$$J(\tau) = \alpha [\tau(1 - \rho) + \rho] + \frac{(1-\alpha)}{\Delta_L} \left(L^* + \frac{A}{\tau} + D\tau \right) \quad (13)$$

Theorem 1. *Let R be fixed, $\rho \in (0,1)$, $A > 0$, $D \geq 0$ and $\alpha \in (0,1)$. Then $J(\tau)$ in Eq. (13) is strictly convex on $\tau > 0$ and attains its unique global minimum at*

$$\tau^* = \sqrt{\frac{(1-\alpha)A}{\alpha(1-\rho)\Delta_L + (1-\alpha)D}}. \quad (14)$$

For $D = 0$ this reduces to $\tau^* = \sqrt{(1-\alpha)A/[\alpha(1-\rho)\Delta_L]}$.

The proof is given in supplementary Appendix A. Since τ must be integral, Corollary 1 (supplementary Appendix A) establishes that the optimal feasible choice is $\tau^+ = \text{argmin}_{\tau \in \{\tau^*, \lceil \tau^* \rceil\}} J(\tau)$ with $\tau^+ \geq 1$, which follows from strict convexity.

Relation to the objective. Eq. (14) is not a separate objective: it is the stationary point of Eq. (13), which is Eq. (10) with the energy and loss models substituted. The model parameters are optimized by stochastic gradient descent as usual, while τ is selected analytically at fixed model parameters within each round. This two-stage treatment avoids solving the joint problem at every round.

Why the τ -dependence belongs in the loss term. An expression for τ^* cannot be obtained from an energy-only objective. Under a fixed total local-computation budget $K = R\tau$, substituting $R = K/\tau$ into Eq. (6) gives $E_{\text{total}} = KA_{\text{comp}} + KA_{\text{comm}}/\tau$, which is monotonically decreasing in τ and has no interior stationary point: total computation energy is fixed by K , and communication energy falls with every additional local epoch. An interior optimum therefore requires a term that increases with τ , and the only quantity with a mechanism for doing so is the loss, through client drift. This is why Eq. (12) carries the τ -dependence and Eq. (6) does not.

Interpretation. τ^* increases with ρ : the more communication-dominated the federation, the larger the optimal amount of local work per round. It decreases with α , since prioritising energy reduces computation at fixed R , and decreases with D , since drift forces frequent synchronization. As $\rho \rightarrow 0$, $\tau^+ \rightarrow 1$: when communication is nearly free, synchronising after every epoch is optimal. This limiting case is a prediction of the theory and is tested in Section IV-A.

3) Energy-weighted objective

The aggregation rule introduced below assigns weights $\alpha_k \neq p_k$. The algorithm therefore does not converge to a minimiser of the sample-weighted empirical risk $F(w) = \sum_k p_k F_k(w)$, and presenting it as an approximate solver of that objective would be incorrect. The target is redefined explicitly as

$$F_E(w) = \sum_{k=1}^N \alpha_k F_k(w), \quad (15)$$

a deliberate design choice that prioritises clients delivering useful updates at lower energy cost. Supplementary Appendix A gives a convergence result for Eq. (15) and,

in Proposition 1, bounds the penalty this incurs on the classical objective:

$$F(w_E^*) - F(w^*) \leq \frac{\zeta^2}{2\mu} \|\mathbf{p} - \boldsymbol{\alpha}\|_1^2, \quad (16)$$

where w_E^* and w^* minimise F_E and F respectively and μ is a Polyak-Lojasiewicz constant. The bound is the product of two measurable quantities, so it can be evaluated during training; Section IV-H reports that evaluation. The bound therefore quantifies the cost of optimizing F_E rather than F , and permits comparison against federated averaging on standard accuracy.

4) Lyapunov drift-plus-penalty control

Theorem 1 gives a static τ^* for known parameters. The proposed method is adaptive: it adjusts τ_r per round subject to a long-run energy budget $\bar{E}$. Define the virtual energy-deficit queue

$$Q(r+1) = \max\{0, Q(r) + E^{(r)} - \bar{E}\} \quad (17)$$

with $Q(0) = 0$.

Define the quadratic Lyapunov function $\mathcal{V}(Q) = 1/2 Q^2$ and one-slot conditional drift $\Delta(r) = \mathbb{E}[\mathcal{V}(Q(r+1)) - \mathcal{V}(Q(r)) | Q(r)]$. Lemma 2 (supplementary Appendix A) bounds the drift, and adding V times the loss penalty yields the control rule

$$\tau_r = \operatorname{argmin}_{\tau \geq 1} \{Q(r) E^{(r)}(\tau) + V L(\tau)\}. \quad (18)$$

Corollary 2. *Dividing the objective in Eq. (18) by $Q(r) + V$ recovers Eq. (10) with $\alpha_r = Q(r)/(Q(r) + V)$, so Theorem 1 applies at each round:*

$$\tau_r^* = \sqrt{\frac{VA}{Q(r)(1-\rho)\Delta_L + VD}}. \quad (19)$$

The per-round decision therefore costs $O(1)$. When the federation overspends its budget, $Q(r)$ grows, $\alpha_r \rightarrow 1$ and τ_r falls; when consumption is below budget, τ_r rises. Equation (19) unifies the static and adaptive formulations: the static result is the constant- α_r special case.

Theorem 3. *Under bounded per-round energy and strict feasibility of the budget (Assumptions 4 to 6, supplementary Appendix A), the rule in Eq. (18) with any $V > 0$ satisfies: (i) $\frac{1}{R} \sum_r \mathbb{E}[Q(r)] = O(V)$; (ii) mean-rate stability, hence $\limsup_R \frac{1}{R} \sum_r \mathbb{E}[E^{(r)}] \leq \bar{E}$; and (iii) $\limsup_R \frac{1}{R} \sum_r \mathbb{E}[L^{(r)}] \leq L^{\text{opt}} + B/V$.*

The proof is given in supplementary Appendix A. Parts (i) and (iii) give the standard trade-off: the optimality gap decays as $O(1/V)$ while backlog grows as $O(V)$. Equation (17), Lemma 2 and Theorem 3 supply the Lyapunov function, the drift analysis and the stability proof respectively.

5) Energy-aware aggregation

Let $\tilde{E}_k = E_k/\max_j E_j$ and $\tilde{F}_k = F_k/\max_j F_j$ be per-client scores normalised over the participating set, both in $(0,1]$. The aggregation weight is

$$\alpha_k = \frac{\frac{n_k}{\eta \tilde{E}_k + (1-\eta) \tilde{F}_k + \epsilon}}{\sum_{j \in \mathcal{S}_r} \frac{n_j}{\eta \tilde{E}_j + (1-\eta) \tilde{F}_j + \epsilon}} \quad (20)$$

with $\eta \in [0,1]$,

where $\epsilon > 0$ is a small constant for numerical safety. The properties of the rule are as follows.

The weights sum to unity by construction, so Eq. (20) defines a convex combination, and both arguments of the denominator are dimensionless ratios, so no term in joules is added to a term in nats. The numerator preserves the sample-size weighting of federated averaging.

The two partial derivatives $\partial \alpha_k / \partial \tilde{E}_k$ and $\partial \alpha_k / \partial \tilde{F}_k$ are both negative: a client that expends more energy for the same contribution, or returns a worse local loss, receives less weight. The parameter η sets the balance between the two penalties.

Proposition 2. *If $\tilde{E}_k = \bar{E}$ and $\tilde{F}_k = \bar{F}$ for all k , then $\alpha_k = n_k/n = p_k$ and the rule reduces exactly to federated averaging.*

Proof. With identical denominators θ , $\alpha_k = (n_k/\theta) / \sum_j (n_j/\theta) = n_k/n$.

Two consequences follow. The method strictly generalizes federated averaging, and client heterogeneity is a precondition for the rule to have any effect. Section IV-D reports the measured effect, which is not the one anticipated.

Weighting by $1/\tilde{E}_k$ systematically down-weights energy-expensive clients, which are typically the most resource-constrained devices. Section IV-G measures the consequence for the worst-served client.

6) Algorithm and system operation

Algorithm 1 states the procedure. Steps 2 and 3 evaluate Eq. (7) and Corollary 2, step 4 the integer projection of Eq. (19), step 5 the energy of Eqs. (3) and (5), step 6 the weights of Eq. (20), step 9 the queue of Eq. (17) and step 10 the emissions of Eq. (8). Fig. 1 shows the operation of one communication round, and Algorithm 1 states the corresponding procedure. The server broadcasts the current global parameters to the participating clients. Each client estimates its own computation and communication energy from Eq. (3) and Eq. (5), performs τ_r local epochs on its private partition, and returns the updated parameters together with its measured energy and local loss. The server forms the weights of Eq. (20), aggregates, updates the virtual queue by Eq. (17), and selects τ_{r+1} from Eq. (19) for the next round. Raw data never leaves the client.

Algorithm 1: E2A-FO

Input: N, R , budget E_{bar} , penalty V , weight ϵ ,
drift parameters (L^*, A, D)

Init: $w(0); Q(0) = 0$

for $r = 0 \dots R-1$:

1 sample participating set $\mathcal{S}_r$

2 estimate ρ from $A_{\text{comp}}, A_{\text{comm}}$

3 $\alpha_r \leftarrow Q(r) / (Q(r) + V)$

4 $\tau_r \leftarrow$ integer projection of $\tau^*, \geq 1$

5 for each client k in $\mathcal{S}_r$, in parallel:

train locally for τ_r epochs

```

    measure E_k and local loss F_k
    6 form weights alpha_k
    7 w(r+1) <- sum_k alpha_k w_k(r)
    8 E(r) <- sum_k E_k
    9 Q(r+1) <- max{0, Q(r) + E(r) - E_bar}
    10 accumulate CO2 with gamma(r)
    
```

Output: $w(R)$, E total, CO2

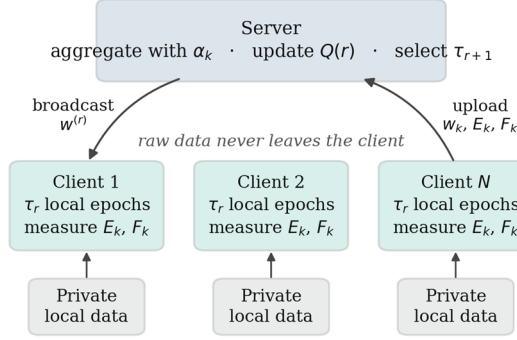


Fig. 1. Operation of a single communication round within the proposed framework.

Complexity. Steps 2-4 are $O(1)$. Step 6 is $O(|S_r|)$. Local training dominates at $O(|S_r| \tau_r n_k d)$, so the control overhead is negligible relative to the computation it schedules.

H. Experimental Setup

1) Datasets and partitioning

Experiments use CICIOT2023 as described in Section III-A. A stratified subsample is drawn in which the per-class quota is capped and classes below the cap are retained whole, preserving the tail of the distribution while keeping the experimental grid tractable. Splits are stratified with shuffling, and standardization statistics are computed on the training partition only and then applied to the held-out partitions.

With 34 classes, a sequential slice of an unshuffled file can exclude classes entirely from one side of a split, and fitting the scaler before splitting leaks held-out statistics into the transformation.

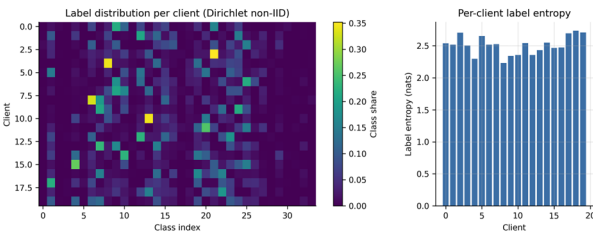


Fig. 2. Client label distribution and per-client label entropy under Dirichlet partitioning.

Training data are distributed across clients by a label-wise Dirichlet partition: for each class c , the proportion vector across clients is drawn as $\mathbf{p}_c \sim \text{Dir}(\alpha_{\text{Dir}} \mathbf{1}_N)$, with lower α_{Dir} producing stronger label skew. Values $\alpha_{\text{Dir}} \in \{0.1, 0.5, 1.0\}$ are evaluated. Fig. 2 shows the resulting partition at $\alpha_{\text{Dir}} = 0.5$: clients hold markedly different class mixtures, and several hold no instances of some

classes, which is the intended property of a non-IID benchmark.

For each seed, one partition is drawn and reused by every method and configuration. Redrawing the partition per run would confound differences between methods with partition variance.

2) Configuration

TABLE IV: EXPERIMENTAL CONFIGURATION

Parameter	Value
Architecture	MLP 46-128-64-34, ReLU; $d = 16,482$
Local optimizer	Adam, learning rate 1×10^{-3}
Loss	Cross-entropy
Batch size	512
Optimizer state	Reinitialized each round
Clients N	20 (5-100 in the scalability study)
Participation C	1.0 (0.2, 0.5 in the participation study)
Local epochs τ	1-32
Lyapunov parameter V	10^0 to 10^5 ; 10^3 unless stated
Energy weight α	0.05-0.5
Aggregation weight η	0.5
Seeds	3-5 per experiment, stated per table
Determinism	Model initialization seeded per run; deterministic kernels enabled

Model initialization is seeded within each run. Without this, the initial weights are drawn from a generator whose state advances across calls, and a configuration's result then depends on the order in which configurations are executed. Table IV presents the experimental configuration.

3) Evaluation metrics

Accuracy, macro-averaged F_1 , cross-entropy loss, total energy in kilojoules, energy to reach a target accuracy, and equivalent emissions. Energy is reported in kilojoules to a precision consistent with the seed-to-seed variation, rather than to five decimal places in joules.

Macro- F_1 is reported alongside accuracy throughout. With 34 classes at an imbalance ratio near 5,800 to 1, aggregate accuracy can remain high while the model fails on minority attack classes, which are precisely the classes an intrusion-detection system exists to identify. The gap between the two metrics in Section IV is substantial and would have been invisible had accuracy been reported alone.

4) Comparison protocols

A comparison confounds method with configuration when the methods are run at different local epoch counts. Three protocols are therefore defined and reported separately.

P1, equal computation budget. All methods run at identical $K = R\tau$, N , participation and partition.

P2, equal energy budget. All methods run until a fixed energy budget is exhausted; final accuracy is compared.

P3, equal target accuracy. All methods run until a target validation accuracy is reached; energy-to-target is compared. This is the primary protocol for an energy-oriented study.

Under P3, baselines are additionally swept over $\tau \in \{1, 2, 4, 8, 16\}$ and compared at **their own best setting**, not

at a fixed one. Section IV-E reports that comparison, which is the decisive test of the method's energy claim.

5) Baselines and reproducibility

Baselines are FedAvg [21], FedProx with $\mu = 0.01$ [25], FedNova [17] and FedAdam [24]. All share architecture, optimiser, learning rate, batch size, participation schedule, partition and seed.

Drift model parameters (L^*, A, D) are estimated on a calibration seed held out from the reported runs, so that τ^* is not tuned on the data against which it is evaluated. Every table and figure is generated directly by a named script from the run output; no numeric result is transcribed manually. Supplementary Appendix B relates the implementation to the reported values.

Energy is computed from the analytical model of Section III-C and is not measured by instrumentation. This is standard practice in the federated learning literature, but the reported values are model outputs rather than measurements.

IV. RESULTS AND DISCUSSION

Results are grouped as follows. Sections IV-A to IV-C establish and test the theory. Sections IV-D to IV-G evaluate the method against baselines, including the comparison in which it does not prevail. Sections IV-H to IV-K report supporting measurements.

Unless stated otherwise: $N = 20$, $\alpha_{\text{Dir}} = 0.5$, NB-IoT radio parameters, $V = 10^3$, three seeds, with standard deviations across seeds.

A. Calibration of the Loss Model

Section III-G left the regime of Eq. (12) as an empirical question. Table V reports the fit obtained by sweeping τ at fixed R on the calibration seed.

TABLE V: FITTED DRIFT MODEL PARAMETERS

L	A	D	95% CI	$p(H_0 : D = 0)$	R^2
0.536	0.235	0.000	$[-0.011, +0.011]$	0.500	0.974

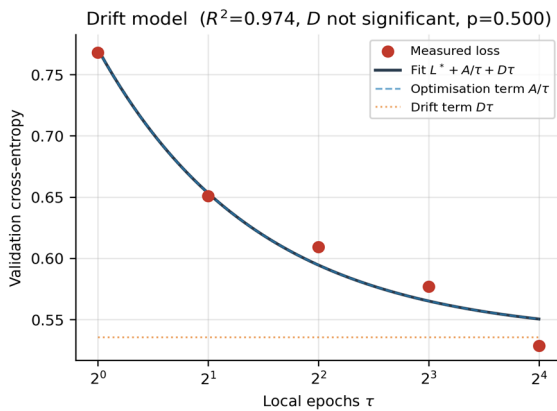


Fig. 3. Fitted loss model versus measured validation loss, with the dashed curve representing the optimization term alone.

The drift coefficient is not distinguishable from zero under a one-sided test. Validation loss falls monotonically across the swept range, taking the values 0.768, 0.651,

0.609, 0.577 and 0.529 at $\tau = 1, 2, 4, 8, 16$, with no turning point. The federation is therefore in the diminishing-returns regime, and the client-drift mechanism, though present in the model and in the literature, is inactive at this scale and heterogeneity. Fig. 3 shows the fitted loss model compared with the measured validation loss, with the dashed curve representing the optimization term alone.

This is reported as a negative result rather than suppressed. It has a direct consequence for the interpretation of Section IV-E: with $D = 0$ the loss continues to fall in τ throughout the feasible range, so from the standpoint of energy-to-target alone, large τ is preferable.

B. Dependence of the Optimal Epoch Count on the Communication Energy Fraction

Theorem 1 predicts that τ^* increases with the communication energy fraction. Fig. 4 tests this by varying the uplink rate to sweep ρ across $[0.01, 0.95]$ at four values of the energy weight α , comparing the closed-form prediction against the empirical minimiser of J over the epoch grid.

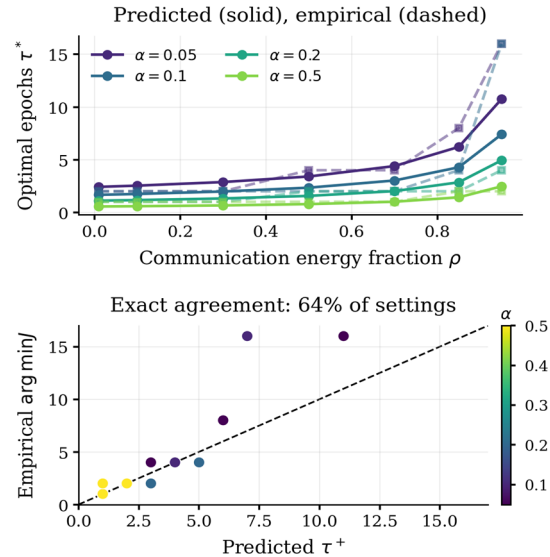


Fig. 4. Optimal local epoch count as a function of communication energy fraction (predicted vs. empirical).

The predicted values increase monotonically in ρ at every α , from 2.42 to 10.76 at $\alpha = 0.05$ and from 0.55 to 2.47 at $\alpha = 0.5$. Agreement with the empirical minimiser is exact for 64% of the points on integers and 86% when evaluated against the dyadic sweep grid $\{1, 2, 4, 8, 16\}$; the residual disagreement arises where the prediction falls between grid points and cannot be matched exactly. Fig. 4 shows the optimal number of local epochs as a function of the communication energy fraction, comparing the predicted values with the corresponding empirical results.

C. An Independent Test of the Same Prediction

Varying the uplink rate is one way to move ρ . Quantisation width is another, physically distinct mechanism: reducing b shrinks the payload $S = bd$ and hence the communication energy, without altering the channel. Table VI reports this test.

TABLE VI: EFFECT OF QUANTIZATION WIDTH SINGLE SEED

b (bit)	ρ	E (kJ)	$\bar{\tau}_r$	Acc.	F_1
8	0.608	1.07	2.45	0.708	0.483
16	0.756	1.55	2.70	0.705	0.481
32	0.861	2.43	2.90	0.698	0.475

Training is unaffected by b , so the differences arise from the energy accounting and the epoch count it selects. The selected epoch count tracks ρ under quantisation as it

TABLE VII: COMPONENT ABLATION RESULTS ACROSS THREE SEEDS, WITH STANDARD DEVIATION IN PARENTHESES AND ENERGY-AWARE WEIGHTS DENOTED BY E-WTS

Arm	Adapt. τ	E-wts	Acc.	F_1	E (kJ)	$\bar{\tau}$	$\ p - \alpha\ _1$
A0 (FedAvg, $\tau = 2$)	No	No	0.678 (0.020)	0.462 (0.017)	2.01	2.00	0
A1	Yes	No	0.689 (0.023)	0.485 (0.022)	2.26	3.42	0
A2	No	Yes	0.643 (0.014)	0.384 (0.003)	1.83	1.00	0.081
A3 (full)	Yes	Yes	0.694 (0.015)	0.491 (0.010)	2.26	3.42	0.100

Table VIII repeats the comparison at three levels of statistical heterogeneity, with paired t -tests across seeds.

TABLE VIII: COMPONENT EFFECTS ACROSS VARIOUS HETEROGENEITY LEVELS BASED ON THREE SEEDS AND PAIRED T-TESTS

α_{dir}	Arm	Acc.	F_1	Contrast	Δ acc.	p
0.1	A0	0.608	0.350			
	A1	0.634	0.394	A1-A0	+0.026	0.046
	A3	0.634	0.387	A3-A1	-0.000	0.997
0.5	A0	0.678	0.462			
	A1	0.689	0.485	A1-A0	+0.011	0.038
	A3	0.694	0.491	A3-A1	+0.005	0.440
1.0	A0	0.699	0.486			
	A1	0.711	0.509	A1-A0	+0.012	0.023
	A3	0.710	0.508	A3-A1	-0.001	0.281

The adaptive epoch count improves on federated averaging at a fixed two epochs at every heterogeneity level, with $p < 0.05$ throughout, and the largest gain occurs under the strongest skew. This is the component that is supported.

The energy-aware aggregation shows no measurable effect. The contrast A3-A1 is within noise at all three levels, the p -values are 0.997, 0.440 and 0.281, and the sign of the effect changes across levels. The measured weight deviation $\|p - \alpha\|_1$ is approximately 0.1, small enough that Proposition 1 bounds the achievable difference tightly. The component is reported here as unsupported rather than presented as a contribution; Section IV-N states the consequence.

E. Comparison Against Tuned Baselines

Section IV-A established that loss falls monotonically in τ on this data. A baseline given a larger fixed epoch count will therefore reach a target accuracy in fewer rounds and, where communication dominates, at lower energy. Testing whether the adaptive controller improves on that is the decisive comparison, and it is reported in full. Table IX summarize the energy required to reach the target (kJ) across three seeds, while Table X compares the optimal configuration of each method, with the best setting highlighted in bold.

Against a tuned federated averaging baseline, the proposed method consumes 74% more energy to reach the same accuracy. Under the same protocol it also consumes

does under rate variation. That two independent physical mechanisms produce the same qualitative response is stronger evidence for Eq. (14) than either alone.

D. Ablation

Table VII isolates the two components of the method. Arm A1 applies the adaptive epoch count with federated averaging weights; arm A2 applies the energy-aware weights at a fixed epoch count.

41% more energy than tuned FedNova and 22% more than tuned FedProx, and 49% less than tuned FedAdam. The excess over federated averaging is positive in every seed (+90.7%, +68.2%, +34.3%), so its direction is unambiguous; at three seeds, however, the 95% interval on its magnitude is wide and the point estimate should be read as indicative rather than precise. Its cause is identified in Section IVA: with $D = 0$ there is no interior optimum in the loss, so the energy-minimal policy is simply the largest feasible epoch count, and a fixed baseline attains that trivially. The adaptive controller holds τ below that value to respect the energy budget.

TABLE IX: ENERGY TO TARGET (kJ) BY FIXED EPOCH COUNT FOR PROTOCOL P3 ACROSS THREE SEEDS, WITH THE BEST SETTING HIGHLIGHTED IN BOLD

Method	Seed	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 8$	$\tau = 16$
FedAvg	0	miss	3.33	2.64	1.52	1.27
	1	miss	2.87	1.61	1.03	0.85
	2	2.90	1.38	0.80	0.58	0.41
FedProx	0	miss	3.33	2.77	1.85	2.04
	1	miss	2.97	1.73	1.18	1.28
	2	2.90	1.48	0.80	0.58	0.62
FedNova	0	3.55	3.75	2.01	1.52	2.03
	1	miss	2.97	1.50	1.03	0.85
	2	2.81	1.57	0.80	0.58	0.41
FedAdam	0	3.27	3.65	4.15	4.72	5.59
	1	2.82	2.48	3.11	3.84	2.55
	2	2.81	2.56	2.50	3.19	4.14

TABLE X: COMPARISON OF ENERGY TO TARGET AT THE OPTIMAL CONFIGURATION FOR EACH METHOD ACROSS THREE SEEDS

Method	Best fixed τ	Energy to target (kJ)	mean τ
FedAvg	16	0.84	16.00
FedProx	8	1.21	8.00
FedNova	8	1.04	8.00
FedAdam	2	2.89	2.00
E2A-FO	adaptive	1.47	4.93

No energy advantage over a tuned fixed schedule is claimed.

Table XI states the setting in which the method retains an advantage. The 0.84 kJ figure is the cost of the winning run, but it is not available without first performing the sweep that identifies $\tau = 16$ as the winner. Counting that search, the adaptive controller reaches a comparable operating point at 81.4% lower total energy. The sweep

cost excludes runs that failed to reach the target, so it understates the true cost of the search. This is the amortised argument standard in automated hyperparameter selection, and it is the form in which the energy claim of this paper is made. It applies where the epoch count must be chosen afresh, as at a new deployment or after a change of hardware or channel, and not where a single setting can be tuned once and reused indefinitely.

TABLE XI: TOTAL ENERGY INCLUDING THE COST OF SELECTING THE EPOCH COUNT

Quantity	Energy (kJ)
FedAvg: sweep over $\tau \in \{1,2,4,8,16\}$	7.07
FedAvg: best single run	0.84
FedAvg: total to identify and deploy	7.91
E2A-FO: single run, no sweep	1.47
Reduction	81.4%

F. Behavior under an Energy Budget

TABLE XII: OPERATION UNDER A PER-ROUND ENERGY BUDGET (IN KJ) EVALUATED ACROSS THREE SEEDS

Config.	E (kJ)	Over-budget	Acc.	F_1	$\bar{\tau}$
FedAvg, $\tau = 2$	2.01	0.50	0.678	0.462	2.00
E2A-FO	2.26	0.75	0.691	0.490	3.38
FedAvg, $\tau = 8$	3.08	1.57	0.716	0.516	8.00
FedAvg, $\tau = 16$	4.49	2.99	0.739	0.545	16.00

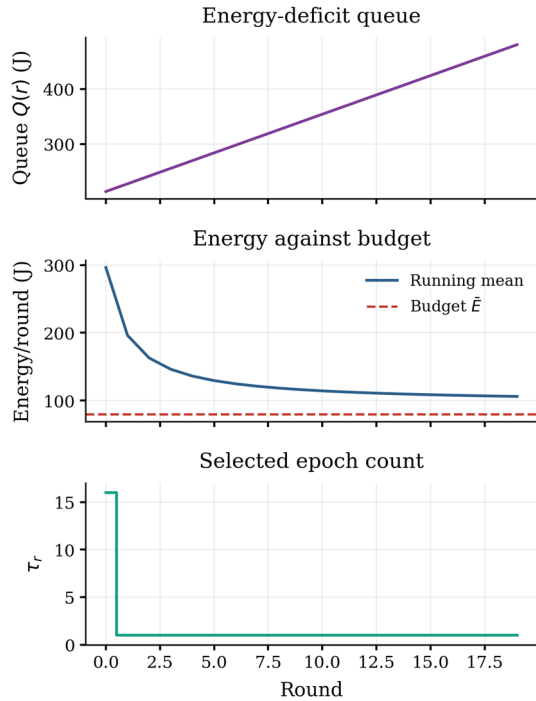


Fig. 5. Queue, energy versus budget, and selected epoch count metrics for a single run.

The budget is set to 75% of the mean per-round energy of federated averaging at $\tau = 2$, and the over-budget column reports the cumulative excess above it. Fixed epoch counts trace a monotone relation between accuracy and budget violation: higher τ buys accuracy at the cost of

exceeding the budget. The adaptive controller occupies an intermediate position. It does not minimize violation, since federated averaging at $\tau = 2$ violates less, and it does not maximise accuracy. What it does is arrive at an interior operating point without being supplied with the epoch count, which is the same property quantified in Table XII.

Fig. 5 shows the mechanism over one run: the virtual queue accumulates when consumption exceeds the budget, the running average energy converges toward the budget line, and the selected epoch count responds to the queue.

G. The Lyapunov Trade-off

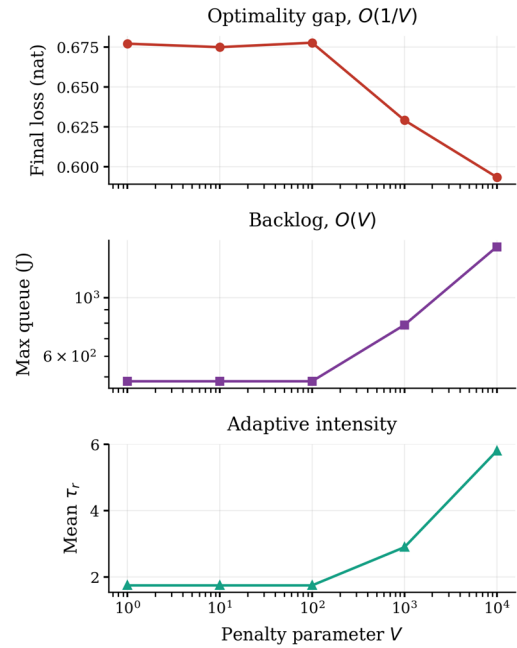


Fig. 6. Effect of the Lyapunov penalty parameter.

Fig. 6 shows the effect of the penalty parameter. Loss decreases and backlog grows as V increases, consistent with the $O(1/V)$ and $O(V)$ scaling of Theorem 3: final loss falls from 0.677 to 0.594 while the maximum queue rises from 481 J to 1561 J and the mean selected epoch count from 1.75 to 5.80 across $V \in [10^0, 10^4]$. Two observations qualify this. The response is flat below $V = 10^3$, so the parameter has no effect over three decades. And the method is sensitive to V in a way that matters practically: at $V = 10^2$ it fails to reach the target accuracy on two of three seeds under protocol P3, while at $V = 10^3$ it reaches it on all three. V is therefore a parameter requiring selection, and the claim that the closed form eliminates hyper parameter search applies to τ and not to V .

H. Verification of the Bias Bound

The constant μ is estimated from the Polyak-Lojasiewicz inequality at the final iterate using a conservative surrogate for F^* , so the margin is indicative rather than tight. The measured difference between the two objectives lies well inside the bound. The same small

weight deviation that makes the bound comfortable also explains the absence of an aggregation effect in

Table XIII weights close to the sample proportions cannot produce large differences in either direction.

TABLE XIII: EMPIRICAL EVALUATION OF PROPOSITION 1

ζ	$\ \mathbf{p} - \alpha\ _1$	$\hat{\mu}$	Bound, Eq. (16)	Measured difference	Satisfied
4.723	0.120	0.547	0.292	0.021	Yes

I. Sensitivity to the Communication Regime

Table XIV illustrates the energy required to reach the target (in kilojoules, kJ) across different communication regimes. The results are evaluated using two random seeds.

Baselines run at $\tau = 2$ throughout. The benefit rises with ρ and reverses at low ρ , where the method consumes

more energy than the baseline. This sign change is predicted by Theorem 1: as $\rho \rightarrow 0$ the optimal epoch count approaches unity and no local-work advantage exists to be exploited. The negative result is retained because a model that identifies its own domain of applicability, and is correct at the boundary, is better supported than one reporting gains everywhere.

 TABLE XIV: ENERGY TO TARGET (kJ) ACROSS COMMUNICATION REGIMES: TWO SEEDS; $\bar{\tau}$ IS THE MEAN SELECTED EPOCH COUNT

Regime	Uplink	ρ	FedAvg	E2A-FO	$\bar{\tau}$	Diff.
WLAN	10 Mbit/s	0.022	0.80	0.98	1.83	-23.4%
NB-IoT	50 kbit/s	0.873	1.45	1.04	4.72	+28.5%
LoRa	5.5 kbit/s	0.970	5.74	2.36	7.29	+58.9%

The WLAN row is retained deliberately. A framework evaluated only where ρ is large would leave its domain of applicability untested; reporting the regime in which the closed form returns $\tau^+ = 1$, and in which the method is accordingly outperformed, establishes that the prediction is falsifiable.

J. Carbon

Emissions differ by more than an order of magnitude across the range of grid intensities, while the ranking of methods is unchanged, as Proposition 3 requires. Absolute emissions at this scale are small; the quantity of interest is the relative ordering and its invariance to γ , as Table XV shows. Claims regarding environmental benefit are confined to what these numbers support: a reduction in estimated emissions proportional to the reduction in

energy, under a stated conversion, in the deployment setting of Table XI.

K. Client-level Outcomes

Fig. 7 shows the distribution of per-client accuracy. Section III-G raised the concern that down-weighting energy-expensive clients could disadvantage resource-poor devices. Relative to federated averaging at $\tau = 2$, the proposed method slightly improves both dispersion (standard deviation 0.129 against 0.139) and the worst-served client (0.408 against 0.403), so that concern is not realized. Relative to federated averaging at $\tau = 8$, it does not: the tuned baseline reaches 0.431 on the worst client and a lower dispersion of 0.103. The pattern matches Table X, and the explanation is the same: more local work per round improves all three quantities on this data.

TABLE XV: EQUIVALENT EMISSIONS UNDER FOUR GRID INTENSITIES

Method	Energy (kJ)	30 gCO ₂ e/kWh	250	500	800
FedAvg	2.12	1.77×10^{-5}	1.47×10^{-4}	2.94×10^{-4}	4.71×10^{-4}
E2A-FO	2.06	1.72×10^{-5}	1.43×10^{-4}	2.86×10^{-4}	4.58×10^{-4}

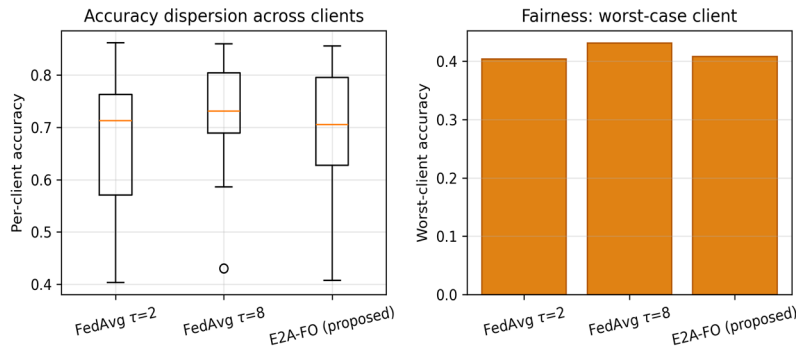


Fig. 7. Per-client accuracy distribution and worst-client accuracy.

L. Discussion

1) Why the optimum moves with the communication fraction?

The dependence of τ^* on ρ follows from the structure of Eq. (13). At fixed R , the energy term grows linearly in τ with slope $\alpha(1 - \rho)$, while the loss term falls as A/τ .

Raising ρ shrinks the slope of the energy term without altering the loss term, so the crossing point moves to a larger τ . When a large share of the energy is spent transmitting parameters, the marginal cost of an additional local epoch is small relative to the cost of an additional synchronization, and the optimum shifts towards more local work.

Two features of Sections IV-B and IV-C distinguish this result from a coincidental agreement. First, the response was produced by two interventions that share no mechanism: lowering the uplink rate lengthens transmission time, while narrowing the quantization width shortens the payload. Both act on ρ and nothing else in the model, and both move the selected epoch count in the predicted direction. Second, the relation has a boundary. At $\rho = 0.022$ the closed form returns $\tau^+ = 1$, and Table XIV shows the method consuming more energy than the baseline in that regime. A prediction that identifies the conditions under which it ceases to hold is more informative than one reporting gains at every setting.

2) Adaptive selection against a tuned schedule

Table XI shows a tuned fixed schedule reaching the target accuracy at lower energy than the adaptive controller. Table V gives the reason. With the fitted drift coefficient at zero, the loss falls monotonically in τ across the whole feasible range, so no interior optimum exists in the loss and the energy-minimal policy is the largest feasible epoch count. A fixed schedule set at the largest feasible value attains that policy directly. The controller does not, because the queue in Eq. (17) reduces τ_r whenever consumption exceeds the budget.

This outcome is a property of the operating regime, not of the control law. The controller is solving a constrained problem while the fixed schedule solves an unconstrained one, and Table XII shows the two occupying different points on the same accuracy-against-violation curve. Where no budget binds and drift is absent, the two problems coincide, and the constrained solution cannot improve on the unconstrained one.

The controller supplies the setting itself. Table X required five runs per method to establish that $\tau = 16$ is the best fixed choice, and that determination is specific to the model, the partition and the channel used here. Table XII gives the resulting comparison: 7.91 kJ to identify and deploy a fixed schedule against 1.47 kJ for a single adaptive run. The advantage is therefore the avoided search, and it applies whenever that search must be repeated.

3) Heterogeneity and the aggregation rule

The aggregation rule of Eq. (20) reweights clients by energy and by local loss, and Table IX shows the resulting effect to be indistinguishable from zero at all three heterogeneity levels. The measured weight deviation explains why: at $\|\mathbf{p} - \boldsymbol{\alpha}\|_1 \approx 0.1$ the weights remain close to the sample proportions, and by Proposition 1 the attainable difference in either direction is correspondingly small.

The deviation is small because the quantities entering Eq. (20) vary little across clients in this configuration. Computation energy is proportional to n_k and to f_k^2 , and

with clock frequencies drawn from a factor-of-four range the normalized scores $\tilde{E}_k$ span a narrow interval. A federation with wider hardware disparity would produce larger deviations, and Proposition 1 indicates that the bias relative to the classical objective would grow with them. These experiments do not establish whether the accuracy effect grows in the same proportion.

4) Practical reading

Three consequences follow for deployment. Where communication is expensive relative to computation, the epoch count is the parameter worth adapting, and Eq. (14) gives it without search. Where bandwidth is plentiful, $\tau^+ = 1$ and the mechanism is inactive, so the added machinery is not warranted. And where a long-run energy budget must be respected, the queue formulation supplies a policy with a proven guarantee, at the cost of the accuracy that an unconstrained schedule would obtain.

The Lyapunov parameter requires attention in all three cases. Fig. 6 shows no response below $V = 10^3$ and a marked response above it, and Section IV-G records that at $V = 10^2$ the controller failed to reach the target accuracy on two of three seeds. The closed form removes the search over τ but not the choice of V .

M. Threats to Validity

Experiments use a stratified subsample of the full dataset rather than its entirety. The qualitative findings, namely the sign of the ρ dependence, the absence of drift and the ranking against tuned baselines, are structural and are not expected to change with scale, but the absolute values will. Between three and five seeds are used per experiment, and the headline figures are reported with 95% confidence intervals across seeds. At these sample sizes the intervals are wide: the amortized saving of Table XI spans approximately 63% to the upper bound of the admissible range, and the comparison of Section IV-E, although positive in every seed, has an interval that includes zero. The conclusions drawn here therefore rest on the sign and ordering of effects, which are stable across seeds, and not on the precise magnitudes, which are not established at this sample size. Paired t -tests are reported where five seeds are available; the Wilcoxon signed-rank test cannot attain $p < 0.0625$ at $n = 5$, which is a property of the test rather than of the effect it is applied to.

Energy is computed from Eq. (3) and Eq. (5) rather than measured on hardware. The functional forms follow established models, but the absolute values depend on parameters taken from the literature, and a deployment with different device or radio characteristics would produce different magnitudes even where the ordering held. All results come from CICIOT2023; Section V states the intended extension to further datasets.

One confound requires explicit attention. In the scalability setting, A_{comm} scales with the number of clients while A_{comp} is fixed by the total sample count, so ρ rises with N automatically. Differences observed across federation sizes therefore partly re-express the ρ dependence of Section IV-B and should not be read as an independent effect of federation size.

N. Limitations

The energy advantage reported here is conditional in two respects. It applies where the local epoch count must be selected rather than assumed, as quantified in Table XI, and where communication accounts for a substantial share of total energy, as Table XIV shows. In high-bandwidth settings the method consumes more energy than federated averaging. The aggregation rule of Eq. (20) is retained for its analytical properties and for the exact reduction to federated averaging established in Proposition 2, but Table IX gives no empirical support for it and no performance claim is made on its behalf. The comparison against tuned fixed schedules now covers all four baselines; the proposed method is not the cheapest of them at a tuned setting, and the advantage it retains is the amortized one of Table XI.

The closed form removes the need to search over τ but not the need to choose V , and Section IV-G shows the method to be sensitive to that choice near $V = 10^3$. Carbon is reported as a derived quantity: under a constant emission factor it carries no information beyond energy, by Proposition 3, and carbon-aware optimization would require the time-varying intensity data that this work does not use. Finally, the verification of Proposition 1 in Table XIII rests on an estimated Polyak-Lojasiewicz constant with a conservative surrogate for the optimum, so the reported margin is indicative rather than a tight bound.

V. CONCLUSION AND FUTURE WORK

This paper formulated the energy-communication-accuracy trade-off in federated learning as a normalized multi-objective problem, derived a closed-form expression for the local epoch count that minimizes it, and embedded that expression in a Lyapunov controller with proven mean-rate stability and an $O(1/V)$ optimality gap. The predicted dependence of the optimum on the communication energy fraction was confirmed through two independent physical mechanisms and shown to reverse at the boundary the theory identifies.

The adaptive controller improves on federated averaging at a fixed epoch count of two at three levels of statistical heterogeneity, with $p < 0.05$ throughout. Against federated averaging tuned across a grid of fixed epoch counts it consumes more energy in a single run; its advantage is that it reaches a comparable operating point without that grid search, corresponding to an 81.4% reduction in total energy once the search is counted. The energy-aware aggregation rule, proposed as a second component, showed no measurable effect and is reported as unsupported.

All energy figures reported in this paper are outputs of the analytical model of Section III-C rather than hardware measurements, and the quantitative claims should be read subject to that scope.

Four directions follow directly. First, evaluation on additional federated benchmarks: FEMNIST and CIFAR-10 under Dirichlet partitioning, and a second IoT security dataset such as N-BaIoT, to establish whether the diminishing-returns regime observed here is a property of

this dataset or of tabular intrusion detection more generally. This extension is planned and requires only the substitution of the model and data loader in the released implementation. Second, incorporation of the round count into the objective, which would allow the controller to account for the effect of the epoch count on rounds-to-target and would address the gap identified in Section IV-E. Third, carbon-aware scheduling under time-varying grid intensity, where Proposition 3 no longer applies and the carbon-optimal schedule diverges from the energy-optimal one. Fourth, evaluation in the drift regime, at larger epoch counts and stronger heterogeneity than those reached here, where the inactive term of Eq. (12) would become operative.

SUPPLEMENTARY MATERIAL AND DATA AVAILABILITY

Two supplementary documents accompany this manuscript.

Supplementary A: Derivations. Assumptions 1 to 7; the proof of Theorem 1 and Corollary 1; Definition 1 and Lemma A.1 for the heterogeneity constant of the energy-weighted objective; Theorem A.1 and the proof of Proposition 1; and the proof of Lemma 2 and Theorem 3.

Supplementary B: Implementation and output. The correspondence between the equations of Sections III and IV and the code that evaluates them, with the program output from which Tables V, VI and X to XI are generated.

The implementation is released under an open license and archived with a persistent identifier. It comprises the federated training engine, the energy and carbon models, the calibration and protocol drivers, and the scripts that generate every table and figure reported here. Configuration is held in a single file; seeds, partitions and energy parameters are not embedded in the analysis code. Raw run logs are released alongside the source, so that each reported value can be traced to the run that produced it. CICIOT2023 is publicly available from the Canadian Institute for Cybersecurity [35].

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The author conceived the study, developed the theoretical framework, implemented the software, conducted the experiments, analyzed the results, wrote the manuscript, and approved the final version.

ACKNOWLEDGMENT

The author thanks colleagues at the Electrical Engineering Department, Al-Balqa Applied University, for helpful discussions.

REFERENCES

- [1] Y. Sun, H. Ochiai, and H. Esaki, "Decentralized deep learning for multi-access edge computing: A survey on Communication Efficiency and Trustworthiness," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 963–972, Dec. 2022. doi: 10.1109/TAI.2021.3133819

- [2] M. R. A. Berkani, A. Chouchane, Y. Himeur *et al.*, "Advances in federated learning: applications and challenges in smart building environments and beyond," *Computers*, vol. 14, no. 4, 2025. doi: 10.3390/computers14040124
- [3] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021. doi: 10.1109/COMST.2021.3075439
- [4] X. Hou, J. Wang, F. Guan, J. Du, C. Jiang, and Y. Ren, "Energy-efficient federated learning for edge real-time vision via joint data, computation, and communication design," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 12, pp. 4000–4014, 2025.
- [5] P. Zheng, Y. Zhu, Y. Hu, Z. Zhang, and A. Schmeink, "Federated learning in heterogeneous networks with unreliable communication," *IEEE Transactions on Wireless Communications*, vol. 23, no. 4, pp. 3823–3838, 2024. doi: 10.1109/TWC.2023.3311824
- [6] J. Palomares, C. Camerota, E. Coronado *et al.*, "Taming Bandwidth Bottlenecks in Federated Learning via ECN-based Gradient Compression," in *Proc. 21st Int. Conf. on Network and Service Management*, Bologna, Italy, 2025. doi: 10.23919/CNSM67658.2025.11297573
- [7] R. Jin, X. He, and H. Dai, "Communication efficient federated learning with energy awareness over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5204–5219, 2022. doi: 10.1109/TWC.2021.3138394
- [8] R. Chen, Q. Wan, X. Zhang, X. Qin *et al.*, "Wireless-Aware energy-efficient federated learning over mobile devices via algorithm and hardware co-design," *IEEE Transactions on Networking*, vol. 34, pp. 4283–4298, 2026. doi: 10.1109/TON.2026.3671076
- [9] Z. Gao, Y. Xiang, S. Lu, and Y. Liu, "An optimized updating adaptive federated learning for pumping unit's collaborative diagnosis with label heterogeneity and communication redundancy," *Engineering Applications of Artificial Intelligence*, vol. 152, no. 11, 2025. doi: 10.1016/j.engappai.2025.110724
- [10] T. Nishio and R. Yonetani, "Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge," in *Proc. 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, 2019, pp. 1–7. doi: 10.1109/ICC.2019.8761315
- [11] Y. Lin, K. Wang, W. Huang, and Z. Ding, "Federated Learning Over Wireless Networks: Optimizing Performance With NOMA and Power Allocation," *IEEE Trans. on Wireless Communications*, vol. 25, pp. 4467–4479, 2026. doi: 10.1109/TWC.2025.3611589
- [12] T. Zhao, X. Chen, Q. Sun, and J. Zhang, "Energy-Efficient Federated Learning Over Cell-Free IoT Networks: Modeling and Optimization," *IEEE Internet of Things Journal*, vol. 10, no. 19, pp. 17436–17449, 2023. doi: 10.1109/IIOT.2023.3273619
- [13] K. Mohamed, D. Badis, and Y. Ali, "An efficient federated learning solution for the artificial intelligence of things," *Future Generation Computer Systems*, vol. 163, issue: 3, 2025. doi: <https://doi.org/10.1016/j.future.2024.107533>
- [14] A. Lameesa, M. Alam, S. Ahmed, G. Liu *et al.*, "Federated learning for sustainable IoT: Energy-efficient applications," *Energy Reports*, vol. 15, 109145, 2026. doi: 10.1016/j.egy.2026.109145
- [15] C. Battiloro, P. Di Lorenzo, M. Merluzzi, and S. Barbarossa, "Lyapunov-Based Optimization of Edge Resources for Energy-Efficient Adaptive Federated Learning," *IEEE Transactions on Green Communications and Networking*, vol. 7, no. 1, pp. 265–280, March 2023. doi: 10.1109/TGCN.2022.3186879
- [16] I. Soomro, H. Rehman, S. Hussain *et al.*, "ROCHE: A Robust and End-to-End Privacy-Preserving Federated Learning Framework for Intrusion Detection in Industrial Internet of Things," *IEEE Internet of Things Journal*, vol. 12, no. 24, pp. 52357–52377, 2025. doi: 10.1109/IIOT.2025.3612944
- [17] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 7611–7623, 2020.
- [18] N. S. E. Hassan, "Energy-Aware FedDeepRiskNet: Optimized Federated Learning for Edge-Enabled Multi-Hospital Systems," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 2026. doi: 10.1177/18758967261435711
- [19] M. Mozafari, M. Hosseini, P. Hosseini, A. Zahedi, J. Abouei, and A. Mohammadi, "Energy-Efficient Federated Learning for IoT Networks With Massive MIMO-Enabled SWIPT," *IEEE Internet of Things Journal*, vol. 13, no. 13, pp. 29871–29886, 2026. doi: 10.1109/IIOT.2026.3686337
- [20] S. P. Singh, M. Omar, D. Sumathi, and P. Karthik, "Federated Learning-Based Multi-Objective Optimization for IoT-Enabled Distributed Environmental Monitoring in Consumer Electronics," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 3, pp. 8979–8988, Aug. 2025. doi: 10.1109/TCE.2025.3594404
- [21] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.
- [22] I. Alzamil, "Federated Deep Learning for Scalable and Explainable Load Forecasting in Privacy-Conscious Smart Cities," *IEEE Access*, vol. 13, pp. 142237–142250, 2025. doi: 10.1109/ACCESS.2025.3587058
- [23] A. Mahmoudi, M. Zaher and E. Björnson, "Low-Latency and Energy-Efficient Federated Learning Over Cell-Free Networks: A Trade-Off Analysis," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 2274–2292, 2025. doi: 10.1109/OJCOMS.2025.3553593
- [24] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konecny, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *Proc. 9th Int. Conf. on Learning Representations (ICLR)*, 2021.
- [25] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Machine Learning and Systems (MLSys)*, vol. 2, 2020, pp. 429–450.
- [26] S. Wang, J. Liu, H. Xu, C. Tang, Q. Ma, and L. Huang, "Toward Communication-Efficient Decentralized Federated Graph Learning Over Non-IID Data," *IEEE Transactions on Mobile Computing*, vol. 25, issue 5, pp. 6929–6947, May 2026. doi: 10.1109/TMC.2025.3641696
- [27] G. Shen, X. Guo, H. Yuan, J. Du, W. Zhang, Y. Neil, and X. Kong, "Communication-efficient decentralized federated graph learning via knowledge distillation under dual heterogeneity," *Information Sciences*, vol. 739, 123192, 2026. doi: 10.1016/j.ins.2026.123192
- [28] M. Mhawish, M. Aljamal, A. Alsarhan, B. Khassawneh, A. Akhuzada, and A. Al-Shamayleh, "A Novel Adaptive Generalization-Guided Federated Learning for Multimodal Industrial IoT Intrusion Detection Under Statistical Heterogeneity," *IEEE Open Journal of the Computer Society*, vol. 7, pp. 1414–1425, 2026. doi: 10.1109/OJCS.2026.3703977
- [29] H. Hashempour, M. Nozari, G. Berardinelli, Y. Li, J. Zhang, and H. Ngo, "Energy-Efficient Federated Learning With Relay-Assisted Aggregation in IIoT Networks," *IEEE Internet of Things Journal*, vol. 13, no. 15, pp. 33403–33418, 2026. doi: 10.1109/IIOT.2026.3696125
- [30] A. S. E. Elahi, S. Khanam, and N. Rahman, "Optimizing federated learning efficiency: accelerating model averaging with cluster-based approach using class occurrences-fedCCO," *Neural Computer and Application*, vol. 37, pp. 15369–15387, 2025. doi: 10.1007/s00521-025-11289-0
- [31] A. Abbasi, F. Dong, X. Wang, H. Leung, J. Zhou, and S. Drew, "FedGreen: Carbon-Aware Federated Learning with Model Size Adaptation," in *Proc. 2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, Denver, CO, USA, 2024, pp. 1352–1358. doi: 10.1109/ICCWorkshops59551.2024.10615720
- [32] P. McEnroe, S. Wang, and M. Liyanage, "A Survey on the Convergence of Edge Computing and AI for UAVs: Opportunities and Challenges," *IEEE Internet of Things Journal*, vol. 9, no. 17, pp. 15435–15459, 2022. doi: 10.1109/IIOT.2022.3176400
- [33] Z. Lin, G. Qu, W. Wei, X. Chen, and K. K. Leung, "AdaptSFL: Adaptive Split Federated Learning in Resource-Constrained Edge Networks," *IEEE Transactions on Networking*, vol. 33, no. 6, pp. 2993–3008, Dec. 2025. doi: 10.1109/TON.2025.3577790
- [34] H. Larian and F. Safi-Esfahani, "InTec: integrated things-edge computing: a framework for distributing machine learning pipelines in edge AI systems," *Computing*, vol. 107, 2025. doi: 10.1007/s00607-024-01388-6

- [35] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," *Sensors*, vol. 23, no. 13, 2023. doi: 10.3390/s23135941

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](#)).



Almamoon Alauthman was born in Jordan in 1985. He received his Bachelor's degree in computer engineering from Al-Balqa Applied University (BAU) in Amman, Jordan, and his master's degree in computer engineering from Jordan University of Science and Technology (JUST) in Irbid, Jordan. He completed his Ph.D. at Sultan Zainal Abidin University (UniSA) in Malaysia. Currently, he serves as an assistant professor in the Electrical Engineering Department at the Faculty of Engineering, Al-

Balqa Applied University. His research interests include federated learning, energy-efficient distributed computing, Internet of Things systems, parallel processing, neural networks, computer architecture and organization, and wireless networks.



Feras Alfaqih is an accomplished academic and researcher with over 45 publications in internationally recognized, peer-reviewed journals indexed in Scopus and Web of Science. He has held several leadership positions at Al-Hussein Bin Talal University, including the dean of the Faculty of Science, the vice dean, and the head of the Department of Mathematics, while contributing significantly to postgraduate education, curriculum development, quality assurance, and academic accreditation. His research focuses on applied mathematics, numerical analysis, and computational modeling, with particular expertise in the numerical simulation of complex fluid flows, Nano fluids, magneto hydrodynamics, and heat-transfer phenomena.