



OPEN ACCESS

EDITED BY

Osama Amin,
King Abdullah University of Science and
Technology, Saudi Arabia

REVIEWED BY

Ahmad Bazzi,
New York University Abu Dhabi, United Arab
Emirates

Selvaraj Kandasamy,
PSNA College of Engineering and Technology,
India

*CORRESPONDENCE

Zhenning Chen,
✉ link_chen@yeah.net

RECEIVED 18 November 2025

REVISED 08 January 2026

ACCEPTED 12 January 2026

PUBLISHED 09 February 2026

CITATION

Chen Z, Xu Z, Ding Y and Wang Y (2026) Data-
and distance-aware clustering for scalable
wireless federated learning.
Front. Commun. Netw. 7:1748815.
doi: 10.3389/frcmn.2026.1748815

COPYRIGHT

© 2026 Chen, Xu, Ding and Wang. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in
other forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Data- and distance-aware clustering for scalable wireless federated learning

Zhenning Chen^{1*}, Zihe Xu², Yihan Ding³ and Youren Wang¹

¹College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, ²Jiangsu Key Laboratory of Wireless Communications and IoT, Nanjing University of Posts and Telecommunications, Nanjing, China, ³School of Information Science and Technology, Dalian Maritime University, Dalian, China

Introduction: Federated learning (FL) enables model training on edge devices using local data while aggregating model updates at a central server without exchanging raw data, thereby preserving privacy. However, achieving satisfactory convergence accuracy with low communication energy remains challenging. This work investigates a three-tier clustered FL (CFL) architecture to improve global training performance and communication efficiency through joint device clustering and resource scheduling.

Methods: We analyze how clustering strategies influence learning convergence and communication energy consumption. Based on this analysis, we propose a clustering method that jointly accounts for gradient cosine similarity and communication distance. A simplified procedure is further developed for device association and cluster-head selection, with the goals of improving intra-cluster data balance and reducing the overall communication distance to the server.

Results: Simulations demonstrate that the proposed method consistently improves model accuracy while reducing communication energy consumption compared with random clustering and similarity-based clustering baselines.

Discussion: These results indicate that jointly considering update similarity and communication distance in CFL can effectively balance learning quality and communication cost, offering a practical approach for energy-efficient federated training in edge networks.

KEYWORDS

device clustering, energy efficiency, federated learning, gradient similarity, wireless communications

1 Introduction

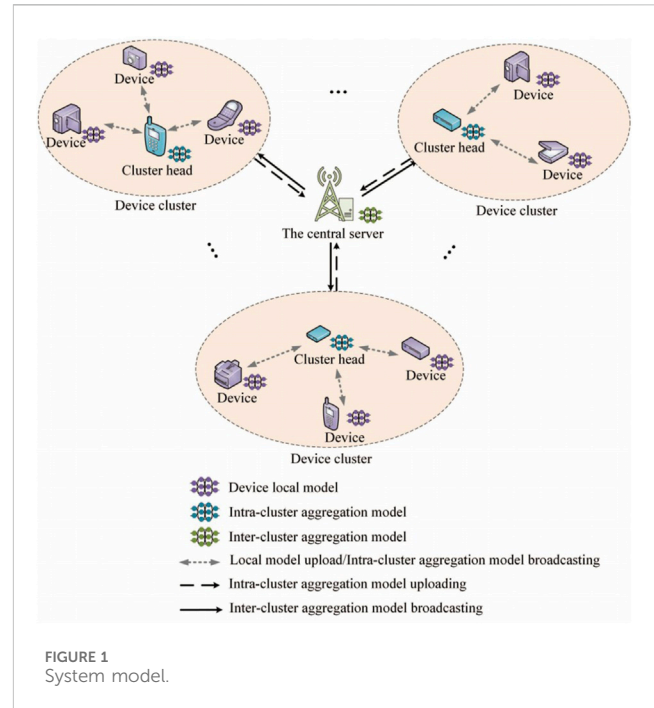
The explosive growth of mobile edge devices and their local storage of data has challenged traditional centralized machine learning paradigms. Centralized approaches necessitate frequent data transmission to the central server, which results in significant privacy risks and high communication overhead (Xia et al., 2020). Federated learning (FL) has emerged as a promising distributed learning paradigm, which enables model training by aggregating locally trained models at the central server without directly transmitting raw data, thereby preserving data privacy (Konečný et al., 2015). However, FL still relies on frequent communication

between devices and the server, which becomes inefficient in wireless environments with heterogeneous data distributions (Lu et al., 2024).

During the evolution of FL, extensive efforts have been devoted to mitigating communication overhead from various perspectives. Hierarchical FL reduces communication frequency through staged aggregation, thereby lowering communication costs (Briggs et al., 2020; Tran et al., 2025); however, its rigid hierarchical design often limits adaptability in dynamic environments. Prototype-based clustered FL is a representative branch of personalized FL, which aligns client features through global or local prototypes, thereby effectively reducing communication costs by minimizing the need for frequent model exchanges (Yang et al., 2024; Tan et al., 2022). However, such methods often compromise global consistency, thereby limiting scalability when tasks are only partially shared. Moreover, resource-aware FL introduces client selection under bandwidth and computation constraints to improve communication efficiency but typically decouples resource optimization from model aggregation, neglecting their intrinsic interplay (Nishio and Yonetani, 2019). Against this backdrop, clustered FL (CFL) has become a more direct and effective solution. CFL reduces communication costs and improves convergence stability by performing intra- and inter-cluster aggregation after grouping clients based on their statistical or geographical characteristics (Ghosh et al., 2021; Zeng et al., 2023). For example, Yan et al. (2024) proposed adaptive clustering strategies that enable flexible client participation and dynamic cluster formation, thereby reducing redundant communication under non-independent and identically distributed (Non-IID) conditions. Similarly, Gao et al. (2023) clustered clients based on the similarity of their local data distributions and used acceleration algorithms to shorten training time and lower communication overhead. In addition, Zhang et al. (2024) developed an adaptive CFL framework that adjusts cluster size and communication intervals through online similarity measurement, thereby improving both robustness and communication efficiency. Despite these advances, existing CFL approaches still suffer from several limitations: 1) existing studies rarely explore how the clustering strategy influences both data richness and convergence dynamics. Over-reliance on data similarity for clustering may reduce intra-cluster diversity, thereby weakening model generalization. 2) Prior CFL methods primarily rely on gradient similarity or geographic proximity for clustering but often ignore joint optimization of learning performance and resource efficiency.

To bridge this research gap, we propose a data- and distance-aware clustering scheme. The proposed scheme exploits data distribution characteristics and geographical information to optimize cluster head selection and device association scheduling prior to the training process. Based on this clustering result, the CFL training procedure is subsequently carried out. The main contributions of this study are summarized as follows.

- We propose a CFL framework that collectively considers learning performance and communication cost. On the learning side, a convergence analysis is conducted to theoretically demonstrate that enhancing the diversity and representativeness of intra-cluster data effectively improves the convergence behavior of CFL under data heterogeneity.



On the communication side, the communication cost is modeled in terms of the transmission distance. Based on this model, a combined optimization problem for cluster head selection and device association scheduling is formulated, which simultaneously accounts for learning performance and communication cost.

- Based on the formulated joint optimization objective, we develop an iterative algorithm to efficiently solve the cluster head selection and device association scheduling problem. The proposed algorithm decomposes the original NP-hard problem into two tractable subproblems, which are solved in an alternating optimization manner.
- Simulation results demonstrate that the proposed method consistently outperforms the three baseline algorithms in terms of model accuracy and communication energy efficiency, thereby validating the effectiveness of the proposed framework.

2 System model

We consider a wireless CFL system, as illustrated in Figure 1, which consists of a central server and a set of devices $\mathcal{U} = \{1, 2, \dots, U\}$. Each device $u \in \mathcal{U}$ has a local dataset $\mathcal{D}_u = \{(x_u^i, y_u^i)\}_{i=1}^{D_u}$, where $D_u = |\mathcal{D}_u|$. The communication distances are denoted by $d_{u,u'}$ for the links between devices and $d_{u,cs}$ for the links between devices and the server. Before training, a subset of high-performance devices $\mathcal{N} = \{1, 2, \dots, N\}$ can be selected from the set of devices as candidates for cluster heads, and all devices are grouped through two scheduling strategies.

A binary variable $a_n \in \{0, 1\}$ indicates whether the device n is selected as a cluster head. The set of cluster heads is \mathcal{N}_{ch} , satisfying $a_n = 1$ for all $n \in \mathcal{N}_{ch}$. The number of clusters is N_{ch} . We also introduce a binary variable $b_{u,n} \in \{0, 1\}$ that

indicates whether the device u is associated with the cluster head n . Each cluster is defined as $C_n = \{u \in \mathcal{U} \mid b_{u,n} = 1\}$, and the overall cluster set is $\mathcal{C} = \{C_1, C_2, \dots, C_N\}$, where $C_n \cap C_{n'} = \emptyset$ for $n \neq n'$ and $|C_n| \geq 1$. The total number of samples in cluster n is $D_n = \sum_{u \in \mathcal{U}} b_{u,n} D_u$.

The global loss function is defined in Equation 1:

$$F(\mathbf{w}) = \sum_{n \in \mathcal{N}} \frac{a_n D_n}{D} \sum_{u \in \mathcal{U}} \frac{b_{u,n} D_u}{D_n} F_u(\mathbf{w}). \quad (1)$$

Here, $F_u(\mathbf{w}) = \frac{1}{D_u} \sum_{i=1}^{D_u} \ell(x_u^i, y_u^i, \mathbf{w})$ is the local loss at device u , and $D = \sum_{u \in \mathcal{U}} D_u$ is the total number of training samples. The intra-cluster aggregation loss at the cluster head n is defined as $F_n(\mathbf{w}) = \sum_{u \in \mathcal{U}} \frac{b_{u,n} D_u}{D_n} F_u(\mathbf{w})$.

2.1 CFL process

The architecture of CFL consists of three layers: intra-cluster devices, cluster heads, and a central server. A synchronous aggregation scheme is used. The overall process comprises the following steps, where Step 1 is executed during the initialization phase and the remaining steps are iteratively performed throughout the training process.

1. Clustering: Before the formal training process begins, each device performs pre-training using the global model \mathbf{w} broadcasted by the central server, computes local gradients with its local dataset, and uploads $\nabla F_u(\mathbf{w})$ to the central server. The central server performs data similarity analysis and communication distance calculation based on the gradients and location information uploaded by all devices. These metrics are used to select cluster heads and establish device associations.
2. Local training: The central server broadcasts the global model \mathbf{w} to all devices. The devices receive the model and unfold the local training. In training round t , each device u updates its own model with its local data using the SGD algorithm, as shown in Equation 2:

$$\mathbf{w}_u^t = \mathbf{w}_u^{t-1} - \gamma \mathbf{g}_u(\mathbf{w}_u^{t-1}), \quad (2)$$

where \mathbf{w}_u^t represents the local model parameters obtained during the t th round of local iteration, $\mathbf{g}_u(\mathbf{w}_u^t) = \mathbf{g}(\zeta_u^t; \mathbf{w}_u^t)$ denotes the stochastic gradient computed on a randomly sampled mini-batch ζ_u^t of local training data \mathcal{D}_u , and γ represents the learning rate.

1. Intra – cluster model aggregations : Devices perform intra-cluster model aggregation during local iteration $t \in \{\tau_1, 2\tau_1, 3\tau_1, \dots, T\}$, where T denotes the predefined total number of global training rounds. All devices within the cluster upload their latest models to their respective cluster heads. Each cluster head then aggregates the local models within the cluster according to Equation 3:

$$\mathbf{w}_n^t = \sum_{u \in \mathcal{U}} \frac{b_{u,n} D_u}{D_n} \mathbf{w}_u^t. \quad (3)$$

After intra-cluster model aggregation is completed, each cluster head broadcasts the aggregated model to its associated devices for continued local training.

1. Inter – cluster model aggregations : Devices conduct inter-cluster model aggregation during local iteration $t \in \{\tau_2, 2\tau_2, 3\tau_2, \dots, T\}$, where τ_2 is an integer multiple of τ_1 . Equation 4 is provided as follows:

$$\mathbf{w}^t = \sum_{n \in \mathcal{N}} \frac{a_n D_n}{D} \mathbf{w}_n^t. \quad (4)$$

Subsequently, the central server broadcasts the latest global model to all devices, which is then used for the next round of local training.

2.2 Problem formula

The pairwise cosine similarity between the gradients of devices is computed, as shown in Equation 5:

$$c_{u,u'}(\mathbf{w}) = \frac{\langle \nabla F_u(\mathbf{w}), \nabla F_{u'}(\mathbf{w}) \rangle}{\|\nabla F_u(\mathbf{w})\| \cdot \|\nabla F_{u'}(\mathbf{w})\|}. \quad (5)$$

Assuming that the spectrum is divided into U orthogonal sub-channels and all devices share similar channel conditions (e.g., equal transmission power and bandwidth), the main factor influencing communication energy consumption becomes the distance between devices and their associated cluster heads, along with distance between cluster heads and the central server. Therefore, minimizing the global communication energy consumption can be transformed into minimizing the total communication distance. Let $d_{n,cs}$ denote the distance between the cluster head n and the central server. The total communication distance across all clusters is defined as shown in Equation 6:

$$d(\mathcal{C}) = \sum_{n \in \mathcal{N}} a_n \left(\sum_{u \in \mathcal{U}} b_{u,n} d_{u,n} + d_{n,cs} \right). \quad (6)$$

The objective is to determine the optimal clustering strategy that minimizes the weighted sum of the final global training loss and the overall communication cost. The optimization problem is formulated as shown in Equation 7:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}} \quad & \alpha F(\mathbf{w}^T) + \beta d(\mathcal{C}) \\ \text{s.t.} \quad & \text{C1: } \sum_{n \in \mathcal{N}} b_{u,n} = 1, \quad \forall u \in \mathcal{U}, \\ & \text{C2: } \sum_{u \in \mathcal{U}} b_{u,n} \leq V_n, \quad \forall n \in \mathcal{N}, \\ & \text{C3: } b_{u,n} \leq a_n, \quad \forall u \in \mathcal{U}, \forall n \in \mathcal{N}, \\ & \text{C4: } a_n \in \{0, 1\}, \quad \forall n \in \mathcal{N}, \\ & \text{C5: } b_{u,n} \in \{0, 1\}, \quad \forall u \in \mathcal{U}, \forall n \in \mathcal{N}. \end{aligned} \quad (7)$$

Here, $\mathbf{A} = \{a_n \mid n \in \mathcal{N}\}$, and $\mathbf{B} = \{b_{u,n} \mid u \in \mathcal{U}, n \in \mathcal{N}\}$ denote the cluster head selection and device association strategies, respectively. The weights α and β are used to balance model performance and communication cost, and $\alpha + \beta = 1$. Constraints C1–C5 ensure valid clustering assignments and binary decisions, where V_n imposes a capacity limit on the number of devices that a cluster head can serve due to constrained communication resources. Since $F(\mathbf{w}^T)$ depends implicitly on the decision variables, direct optimization is

intractable. We, therefore, analyze the convergence behavior to derive a tractable formulation.

3 Convergence analysis

To evaluate how clustering strategies influence learning performance, we conduct a convergence analysis to understand how they influence the convergence performance of CFL. To obtain the expected convergence rate of CFL, we first make the following assumptions (Wang et al., 2020; Wan et al., 2021).

- Assumption 1: Assume that the global loss function is differentiable, its gradient is uniformly Lipschitz continuous, and there exists a positive constant L that satisfies Equation 8:

$$\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|. \tag{8}$$

The equation is equivalent to Equation 9:

$$F(\mathbf{w}) \leq F(\mathbf{w}') + (\mathbf{w} - \mathbf{w}')^T \nabla F(\mathbf{w}') + \frac{L}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2. \tag{9}$$

Here, $\|\cdot\|_2$ denotes the calculation of the Euclidean norm.

- Assumption 2: Global divergence is bounded, as shown in Equation 10:

$$\sum_{n \in \mathcal{N}} \frac{a_n D_n}{D} \|\nabla F_n(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq \epsilon^2, \forall n, \mathbf{w}. \tag{10}$$

- Assumption 3: Local divergence is bounded, as presented in Equation 11:

$$\sum_{u \in \mathcal{U}} \frac{b_{u,n} D_u}{D_n} \|\nabla F_u(\mathbf{w}) - \nabla F_n(\mathbf{w})\|^2 \leq \epsilon_n^2, \forall n, \mathbf{w}. \tag{11}$$

- Assumption 4: Local variance is bounded, as shown in Equation 12:

$$\mathbb{E}\|\mathbf{g}_u(\mathbf{w}) - \nabla F_u(\mathbf{w})\|^2 \leq \sigma^2, \forall u, \mathbf{w}. \tag{12}$$

Based on the aforementioned assumptions and the description of the global model, we present the following convergence results. Given the optimal global model \mathbf{w}^* , the learning rate is set to satisfy $\gamma \leq \frac{1}{2\sqrt{6}GL}$. Then, it should satisfy $0 < \frac{4\gamma^2 L^2 \tau_1^2}{1 - 12\gamma^2 L^2 \tau_1^2} \leq \frac{1}{3}$ and $1 < \frac{1}{1 - 12\gamma^2 L^2 \tau_1^2} < \frac{1}{1 - 12\gamma^2 L^2 \tau_2^2} \leq 2$. The optimality gap between the expected global loss function value and the optimal global loss function value is presented in Equation 13:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E}\|\nabla f(\mathbf{w}^t)\|^2 &\leq \frac{2(f(\mathbf{w}^1) - f(\mathbf{w}^*))}{\gamma T} + 36\gamma^2 L^2 \tau_2^2 \epsilon^2 + 3 \\ &\times \sum_{n \in \mathcal{N}} \left(\frac{a_n D_n}{D}\right)^2 \sum_{u \in \mathcal{U}} \left(\frac{b_{u,n} D_u}{D_n}\right)^2 \sigma^2 + 48\gamma^2 L^2 \tau_1^2 \sum_{n \in \mathcal{N}} \frac{a_n D_n}{D} \epsilon_n^2 \\ &+ 24\gamma^2 L^2 \tau_2 \sum_{n \in \mathcal{N}} a_n \left(\frac{D_n}{D} \left(1 - \frac{D_n}{D}\right)\right) \sum_{u \in \mathcal{U}} \left(\frac{b_{u,n} D_u}{D_n}\right)^2 \sigma^2 \\ &+ 16\gamma^2 L^2 \tau_1 \sum_{n \in \mathcal{N}} \frac{a_n D_n}{D} \sum_{u \in \mathcal{U}} \left(\frac{b_{u,n} D_u}{D_n} \left(1 - \frac{b_{u,n} D_u}{D_n}\right)\right)^2 \sigma^2. \end{aligned} \tag{13}$$

From Equation 13, we can analyze how each key parameter would affect the convergence of the proposed CFL algorithm. The learning rate γ determines the gradient step size, while the smoothness constant L characterizes the curvature of the loss function; increasing either parameter would enlarge the convergence upper bound, whereas decreasing them would tighten it, thereby improving convergence. Participation weights a_n and $b_{u,n}$, along with data sizes D_n and D_u , determine each device's contribution to the aggregated gradient; uneven participation or disproportionate data sizes would further increase the bound. The total number of communication rounds T reduces the contribution of the initial optimality gap, so more rounds would tighten the bound. The parameter τ_1 specifies how many times each device updates its model before a single intra-cluster aggregation; a larger τ_1 would amplify the effects of ϵ_n^2 and σ^2 , thus increasing the bound. Similarly, the parameter τ_2 specifies how many times each device updates its model before a single inter-cluster aggregation; a larger τ_2 value would amplify the effects of ϵ^2 and σ^2 , thus also increasing the bound.

Notably, by combining the convergence bound in Equation 13 with assumptions 2 and 3, it can be observed that client drift—defined as the deviation of local gradient updates from the global gradient direction—increases the convergence upper bound, thereby degrading learning performance. In particular, assumptions 2 and 3 characterize such drift through the deviation measures ϵ and ϵ_n , which quantify the discrepancy between cluster-level gradients and the global gradient, along with between-individual client gradients and their corresponding cluster-level gradients. Since the derived convergence upper bound increases monotonically with both ϵ and ϵ_n , stronger client drift results in a looser convergence bound and, consequently, slower convergence. We further observe that reducing ϵ^2 , which measures the discrepancy between the aggregated data distribution within each cluster and the global data distribution, yields a more significant improvement in learning performance than reducing ϵ_n^2 . This is because the period of cluster-level updates, τ_2 , is typically larger than the period of device-level updates, τ_1 , causing the effect of intra-cluster heterogeneity to accumulate more strongly over τ_2 steps. This indicates that enhancing intra-cluster data representativeness plays a dominant role in mitigating client drift and motivates our focus on cluster-level optimization. Consequently, we focus on minimizing the following objective $\min \sum_{n \in \mathcal{N}} \frac{a_n D_n}{D} \|\nabla F_n(\mathbf{w}) - \nabla F(\mathbf{w})\|^2$. For analytical convenience, we assume uniform data sizes across devices and simple averaging during global aggregation. Under these assumptions, the above divergence objective can be reformulated using norm properties and further bounded via the Cauchy-Schwarz inequality. This results in an upper bound that relates the discrepancy between local and global gradients to the pairwise similarity among local gradients. As a result, minimizing the sum of intra-cluster gradient cosine similarity serves as a tractable surrogate for reducing client drift induced by data heterogeneity. Accordingly, based on the definition of cosine similarity in Equation 2, the objective is transformed into the maximization problem shown in Equation 14:

$$\min \frac{1}{N_{ch}} \sum_{n \in \mathcal{N}_{ch}} \frac{1}{C_n^2} \sum_{u, u' \in \mathcal{C}_n} c_{u, u'}(\mathbf{w}). \tag{14}$$

4 Algorithm design

Based on the above convergence analysis, the cluster-level data representativeness metric is defined as follows (Equation 15):

$$c(\mathcal{C}) = \frac{1}{N_{ch}} \sum_{n \in \mathcal{N}_{ch}} \frac{1}{C_n^2} \sum_{u, u' \in \mathcal{C}_n} c_{u, u'}(\mathbf{w}). \quad (15)$$

By substituting the original global loss function with $c(\mathcal{C})$, the optimization problem in Equation 4 can be reformulated as presented in Equation 16:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}} \quad & \beta d(\mathcal{C}) + \alpha c(\mathcal{C}) \\ \text{s.t.} \quad & \text{C1-C5,} \end{aligned} \quad (16)$$

where Equation 14 is NP-hard and is addressed by decomposing it into device association and cluster-head selection strategies based on cosine similarity and communication distance.

Given an initial cluster head set $\mathcal{N}_{ch} \subseteq \mathcal{N}$, the device association problem is formulated as shown in Equation 17:

$$\begin{aligned} \min_{\mathbf{B}} \quad & \beta \sum_{n \in \mathcal{N}_{ch}} \sum_{u \in \mathcal{U}} b_{u, n} d_{u, n} + \alpha \frac{1}{N_{ch}} \sum_{n \in \mathcal{N}_{ch}} \frac{1}{C_n^2} \sum_{u, u' \in \mathcal{C}_n} c_{u, u'}(\mathbf{w}) \\ \text{s.t.} \quad & \text{C1, C2, C5,} \end{aligned} \quad (17)$$

which is an integer nonlinear programming problem. We adopt the Gurobi solver for optimal device association under a fixed cluster head set. The overall clustering utility is defined in Equation 18:

$$\Psi(\mathcal{N}) = \psi(\mathcal{N}) + \sum_{n \in \mathcal{N}} a_n d_{n, cs}, \quad (18)$$

where $\psi(\mathcal{N})$ is the device association utility obtained by solving Equation 17 with the given cluster head set. The cluster head selection problem can be formulated as shown in Equation 19:

$$\begin{aligned} \min_{\mathbf{A}} \quad & \Psi(\mathcal{N}) \\ \text{s.t.} \quad & \text{C4.} \end{aligned} \quad (19)$$

To minimize $\Psi(\mathcal{N}_{ch})$, a greedy iterative strategy is used to adjust the cluster head set, which includes the following three main update operations:

- **Cluster head addition:** For any device $n \in \mathcal{N} \setminus \mathcal{N}_{ch}$, compute the utility $\Psi(\mathcal{N}_{ch} + \{n\})$; if it improves the objective, add n to \mathcal{N}_{ch} .
- **Cluster head exchange:** For $n \in \mathcal{N} \setminus \mathcal{N}_{ch}$ and $n' \in \mathcal{N}_{ch}$, evaluate $\Psi(\mathcal{N}_{ch} + \{n\} - \{n'\})$; replace n' with n if utility is improved.
- **Cluster head removal.** For $n \in \mathcal{N}_{ch}$, compute $\Psi(\mathcal{N}_{ch} - \{n\})$; if this reduces the objective, remove n from the set.

Therefore, in the greedy iterative strategy for cluster head selection, the number of cluster heads varies dynamically until convergence. We summarize the alternating optimization process of device association and cluster head selection in Algorithm 1. The overall complexity is mainly determined by two components: (i) solving the device association problem using the Gurobi solver for a fixed cluster head set and (ii) the greedy iterative updates for cluster head selection, including addition, exchange, and removal operations. In particular, let $N = |\mathcal{N}|$ be the total number of candidate nodes and $U = |\mathcal{U}|$ be the number of devices. Computing the device association utility $\psi(\mathcal{N}_{ch})$ via Gurobi involves all devices in \mathcal{U} , with complexity $\mathcal{O}(U)$ per evaluation.

During each iteration of the greedy cluster head updates, the addition step requires up to $N - |\mathcal{N}_{ch}|$ evaluations of $\Psi(\mathcal{N}_{ch})$, the exchange step requires up to $(N - |\mathcal{N}_{ch}|)|\mathcal{N}_{ch}|$ evaluations, and the removal step requires up to $|\mathcal{N}_{ch}|$ evaluations. Hence, the per-iteration complexity can be expressed as $\mathcal{O}((N - |\mathcal{N}_{ch}|) + (N - |\mathcal{N}_{ch}|)|\mathcal{N}_{ch}| + |\mathcal{N}_{ch}|)U$. Since $|\mathcal{N}_{ch}| \leq N$, this can be simplified and upper-bounded by $\mathcal{O}(N^2U)$ per iteration. Therefore, for a total of T iterations until convergence, the overall computational complexity of the alternating optimization algorithm is $\mathcal{O}(TN^2U)$, which clearly highlights its scalability with respect to the number of candidate nodes and the number of devices.

Input: Initial cluster head set \mathcal{N}_{ch} , candidate cluster head set \mathcal{N} , and device set \mathcal{U} .

Output: Optimal clustering set and device association strategy.

1: Compute the cosine similarity between devices, communication distances, and server distances.

2: Solve Equation 17 using the Gurobi optimizer to obtain $\psi(\mathcal{N}_{ch})$.

3: Compute $\Psi(\mathcal{N}_{ch})$.

4: Repeat the following operations until the objective value of Equation 19 converges.

5: **for** $n \in \mathcal{N} \setminus \mathcal{N}_{ch}$ **do**

6: Compute $\Psi(\mathcal{N}_{ch} + \{n\})$

7: **if** $\Psi(\mathcal{N}_{ch} + \{n\}) < \Psi(\mathcal{N}_{ch})$ **then**

8: $\mathcal{N}_{ch} \leftarrow \mathcal{N}_{ch} + \{n\}$

9: **end if**

10: **end for**

11: **for** $n \in \mathcal{N} \setminus \mathcal{N}_{ch}$, $n' \in \mathcal{N}_{ch}$ **do**

12: Compute $\Psi(\mathcal{N}_{ch} + \{n\} - \{n'\})$

13: **if** $\Psi(\mathcal{N}_{ch} + \{n\} - \{n'\}) < \Psi(\mathcal{N}_{ch})$ **then**

14: $\mathcal{N}_{ch} \leftarrow \mathcal{N}_{ch} + \{n\} - \{n'\}$

15: **end if**

16: **end for**

17: **for** $n \in \mathcal{N}_{ch}$ **do**

18: Compute $\Psi(\mathcal{N}_{ch} - \{n\})$

19: **if** $\Psi(\mathcal{N}_{ch} - \{n\}) < \Psi(\mathcal{N}_{ch})$ **then**

20: $\mathcal{N}_{ch} \leftarrow \mathcal{N}_{ch} - \{n\}$

21: **end if**

22: **end for**

Algorithm 1. Clustering strategy based on device association and cluster head selection.

5 Numerical results

We simulate a wireless FL system consisting of a central server and 100 devices uniformly distributed within a circular area of 100 m radius. To model long-range communication, the server is positioned 1 km away from the device cluster center. The total number of global training rounds is fixed at 200, and the intra- and inter-cluster model aggregation periods are set to $\tau_1 = 1$ and $\tau_2 = 2$, respectively. Each device is allocated to a communication bandwidth of 1 Mbps and transmits at a power of 10 mW. The noise power spectral density is set to -173 dBm/Hz, and the path loss model

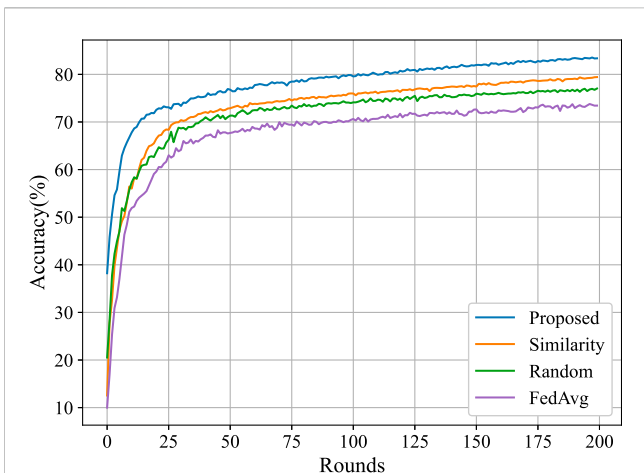


FIGURE 2 Test accuracy on Fashion-MNIST vs. local training rounds.

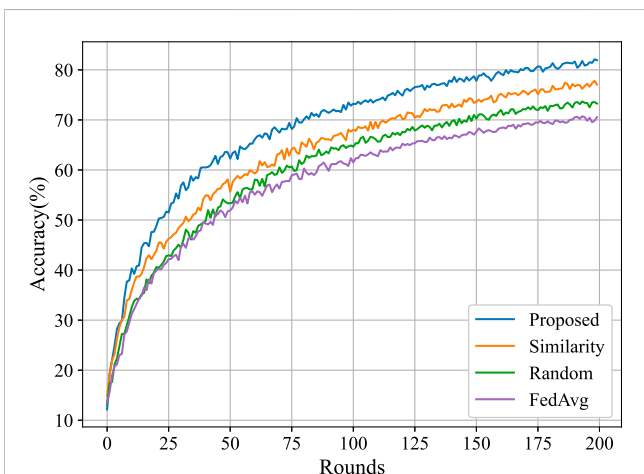


FIGURE 3 Test accuracy on CIFAR-10 vs. local training rounds.

TABLE 1 Test accuracy of different algorithms on Fashion-MNIST and CIFAR-10 datasets.

Algorithm	Fashion-MNIST	CIFAR-10
Proposed	83.4%	82.1%
Similarity	79.8%	78.3%
Random	76.7%	73.8%
FedAvg	73.6%	70.2%

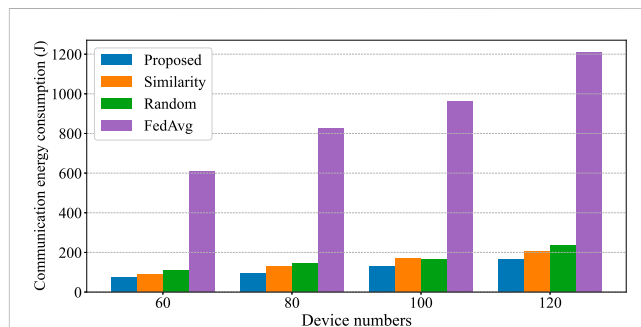


FIGURE 4 Global energy consumption vs. number of devices.

follows $128.1 + 37.6 \log_{10}(d_{km})$. The model size is 28 kB, and each local training batch consists of 32 samples, with a learning rate of 0.01. Communication energy consumption is calculated using the Shannon formula, considering both device-to-cluster-head and cluster-head-to-server distances (Chen et al., 2020; Taik et al., 2022). For simplicity, we assume identical transmission power and bandwidth across all devices so that both communication energy and delay depend solely on transmission distance and exhibit the same trend. In addition, we consider the single-hop communication model between devices and their associated cluster heads, which is commonly adopted in cluster-based wireless FL architectures. Under this assumption, routing overhead is negligible and does not affect the relative performance comparison. Experiments are conducted on the Fashion-MNIST and CIFAR-10 datasets. To emulate a statistically heterogeneous non-IID environment, data are partitioned across devices using a Dirichlet distribution, with a concentration parameter set to 0.5, which

corresponds to a moderate-to-high level of data heterogeneity and has been widely adopted in prior federated learning studies to evaluate robustness under non-IID settings (Meng et al., 2023). We set the weights to $\alpha = \beta = 0.5$.

To evaluate the proposed method, three baseline algorithms are considered for comparison: (i) *random clustering*, where devices are grouped based on the geographical proximity and cluster heads are selected randomly within each cluster; (ii) *similarity-based clustering*, which groups devices with similar local data distributions using statistical distance metrics, with heads randomly assigned; and (iii) *FedAvg*, the conventional FL scheme without clustering, where all devices communicate directly with the central server in each round.

To verify the effectiveness of the proposed clustering algorithm in enhancing learning performance in FL, we conduct test accuracy comparison experiments on the Fashion-MNIST and CIFAR-10 datasets. The experiments evaluate the impact of different device clustering algorithms, along with the classical FedAvg, on the model’s training accuracy. Figures 2, 3 illustrate the evolution of test accuracy during training, and the corresponding test accuracies at convergence are summarized in Table 1. As shown, the proposed algorithms consistently achieve the highest test accuracies for a given number of training rounds and maintain significant advantages throughout the training process. The similarity-based clustering algorithm ranks second, suggesting that adjusting data within clusters to a more balanced distribution—i.e., aligning the data distribution across clusters with the global distribution—can lead to better convergence performance than clustering purely based on intra-cluster data similarity, which is in line with expectations. The randomized clustering algorithm ranks third because it does not account for

TABLE 2 Impact of the weighting factor α on learning performance and communication energy.

α	Learning performance	Communication energy
0	70.4%	101J
0.3	79.5%	117J
0.5	83.4%	132J
0.7	84.9%	154J
1	85.1%	168J

data distribution or similarity within clusters, leading to less balanced clusters and consequently slower convergence. The FedAvg algorithm exhibits the worst performance, primarily due to the heterogeneity of local data distributions across devices under $\alpha = 0.5$, which leads to inconsistent update directions in local models and partially counteracts gradients during global model aggregation, thereby reducing convergence efficiency. Overall, these results further highlight the effectiveness of the proposed clustering algorithm in improving the training of devices in heterogeneous environments.

To verify the effectiveness of the proposed algorithm in reducing global communication energy consumption, energy simulations are conducted on the Fashion-MNIST dataset under different numbers of devices, as shown in Figure 4. FedAvg incurs the highest energy consumption due to frequent long-distance communication with the server. In contrast, the other three algorithms use intra-cluster aggregation, which shortens communication distances and reduces upload frequency, thereby lowering energy usage. The proposed method performs best by jointly optimizing inter-device distances and balancing intra-cluster data, further reducing global communication energy.

As shown in Figures 2, 4, the proposed algorithm consistently achieves high training accuracy with low communication energy consumption, highlighting its advantage in maintaining model performance while reducing communication cost. In comparison, similarity-based and random clustering exhibit slower convergence and higher energy consumption. The results demonstrate that balancing intra-cluster data enhances cluster representativeness, mitigates aggregation conflicts, and, when combined with geographical proximity, contributes to reducing overall energy consumption.

Table 2 presents the learning performance and communication energy under different values of the weighting factor α . As α increases, giving more importance to model accuracy in the optimization problem, the learning performance improves, reaching up to 85.1% for $\alpha = 1$. However, this comes at the cost of higher communication energy, which increases from 101 J to 168 J over the same range. These results clearly illustrate the trade-off controlled by the weighting factor: larger α prioritizes learning accuracy at the expense of communication efficiency, while smaller α reduces energy consumption but yields lower model performance. This demonstrates that α serves as an effective tuning parameter to balance learning quality and communication cost in the proposed CFL framework.

6 Conclusion

This study investigates the trade-off between learning performance and communication energy consumption in CFL, focusing on how cluster head selection and device association affect model training and energy overhead. To collectively optimize model performance and communication efficiency in wireless FL, we first conduct a convergence analysis linking global loss to inter-cluster data imbalance and use cosine similarity to quantify distributional dissimilarity. An optimization model of training loss is then constructed based on cosine gradient similarity, while communication energy is modeled as a function of transmission distance. Finally, a clustering algorithm is proposed to jointly schedule cosine similarity and communication distance for solving the reformulated combined optimization problem. The simulation results show that the proposed method markedly reduces communication energy while improving model accuracy.

In future work, we aim to introduce data-size-aware weighting mechanisms to further optimize client selection and matching, along with adaptive channel allocation strategies to extend the applicability of the method to heterogeneous devices and non-uniform wireless environments. These directions aim to improve both the scalability and robustness of the CFL framework, providing a more comprehensive solution for real-world federated learning scenarios.

Data availability statement

The datasets generated during this study are not publicly available due to the following restrictions: 1. the simulation data and model parameters are integral to the proprietary research methodology developed in this work. 2. The training data consist of standard benchmark datasets that are already publicly available through their original sources. 3. Raw gradient information and device-specific data cannot be shared as they may contain sensitive information about the federated learning system architecture. Requests to access the datasets should be directed to Zhenning Chen, link_chen@yeah.net.

Author contributions

ZC: Writing – original draft. ZX: Writing – review and editing. YD: Data curation, Investigation, Validation, Writing – review and editing. YW: Methodology, Supervision, Validation, Writing – review and editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial

intelligence, and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Briggs, C., Fan, Z., and Andras, P. (2020). "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 international joint conference on neural networks (IJCNN)*, 1–9. doi:10.1109/IJCNN48605.2020.9207469
- Chen, M., Yang, Z., Saad, W., Yin, C., Poor, H. V., and Cui, S. (2020). A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions Wireless Communications* 20, 269–283. doi:10.1109/twc.2020.3024629
- Gao, Z., Xiong, Z., Zhao, C., and Feng, F. (2023). "Clustered federated learning framework with acceleration based on data similarity," in *International conference on algorithms and architectures for parallel processing*, 80–92.
- Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. (2021). An efficient framework for clustered federated learning. doi:10.48550/arXiv.2006.04088
- Konečný, J., McMahan, B., and Ramage, D. (2015). Federated optimization: distributed optimization beyond the datacenter. *arXiv Preprint arXiv:1511.03575*. doi:10.48550/arXiv.1511.03575
- Lu, Z., Pan, H., Dai, Y., Si, X., and Zhang, Y. (2024). Federated learning with non-iid data: a survey. *IEEE Internet Things J.* 11, 19188–19209. doi:10.1109/JIOT.2024.3376548
- Meng, X., Li, Y., Lu, J., and Ren, X. (2023). An optimization method for non-iid federated learning based on deep reinforcement learning. *Sensors* 23, 9226. doi:10.3390/s23229226
- Nishio, T., and Yonetani, R. (2019). "Client selection for federated learning with heterogeneous resources in Mobile edge," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 1–7. doi:10.1109/ICC.2019.8761315
- Taik, A., Mlika, Z., and Cherkaoui, S. (2022). Clustered vehicular federated learning: process and optimization. *IEEE Trans. Intelligent Transp. Syst.* 23, 25371–25383. doi:10.1109/TITS.2022.3149860
- Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., et al. (2022). Fedproto: Federated prototype learning across heterogeneous clients. *Proc. AAAI Conference Artificial Intelligence* 36, 8432–8440. doi:10.1609/aaai.v36i8.20819
- Tran, D. T., Ha, N. B., Nguyen, V.-D., and Wong, K.-S. (2025). Sherl-fl: when representation learning meets split learning in hierarchical federated learning. *arXiv Preprint arXiv:2508.08339*. doi:10.48550/arXiv.2508.08339
- Wan, S., Lu, J., Fan, P., Shao, Y., Peng, C., and Letaief, K. B. (2021). Convergence analysis and system design for federated learning over wireless networks. *IEEE J. Sel. Areas Commun.* 39, 3622–3639. doi:10.1109/jsac.2021.3118351
- Wang, J., Wang, S., Chen, R.-R., and Ji, M. (2020). Local averaging helps: hierarchical federated learning and convergence analysis. *arXiv Preprint arXiv:2010*. doi:10.48550/arXiv.2010.12998
- Xia, W., Quek, T. Q. S., Guo, K., Wen, W., Yang, H. H., and Zhu, H. (2020). Multi-armed bandit-based client scheduling for federated learning. *IEEE Trans. Wirel. Commun.* 19, 7108–7123. doi:10.1109/TWC.2020.3008091
- Yan, Y., Tong, X., and Wang, S. (2024). Clustered federated learning in heterogeneous environment. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 12796–12809. doi:10.1109/TNNLS.2023.3264740
- Yang, M., Xu, J., Ding, W., and Liu, Y. (2024). Fedhap: federated hashing with global prototypes for cross-silo retrieval. *IEEE Trans. Parallel Distributed Syst.* 35, 592–603. doi:10.1109/TPDS.2023.3324426
- Zeng, D., Hu, X., Liu, S., Yu, Y., Wang, Q., and Xu, Z. (2023). Stochastic clustered federated learning. doi:10.48550/arXiv.2303.00897
- Zhang, Y., Chen, H., Lin, Z., Chen, Z., and Zhao, J. (2024). Fedac: an adaptive clustered federated learning framework for heterogeneous data. doi:10.48550/arXiv.2403.16460