Information Fusion 135 (2026) 104482

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

SCMoE-PFL: A soft-clustering mixture-of-experts framework for personalized federated learning

Gongli Li[a,b], Xianzhong Jia[a], Weichen Liu[a], En Zhang[a,b], Zidong Wang[c,*]

[a] College of Computer and Information Engineering, Henan Normal University, Xinxiang, 453007, Henan, China
[b] Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province, Xinxiang, 453007, Henan, China
[c] Department of Computer Science, Brunel University of London, Uxbridge, UB8 3PH, Middlesex, UK

ARTICLE INFO

Keywords:
Personalized federated learning
Soft clustering
Mixture of experts
Gating network
Data heterogeneity
Privacy preservation

ABSTRACT

Traditional federated learning (FL) methods rely on a single global model, which often collapses under heterogeneous and non-IID client data distributions. Personalized federated learning (PFL) alleviates this limitation, yet existing approaches either overfit to local data or fail to exploit shared knowledge effectively. To address these challenges, this paper presents SCMoE-PFL, a personalized federated learning framework that integrates soft clustering and a mixture-of-experts (MoE) mechanism to reconcile global generalization with local personalization. First, we introduce a multi-center threshold-based soft clustering (MCTC) method that enables clients to participate in multiple clusters, improving data utilization and cluster quality. Second, intra-cluster aggregation yields a set of expert models, while each client separately trains a private model on its high-sensitivity data, ensuring privacy preservation. Finally, a lightweight energy-aware gating network adaptively fuses expert and private models. By calibrating initial feature-matching weights with energy-based predictive confidence, this dual-check mechanism effectively prevents over-reliance on uncertain experts, thereby producing highly reliable personalized predictions. Experiments on four benchmark datasets demonstrate that SCMoE-PFL substantially improves accuracy, convergence, and fairness under both moderate and extreme heterogeneity, achieving maximum accuracy improvements of 24.71 and 26.01 percentage points over FedAvg, respectively. Theoretical analysis further establishes performance lower bounds and clarifies the framework's advantages in privacy protection, computational efficiency, and system reliability. These results show that SCMoE-PFL offers a robust and flexible solution for personalized federated learning in heterogeneous environments.

1. Introduction

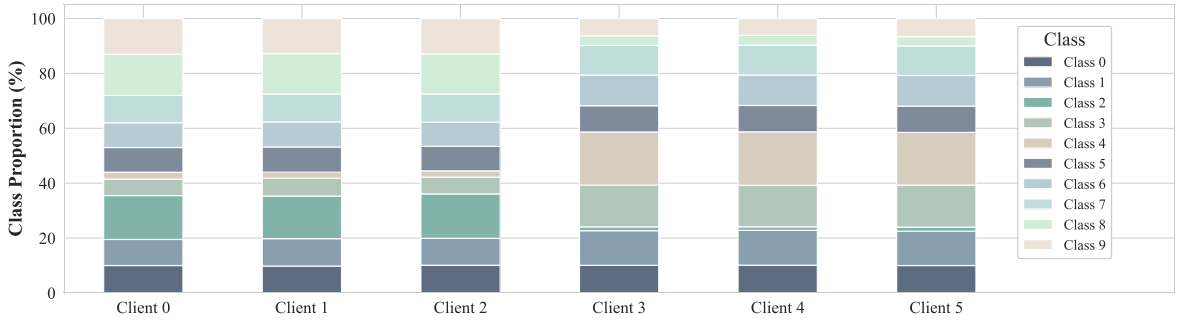
Federated Learning (FL) has been recognized as an important machine learning paradigm that attracts considerable attention across domains such as healthcare, finance, and the Internet of Things (IoT) [1–4]. This paradigm enables distributed devices or clients to collaboratively train a shared global model while retaining data locally, thereby enhancing privacy protection. By allowing collaborative learning without transmitting raw data to a central server, FL addresses long-standing concerns related to data privacy, security, and the difficulties induced by data silos.

In real-world applications, the deployment of federated learning continues to face significant challenges. Distributed devices generate large volumes of data that often exhibit pronounced non-independent and non-identically distributed (non-IID) characteristics, which may cause the global model to perform poorly on certain clients and yield results far below those obtained from locally trained models [5–7]. These observations indicate that a single global model is frequently inadequate for heterogeneous clients. As a result, increasing attention has been directed toward developing personalized models that better reflect individual client characteristics, forming the field known as personalized federated learning (PFL) [8,9]. For example, Clustered Distributed Co-Meta-Learning Personalized FL (CDC-PFL) has been proposed by Ren et al. in [10], where clustering strategies and meta-learning concepts have been combined to group similar clients and train a shared personalized model for each cluster, thereby improving performance when dealing with heterogeneous data and limited sample sizes.

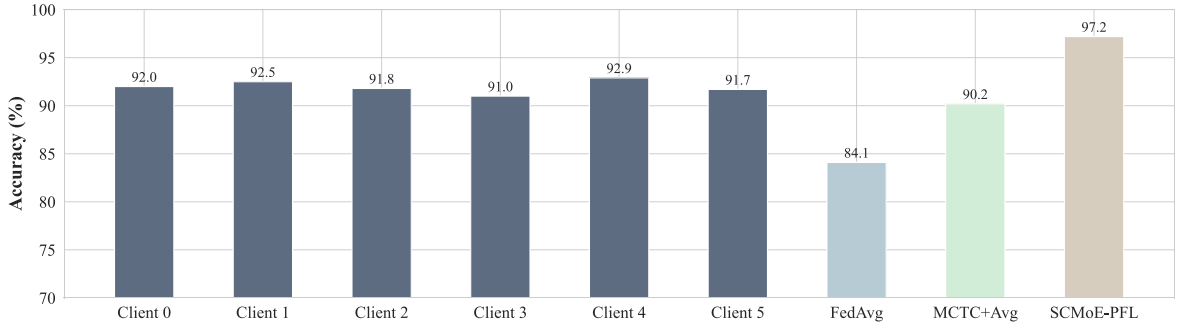
A widely adopted approach for implementing personalized federated learning has been clustering, whose central idea is to partition clients into multiple clusters based on similarities in their data distributions and then train a personalized model for each cluster [11]. By promoting collaboration among clients with similar characteristics, the effects of non-IID data on model performance can be mitigated. However, several limitations remain. In complex real-world environments, the accuracy

* Corresponding author.
E-mail address: zidong.wang@brunel.ac.uk (Z. Wang).

https://doi.org/10.1016/j.inffus.2026.104482
Received 26 December 2025; Received in revised form 30 April 2026; Accepted 13 May 2026
Available online 15 May 2026



(a) Client 0-5 - Label Distribution



(b) Accuracy Comparison: Clients and Methods

Fig. 1. Comparison of client data category distribution and model performance.

of clustering is often insufficient; moreover, the "hard assignment" strategy, in which each client is assigned to a single cluster, restricts data utilization and prevents the exploration of potential cross-cluster model fusion.

Optimizing model aggregation strategies is another key approach for achieving personalized federated learning. Methods such as FedSmooth [12], FedVa [13], and AMA [14] improve aggregation weights by incorporating information such as model overfitting, prediction accuracy, and historical model states. Although these approaches yield moderate gains in accuracy and stability, they often introduce high computational costs or rely on specific prior knowledge, which limits their applicability. More fundamentally, most of these strategies still follow a "single model return" paradigm, in which all clients receive the same aggregated model. Such an approach does not accommodate personalized needs, overlooks the benefits of fusing multiple cluster-level models, and restricts both aggregation flexibility and model expressiveness.

In practical applications of FL, data distributions among clients are often heterogeneous, with substantial imbalances in both category proportions and sample sizes. This heterogeneity raises an important question: in scenarios where category and sample distributions are highly imbalanced, can traditional clustering- and aggregation-based personalized federated learning methods effectively integrate knowledge across diverse clients and clusters? Existing methods have not fully resolved this issue. To investigate this challenge and provide a solution, a novel personalized federated learning method is introduced that incorporates a soft clustering mechanism and a mixture-of-experts (MoE) approach. Clustering quality and data utilization are enhanced by allowing clients to belong to multiple overlapping clusters. Expert models are obtained through intra-cluster aggregation and returned to the clients in the corresponding clusters. For each client, these expert models, along with its private model, constitute its candidate expert set. To robustly fuse these models under extreme data imbalance, an energy-aware gating network is trained for prediction. By calibrating initial feature-matching weights with energy-based predictive confidence, this network effectively mitigates the risk of over-relying on uncertain experts, ensuring highly reliable personalization.

To evaluate the effectiveness of the proposed SCMoE-PFL, a comparative experiment has been conducted using FedAvg [15], with results shown in Fig. 1. Fig. 1(a) illustrates the data distribution across six clients (client 0–5), covering all categories (class 0–9), with different proportions of categories. Clients 0–2 exhibit relatively balanced distributions, whereas clients 3–5 show significant sample-size imbalance, representing typical heterogeneity in federated learning environments. Fig. 1(b) compares the accuracy of each client's locally trained model with the average performance of three federated learning methods. FedAvg employs average aggregation to form a global model; MCTC + Avg applies the MCTC soft clustering scheme followed by intra-cluster average aggregation to obtain cluster-level models, which are then averaged; SCMoE-PFL represents the proposed soft-clustering mixture-of-experts method. The results demonstrate that the MCTC-based soft clustering scheme significantly improves average client accuracy over traditional global aggregation. When the mixture-of-experts mechanism is added, SCMoE-PFL achieves the best performance, confirming that the proposed method strengthens both global generalization and local personalization despite substantial data heterogeneity.

In summary, the main contributions of this paper are as follows.

1. The SCMoE-PFL framework is proposed, comprising MCTC soft clustering, intra-cluster aggregation, private model training, and an energy-aware gating mechanism. It is designed to address data heterogeneity, model uncertainty, and limited personalization under strict privacy requirements.
2. An MCTC soft clustering scheme is designed to improve data utilization and alleviate ambiguous boundaries in heterogeneous scenarios. The scheme supports multi-cluster assignment through adjustable thresholds, while a clustering verification mechanism is used to maintain clustering quality.

3. A personalized MoE module featuring a novel energy-aware gating network is constructed for each client. By dynamically calibrating feature-matching priors with energy-based predictive confidence, this dual-check mechanism effectively mitigates the risk of over-relying on uncertain experts under extreme data imbalance, enabling highly reliable personalized predictions while preserving generalization. Furthermore, since private models are not uploaded to the server, the risk of privacy leakage is reduced.
4. Experiments on multiple benchmark datasets demonstrate the fairness and accuracy of SCMoE-PFL, while ablation studies confirm the importance of its components. Theoretical analysis further establishes its performance lower bounds and its advantages in privacy protection and efficiency.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the design of SCMoE-PFL. Section 4 presents the experimental results. Section 5 concludes this paper.

2. Related work

2.1. Federated learning

Federated learning has gained significant attention in recent years and has been successfully applied across various domains, including medical image analysis, industrial engineering, and edge computing [16–20]. Extensive research efforts have been devoted to exploring different facets of this paradigm [21]. For instance, FedRFC has been proposed by Deng et al. [22] as a federated learning framework based on recursive fuzzy clustering, and training performance has been enhanced by over 10% on highly non-IID data through iterative partitioning of clients into overlapping clusters. Security and privacy risks associated with federated learning have been systematically reviewed by Hu et al. [23], and corresponding defense mechanisms have been proposed. A reliable anomaly detection strategy for the Industrial Internet of Things (IIoT) has been proposed by Wang et al. [24], utilizing federated and deep reinforcement learning to mitigate privacy leakage. Furthermore, to defend against model poisoning attacks, FedHAN has been introduced by Wang et al. [25], which dynamically detects update deviations and recovers damaged models to ensure system reliability under asynchronous communication and device heterogeneity. In terms of applications, PDP-PFL has been proposed by Qu et al. [26], in which noise has been injected according to users' privacy preferences and local fine-tuning of the final layer has been performed to mitigate privacy leakage in vehicular networks. IFGNN has been proposed by Tang et al. [27], where a federated graph neural network has been employed to enable collaborative learning across supply chain nodes while preserving data privacy.

Despite its wide application potential, FL faces a fundamental challenge in practical deployment: statistical heterogeneity, which arises from the non-IID characteristics commonly observed in client data [28,29]. This heterogeneity has been reflected in four primary aspects.

1) Feature distribution skew: shifts in the feature distributions of samples from the same category across different clients.
2) Label distribution skew: significant differences in the proportion of samples from various categories across clients.
3) Quantity imbalance: discrepancies in the total number of samples available to different clients.
4) Quality heterogeneity: variations in data quality (e.g., noise and labeling accuracy) between clients.

Addressing statistical heterogeneity is essential for improving federated learning performance, and considerable research has been motivated to focus on personalized federated learning, where individualized models are generated for each client. FedProx has been proposed by Li et al. [30] to mitigate the impact of non-IID data by introducing a proximal term into FedAvg, resulting in an 18.8% improvement in average test accuracy compared to FedAvg. FedAS has been introduced by Yang et al. [31], where parameter alignment and client synchronization mechanisms have been employed to enhance the adaptability of shared parameters. The negative effects of "straggler clients" have also been reduced through a robust aggregation strategy, and improved performance and robustness have been demonstrated under data heterogeneity. The FairDPFL-SCS framework has been proposed by Sabah et al. [32], where dynamic adjustments of the learning rate and client selection mechanism have been employed. This framework has enhanced model personalization and efficiency while effectively addressing data heterogeneity, achieving an accuracy improvement of up to 16.74% on the SVHN dataset. Despite these advancements, effectively balancing the trade-off between handling statistical heterogeneity and ensuring data privacy remains a critical challenge.

2.2. Clustering methods in PFL

The non-IID data problem, commonly encountered in practical applications, has presented a significant challenge for federated learning, as traditional methods have often been found inadequate [33–35]. Consequently, PFL has received increasing attention. In response to this issue, various personalized strategies have been proposed, with clustering methods having emerged as a prominent approach, whose fundamental idea is to partition clients into multiple clusters based on the similarity of their data distributions or model update characteristics, ensuring that the data environment within each cluster is as homogeneous as possible, thereby reducing the training bias induced by heterogeneity. The information sources utilized for client clustering generally fall into three categories: raw data distributions, model parameters, and gradient updates. Although clustering based on raw data distributions intuitively provides the most direct measure of client heterogeneity, this approach fundamentally conflicts with the privacy-preserving principles of federated learning, as sensitive local data statistics cannot be exposed to the central server. On the other hand, model parameters offer a stable and comprehensive snapshot of the local learning state; however, due to the continuous accumulation of historical updates across communication rounds, the parameters themselves exhibit significant "inertia". In contrast, gradients capture the instantaneous local optimization direction and can more sensitively reflect how a client model responds to its specific data distribution in the current state. Gradient-based clustering thus provides a scheme that satisfies privacy requirements while maintaining high sensitivity.

To enhance the robustness and adaptability of clustering, the StoCFL framework has been proposed by Zeng et al. [36], supporting dynamic client participation and the onboarding of new clients. Low data utilization efficiency has been mitigated through a cross-cluster information-sharing mechanism, and model performance has been significantly improved under various non-IID scenarios. To address data heterogeneity and bias, IFCEA has been proposed by Wei et al. [37], introducing an endpoint evaluation mechanism, a clustering filtering strategy, data distribution-aware aggregation weights, and the adaptive initialization method OneBiPartition. This approach has substantially improved data utilization efficiency and model performance. CFL-ICCV has been introduced by Liu et al. [38], employing secret sharing to enable privacy-preserving distributed client clustering. An intra-cluster cross-validation mechanism has been designed, in which clients not participating in training validate local updates to identify malicious models, thereby addressing data heterogeneity and byzantine attacks in distributed energy forecasting. Du et al. [39] have proposed AICFL, a dynamic adaptive clustering federated learning method that performs clustering division and adjustment operations during early iterations to address client heterogeneity and changes in data distribution. This method has required neither a predefined number of clusters nor full client availability, demonstrating strong real-time performance, adaptability, and

flexibility. FCFLA has been introduced by Yoo et al. [40], achieving multi-cluster membership through a fuzzy membership function. This method has utilized data features more effectively, alleviated data scarcity issues, improved prediction accuracy, and accelerated model convergence.

Clustering strategies have demonstrated notable potential in mitigating data heterogeneity, improving model stability, and accelerating convergence. Recent studies further emphasize overcoming the limitations of rigid clustering architectures. For instance, SnapCFL has been proposed by Cheng et al. [41] to decouple the clustering process from the main federated training stage, thereby improving framework flexibility and alleviating complex data heterogeneities. Moreover, to address the privacy vulnerabilities and robustness issues inherent in standard gradient-based clustering, ProCFL has been introduced by Xu et al. [42]. By utilizing gradient-free similarity measurements and formulating the client clustering process as a weighted set-covering problem, ProCFL achieves diversity-optimized clustering, further highlighting the critical trend toward flexible and overlapping cluster architectures. However, despite these recent advances, mainstream existing methods continue to face practical challenges, including low local data utilization, high computational overhead, and limited flexibility stemming from the strict single-cluster constraint (each client being forcibly assigned to only a single cluster). To address these limitations, this paper introduces the multi-center threshold-based clustering (MCTC) scheme. MCTC offers several key advantages: clients can belong to multiple clusters simultaneously, overcoming the traditional single-cluster constraint; initial cluster centers are constructed by selecting the "client pair" with the lowest cosine similarity, thereby improving clustering diversity and stability; adjustable thresholds are incorporated to determine membership conditions, optimally balancing clustering purity and overlap. The resulting multi-cluster membership structure provides a diverse set of highly relevant candidates for expert model selection within the subsequent energy-aware gating network, thereby significantly enhancing model fusion effectiveness. In summary, MCTC not only improves clustering quality but also establishes a robust topological foundation better suited for extremely heterogeneous data distributions, effectively satisfying stringent personalized requirements.

2.3. Aggregation strategies in PFL

Optimizing model aggregation strategies has been recognized as an effective means of enhancing the adaptability and performance of the global model, thereby mitigating the negative effects arising from discrepancies among client models. In environments with highly heterogeneous data distributions, appropriate adjustments to aggregation strategies have been shown to improve both convergence speed and global model accuracy. A comprehensive survey of aggregation techniques in federated learning has been provided by Qi et al. [43], covering 201 related studies and summarizing approaches ranging from standard to personalized, robust, secure, and weighted aggregation strategies. The need for more robust, personalized, and secure aggregation mechanisms has been emphasized in this survey.

Among the existing strategies, FedSmooth has been proposed by Wang et al. [12], introducing a scoring function to evaluate the overfitting degree of local models together with a smoothed moving-average weighting scheme to optimize aggregation. On the CIFAR-10 dataset, FedSmooth has outperformed FedAvg and FedProx by 5.4% and 3.3%, respectively. RLFL has been introduced by Imani et al. [44], combining mixed-precision models with a reinforcement-learning-based aggregation strategy tailored for resource-heterogeneous mobile edge environments, and accuracy improvements of 5% to 19% have been reported over existing methods. FedVa has been proposed by Liu et al. [13], incorporating both client data size and local model accuracy into weight assignment; on CIFAR-10, its final accuracy has improved by 2.3% in comparison with FedAvg. To balance stability and update speed, adaptive mixed aggregation (AMA) has been proposed by Li et al. [14], forming a weighted combination of the previous global model and current local models. Compared to approaches such as FedALA, AMA has achieved at least a 10% improvement in accuracy while significantly reducing communication overhead.

Despite the aforementioned advances, most aggregation strategies have continued to rely on the "single model return" approach, under which all clients receive only one global model at the end of training. A unified model, however, has struggled to meet the personalized requirements of diverse clients. Moreover, fixed aggregation approaches have failed to exploit knowledge representations derived from different clusters. Recent advancements have introduced techniques like parameter decoupling, hierarchical feature distillation, and cross-layer feature fusion. For example, the FedCPD framework proposed by Jin et al. [45] leverages these approaches to reduce information loss and better capture local insights. However, in soft clustering-based methods where clients obtain aggregation models from multiple clusters, effectively integrating multiple models still represents a core challenge.

To address this limitation, this paper proposes an energy-aware gating network-based mixed expert model strategy. Through this approach, private models and multiple expert models (cluster-level models returned by the server) are dynamically selected by training a local gating network. Crucially, rather than relying solely on input features, this gating mechanism incorporates an energy-based confidence calibration to evaluate the absolute predictive certainty of each expert. The outputs of these dynamically validated models are weighted and combined to produce more accurate and highly reliable personalized predictions. This mechanism enables clients to train private models locally using high-sensitivity data, thereby reducing the risk of privacy leakage. Furthermore, the energy-aware gating network has been designed to select the most suitable and confident expert models for the current task, enhancing generalization and robustness while ensuring that the private model continues to preserve personalized behavior.

3. Proposed method

3.1. Framework design

In non-IID data scenarios, traditional federated learning methods have struggled to balance client-specific requirements and generalization capabilities. This imbalance has led to poor global model performance, reflected in reduced accuracy and weak adaptation to local data. To address this challenge, a soft-clustering mixture-of-experts framework, SCMoE-PFL, is proposed. The framework is designed to reconcile generalization and personalization by combining an overlapping clustering scheme with an energy-aware mixture-of-experts fusion mechanism, while simultaneously mitigating privacy risks and preventing over-reliance on uncertain models.

The general workflow for federated training in the SCMoE-PFL framework is described as follows. The federated server initializes a global model and distributes it to each client. Each client i, according to its data characteristics and privacy preferences, partitions its local data D_i into two subsets: a low-sensitivity subset $D_{i,L}$ used to train the shared model, and a high-sensitivity subset $D_{i,H}$ used to train the private model. The private model is not uploaded to the server to reduce privacy leakage risks. After the shared model has been trained on $D_{i,L}$, the client uploads its gradients and model parameters to the server. The server applies the MCTC scheme to perform overlapping client clustering, carries out average aggregation within each cluster to generate expert models, and returns these expert models to clients belonging to the corresponding clusters. Each client selects the expert model with the highest accuracy from those it receives, and this model becomes its shared model for subsequent training. Once the shared model has converged or has reached the predefined number of training rounds, the client uploads its updated model information. This process is repeated across multiple rounds until the expert models converge. In parallel, each client uses the high-sensitivity subset $D_{i,H}$ to train a private model. Finally, client

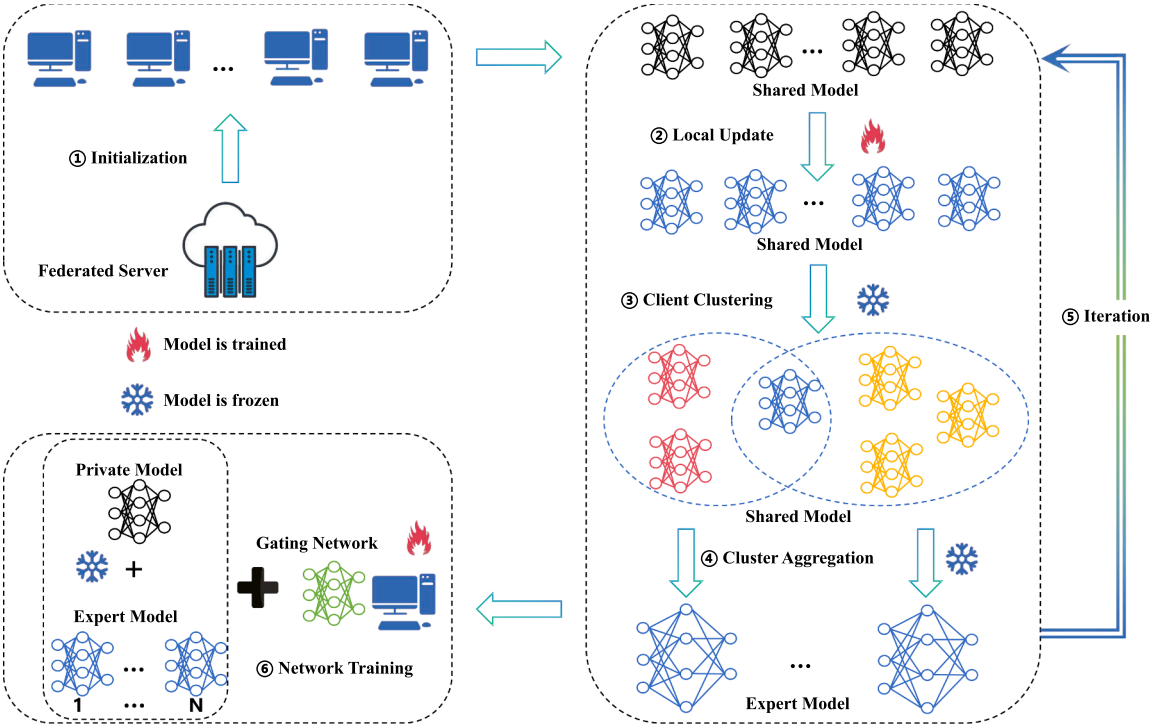


Fig. 2. SCMoE-PFL framework design flowchart.

i performs personalized prediction by dynamically combining the expert model(s) and its private model through an energy-aware gating network $G_i(\cdot)$, which leverages energy-based confidence calibration to improve the reliability and stability of the model fusion process. The overall pipeline is illustrated in Fig. 2, and the six steps are outlined below.

1) Initialization phase: The federated server initializes the global model parameters and distributes them to all participating clients. Upon receiving the initial model, clients begin local training. This phase ensures uniformity in model architecture and establishes the foundation for subsequent federated training.
2) Local training phase: Clients train the shared model (the initial model distributed by the server) using low-sensitivity data. After several iterations, updated gradients and model parameters are uploaded to the server. Meanwhile, clients leverage high-sensitivity data to train their private models, which are maintained locally and not shared with the server.
3) Client clustering phase: Based on the gradients uploaded by clients, the MCTC scheme is applied by the server to perform overlapping clustering. This approach allows clients to belong to multiple clusters simultaneously, thereby enhancing collaboration in heterogeneous data environments.
4) Intra-cluster aggregation phase: Average aggregation is performed within each cluster to generate an expert model, which is then returned to all clients belonging to that cluster. Since a client may belong to multiple clusters, it may receive several expert models and selects the one with the highest accuracy to use for the next round of local updates.
5) Iteration phase: The steps above constitute one complete round of federated training. This process is repeated iteratively, improving model performance until a convergence criterion is satisfied or the predetermined training rounds are completed.
6) Gating network training phase: Each client uses all of its local data to train an energy-aware gating network that dynamically selects the private model and expert model(s) and produces a weighted combination of their outputs as the final prediction. In the gating mechanism, the private model's prediction is always incorporated. Depending on the number of expert models, either all experts or the two most relevant experts with the highest calibrated weights are activated. Crucially, this selection is driven by an energy-based confidence calibration that jointly evaluates feature relevance and absolute predictive certainty, thereby improving the reliability of personalized predictions.

The above framework improves data utilization through overlapping clustering and addresses the problem of ambiguous cluster boundaries. By constructing a MoE module composed of a private model, expert models, and an energy-aware gating network on the client side, the framework satisfies client-specific requirements while maintaining generalization capability and robustness. Ultimately, it addresses the limitations of traditional federated learning methods in non-IID scenarios. The strong adaptability, uncertainty-aware fusion, and privacy protection mechanism enable SCMoE-PFL to exhibit robustness and flexibility in heterogeneous environments with stringent privacy requirements, making it suitable for a wide range of applications.

3.2. Client clustering

3.2.1. Gradient preprocessing

Existing clustering strategies are typically based on gradient vectors. In each federated training round, client i uses low-sensitivity data to train the shared model and uploads the updated model gradients $g_i \in \mathbb{R}^d$ to the server. The server then applies the MCTC scheme to assess similarity in data distributions between clients based on these gradients. However, gradient vectors are often high-dimensional, and directly computing cosine similarity in such a space incurs substantial computational overhead and is susceptible to the "curse of dimensionality", which results in unstable similarity measurements. Therefore, prior to clustering, dimensionality reduction must be performed on the gradient vectors to improve computational efficiency and enhance the representational capacity of the similarity structure.

Specifically, in uncompressed high-dimensional spaces, the distance or angular difference between any pair of data points tends to become

uniform due to the distance concentration effect. This phenomenon renders similarity metrics such as cosine similarity largely indistinguishable and unreliable [46]. By applying principal component analysis (PCA) [47], the gradients are projected onto a compact, lower-dimensional subspace that preserves the directions of maximum variance. This projection effectively filters out the sparse, isotropic noise dimensions responsible for the aforementioned concentration effect. Consequently, the cosine similarity computed in this reduced-dimensional space avoids geometric distortion and accurately reflects the intrinsic divergence in client data distributions, thereby ensuring a highly reliable similarity measurement for clustering.

In this work, PCA is employed to reduce the dimensionality of the gradient vectors uploaded by the clients. Specifically, to standardize the influence of gradients with different magnitudes, the gradient vector g_i of client i is first normalized to remove magnitude information while retaining its direction. The normalization is carried out as follows:

$$\hat{g}_i = \frac{g_i}{\|g_i\|_2}, \quad \|g_i\|_2 = \sqrt{\sum_{k=1}^{d} g_{i,k}^2} \tag{1}$$

Here $\|g_i\|_2$ represents the L_2 norm of the vector, corresponding to the Euclidean length of the gradient vector. Through normalization, all gradient vectors are standardized to unit length while preserving their directional characteristics.

Next, PCA is applied to extract the most representative principal components from the gradient data, thereby mapping the original high-dimensional vectors into a lower-dimensional embedding space. Letting the projection matrix be $V \in \mathbb{R}^{d\times m}$, the reduced-dimensional representation is given by:

$$z_i = \hat{g}_i V, \quad z_i \in \mathbb{R}^m \, (m \ll d) \tag{2}$$

Gradient preprocessing not only reduces the dimensionality of gradient vectors but also eliminates redundant noisy components, thus improving the efficiency and accuracy of similarity calculations in subsequent clustering operations. This step is particularly suitable for federated learning scenarios involving a large number of clients and frequent clustering. The reduced-dimensional embedding vectors are finally used as inputs to the MCTC module, where cosine similarity is computed between clients and clustering is performed, establishing the foundation for subsequent model fusion.

Algorithm 1 MCTC.

Input: Set of clients $C = \{c_1, c_2, \dots, c_N\}$, client gradients $\Delta G = \{\Delta z_1, \Delta z_2, \dots, \Delta z_N\}$, initial cluster count K, similarity threshold $\tau \in [0, 1]$

Output: Set of clusters $S = \{S_1, S_2, \dots, S_K\}$

Compute cosine similarity matrix $Sim \in \mathbb{R}^{N\times N}$

$Sim_{ij} = \cos(\theta_{ij}) = \frac{\Delta z_i \cdot \Delta z_j}{\|\Delta z_i\| \|\Delta z_j\|}$, $\forall i, j = 1, 2, \dots, N$

Select initial cluster centers

Initialize a set $C_{\text{centers}} = \emptyset$

Initialize a set of empty clusters $S = \{S_1, S_2, \dots, S_K\}$

$k \leftarrow 1$

while $k \le K$ **do**

  Find the pair of clients (c_i, c_j) with the smallest Sim_{ij}

  for client c_m in $\{c_i, c_j\}$ **do**

    if $c_m \notin C_{\text{centers}}$ **and** $k \le K$ **then**

      Add c_m to C_{centers} and S_k

      Initialize cluster center $v_k \leftarrow \Delta z_m$

      $k \leftarrow k + 1$

    end if

  end for

  $Sim_{ij} \leftarrow \infty$ ▷ Prevent re-selecting the same pair

end while

Create remaining unassigned client set $C_{\text{others}} = C - C_{\text{centers}}$

Assign C_{others} **to** S

for all client $c_i \in C_{\text{others}}$ **do**

  $IsAssigned \leftarrow$ False

  for all cluster center v_k of S_k **in parallel do**

    if $\cos(\theta_{ik}) \ge \tau$ **then**

      Assign c_i to S_k

      $IsAssigned \leftarrow$ True

    end if

  end for

  if $IsAssigned ==$ False **then**

    Find $k^* = \arg\max_k \cos(\theta_{ik})$

    Assign c_i to S_{k^*}

  end if

  Update all cluster center vectors v_k according to Equation (4)

end for

Return $S = \{S_1, S_2, \dots, S_K\}$

3.2.2. MCTC

In federated learning, the highly heterogeneous distribution of client data and diverse personalized requirements often cause traditional hard clustering methods to perform inadequately. These methods assign each client to a single cluster; however, this design has limitations because a client's information can only be utilized within its assigned cluster. As a result, potential multiple similarities across clusters cannot be fully exploited, and the performance of the final model may be adversely affected. Furthermore, gradient vectors are a type of data whose categorization is inherently ambiguous, and hard clustering methods have often struggled to handle uncertain cluster boundaries, leading to misclassification or underrepresentation. Overlapping clustering methods are better suited for such cases as they enable clients to participate in multiple clusters simultaneously. The SCMoE-PFL framework adopts the MCTC method, which improves data utilization and enhances robustness when handling complex, irregular, or noisy data distributions, thereby yielding superior performance under non-IID conditions.

Specifically, the MCTC scheme computes the cosine similarity between the gradients of all clients $\{c_1, c_2, \dots, c_N\}$ and selects K initial cluster centers from several client pairs (c_i, c_j) exhibiting the smallest cosine similarities. The number of clusters K is determined based on domain knowledge and clustering validation results. Subsequently, the cosine similarity between each client and all cluster centers is computed, and clients are assigned to the corresponding clusters based on the predefined similarity threshold τ. In addition, MCTC adaptively adjusts the threshold through a clustering validation mechanism, offering flexible control over clustering granularity. This adaptation ensures that compact clusters can be maintained while cross-cluster information is incorporated. As a result, MCTC provides more effective client partitioning in heterogeneous environments and facilitates personalized model fusion in subsequent energy-aware gating networks, thereby enhancing both personalization and generalization in federated learning.

The algorithm proceeds as follows. First, the cosine similarity matrix $\text{Sim} \in \mathbb{R}^{N\times N}$ is calculated. Then, several of the most dissimilar client pairs (c_i, c_j) are selected, from which the initial cluster centers are chosen until K clusters have been initialized, with each cluster initially containing only one client. Remaining clients are then assigned to one or more clusters that meet the similarity threshold τ. If no cluster satisfies the threshold, the client is assigned to the most similar cluster S_{k^*}, where $k^* = \arg\max_k \cos(\theta_{ik})$. When a new client joins, cluster centers are updated according to Eq. (4). Finally, the cluster set $S = \{S_1, S_2, \dots, S_K\}$ is generated. The complete procedure is shown in Algorithm 1.

Through this process, MCTC enables clients to be assigned to multiple clusters, thus balancing local compactness with global diversity. Compared with traditional hard clustering, this method mitigates challenges arising from non-IID data, preserves client privacy, and fully utilizes cross-cluster information. MCTC provides a robust clustering foundation for the gating-network-driven personalized model fusion and

improves the overall convergence efficiency and model generalization in federated learning.

3.2.3. Clustering validation

After MCTC, a clustering validation mechanism based on inter-cluster similarity is introduced to ensure the validity and stability of the partitioning results and to prevent the formation of clusters that are too small or insufficiently discriminative. Specifically, after each clustering round, the overall inter-cluster separability is assessed by computing the average cosine similarity among the current cluster-center vectors.

Let the number of clusters be denoted as K, and denote the center vector of cluster S_k by v_k. The average cosine similarity among all cluster-center vectors is defined as:

$$\cos_{\text{avg}} = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \cos(v_i, v_j) \tag{3}$$

The cluster center vector v_k is calculated as the average of the gradient vectors z_i:

$$v_k = \frac{1}{|S_k|} \sum_{c_i \in S_k} z_i \tag{4}$$

Here, v_k can be regarded as the overall "direction" or "representation" of all client gradient features within the cluster S_k. By calculating $\cos_{\text{avg}}$, the overall similarity of the cluster partition can be assessed. This metric reflects whether clusters are sufficiently distinct from one another. If the center vectors of multiple clusters point in nearly the same direction, the average similarity approaches 1, indicating minimal differences among clusters and insufficient separability. In such cases, the clustering result is reverted, and the initial cluster centers or similarity threshold is adjusted for re-partitioning. Conversely, a low average similarity indicates that the clusters are more dispersed in the feature space, resulting in clearer and more meaningful clustering outcomes.

The method outlined above considers factors such as the number of clusters and sample distribution, enabling adaptive adjustment of clustering criteria. In this way, the effectiveness of the clustering process is maintained while preventing excessive partitioning, thereby providing a stable and distinguishable foundation for subsequent model fusion.

3.2.4. Intra-cluster aggregation

After the server completes the MCTC scheme and passes clustering validation, the process proceeds to intra-cluster client model aggregation. The server performs a weighted aggregation of the model parameters uploaded by clients for every cluster, with weights determined by the number of data samples held by each client, thereby generating a cluster-level model that represents the entire cluster and serves as the expert model for the client's MoE module. This weighting mechanism ensures that clients with larger datasets contribute more significantly to the updated expert model. The mathematical expression for the weighted aggregation is:

$$\omega_{t+1}^{S_k} = \sum_{i=1}^{|S_k|} \frac{n_i}{n_{S_k}} \omega_t^i \tag{5}$$

Here, $|S_k|$ represents the number of clients in cluster S_k, ω_t^i denotes the model parameters uploaded by the client c_i in cluster S_k during the tth communication round, n_i is the number of data samples from the client c_i in cluster S_k, and n_{S_k} represents the total number of data samples from all clients in cluster S_k, i.e., $n_{S_k} = \sum_{i=1}^{|S_k|} n_i$. The result $\omega_{t+1}^{S_k}$ denotes the expert model parameters obtained after this round of weighted aggregation.

Since a client may belong to multiple clusters, it may receive one or more expert models. Each client then selects the expert model with the highest accuracy for the next round of local training. This alternating process of local updates and intra-cluster aggregation is repeated until the expert model reaches the target accuracy or the predetermined number of training rounds is completed.

3.3. Energy-aware gating network

In the SCMoE-PFL framework, the MCTC scheme allows clients to belong to multiple clusters simultaneously, and the server distributes each cluster's aggregated expert model to the corresponding clients. As a result, a single client may receive multiple expert models. Under such conditions, relying on a single expert model becomes restrictive. Similarly, producing one unified model through a fixed aggregation strategy also proves insufficient, as it cannot balance cross-client generalization with the personalization required for local data adaptation. Meanwhile, the private model trained locally on high-sensitivity data excels at fitting the local distribution but lacks the ability to integrate knowledge from other clients. To address this issue, an energy-aware gating network is introduced to dynamically weight the private model and multiple expert models. Unlike traditional routers that rely solely on input features, this mechanism jointly evaluates the input representation and the absolute predictive confidence (via energy scores) of each expert. Consequently, it securely filters out 'guessing' models for specific samples, enhancing both personalization and reliability while preserving generalization capability.

The working mechanism of the energy-aware gating network can be summarized as follows. First, client i constructs a candidate model set $M_i = \{m_i^{PM}\} \cup M_i^{EM}$, which consists of the private model m_i^{PM} trained on its high-sensitivity data $D_{i,H}$ and the expert models $M_i^{EM} = \{m_{i1}, m_{i2}, \dots, m_{iJ_i}\}$ returned by the server. In this way, both local knowledge and cross-cluster knowledge are incorporated into the fusion process. Next, the sample feature x is fed into the gating network $G_i(\cdot)$ to generate initial importance scores $s_i = s_i^{PM} \cup [s_{i1}, s_{i2}, \dots, s_{iJ_i}]$ for each candidate model, which are subsequently transformed into prior feature-matching weights $\alpha_{ik}^{\text{prior}}$. Crucially, instead of the direct selection of experts based on these initial priors, an energy-based confidence calibration step [48] is introduced. For each candidate model m_{ik}, we extract its unnormalized logits $l_{ik} = [l_{ik}^1, \dots, l_{ik}^C]$ produced for the input sample x (where C is the number of classes) to evaluate its Free Energy, defined as $E_{ik} = -T \log \sum_{c=1}^{C} \exp(l_{ik}^c / T)$. Here, the temperature parameter T is typically set to 1 to preserve the original magnitude of the logits without artificially smoothing or sharpening the energy surface. This Free Energy serves as a robust indicator of out-of-distribution uncertainty [48]. Rather than relying on arbitrary mapping functions to convert this unbounded energy metric into a usable routing weight, we draw upon the established conventions of Energy-Based Models (EBMs) to naturally bridge the gap between energy and probability. Under this paradigm, the Gibbs distribution dictates that a state's likelihood is proportional to the negative exponential of its energy. Guided by this physical intuition, we define our posterior confidence multiplier simply as the negative exponential of the Free Energy. Conveniently, this formulation mathematically resolves to the partition function of the expert's logit space: $\beta_{ik} = \exp(-E_{ik}/T) = \sum_{c=1}^{C} \exp(l_{ik}^c / T)$. This transformation effectively converts an uncertainty penalty into a strictly positive confidence score. As a result, experts with higher overall logit activations, a strong indicator of in-distribution familiarity, are exponentially rewarded. The prior weights are dynamically calibrated by this multiplier to yield the uncertainty-aware weights $\tilde{\alpha}_{ik} = \alpha_{ik}^{\text{prior}} \cdot \beta_{ik}$. During output fusion, the private model is always retained, and the expert models are selected based on these calibrated weights $\tilde{\alpha}_{ik}$. If $J_i > 2$, only the two experts with the highest calibrated weights are chosen (as empirical evaluations in Section 4.5.2 demonstrate this Top-2 strategy to be Pareto optimal for preventing out-of-distribution noise from lower-ranked experts while strictly bounding computational overhead). Finally, the calibrated weights of the selected models are normalized to produce the final output p.

During this process, only the parameters of $G_i(\cdot)$ are updated, while the parameters of m_i^{PM} and M_i^{EM} remain frozen to avoid disrupting their learned distribution characteristics. The formalized mechanism is presented in Algorithm 2, which meticulously describes the workflow from constructing the candidate expert model set and initial scoring, to

Algorithm 2 Energy-aware gating network.

Input: Client i's local dataset $(x, y) \in D_i$, client private model m_i^{PM} trained on highly sensitive subset $D_{i,H}$, server-returned expert models $M_i^{EM} = \{m_{i1}, m_{i2}, \dots, m_{iJ_i}\}$, J_i represents the number of clusters to which client i belongs, temperature parameter T

Output: Gating network $G_i(\cdot)$ that outputs energy-calibrated model-selection weights

The client i constructs candidate expert set

$M_i = \{m_i^{PM}\} \cup M_i^{EM}$

Gating network scoring

Input sample x into $G_i(\cdot)$ to obtain importance scores:

$s_i = s_i^{PM} \cup [s_{i1}, s_{i2}, \dots, s_{iJ_i}]$

Energy-based confidence calibration

Compute prior weights for all $J_i + 1$ candidate models:

$\alpha_{ik}^{\text{prior}} = \frac{\exp(s_{ik})}{\sum_{j=1}^{J_i+1} \exp(s_{ij})}, \quad \forall k \in \{1, 2, \dots, J_i + 1\}$

for each $m_{ik} \in M_i$ **do**

  Obtain raw logits $l_{ik} = [l_{ik}^1, \dots, l_{ik}^C]$ from $m_{ik}(x)$, where C is the number of classes

  Compute confidence multiplier:

  $\beta_{ik} = \sum_{c=1}^{C} \exp(l_{ik}^c / T)$

  Calibrate weight: $\tilde{\alpha}_{ik} = \alpha_{ik}^{\text{prior}} \cdot \beta_{ik}$

end for

Expert selection strategy

if $1 \le J_i \le 2$ **then**

  $M_i^{selected} = M_i$

else

  Let $\mathcal{K}$ be the indices of the top-2 calibrated weights in $\{\tilde{\alpha}_{i1}, \dots, \tilde{\alpha}_{iJ_i}\}$

  $M_i^{selected} = \{m_i^{PM}\} \cup \{m_{ik} \mid k \in \mathcal{K}\}$

end if

Weight normalization

Compute final normalized weights for selected experts:

$\alpha_{ik} = \frac{\tilde{\alpha}_{ik}}{\sum_{m_{ij} \in M_i^{selected}} \tilde{\alpha}_{ij}}, \quad \forall m_{ik} \in M_i^{selected}$

Generate ensemble prediction

$p = \sum_{m_{ik} \in M_i^{selected}} \alpha_{ik} \cdot m_{ik}(x)$

Gating network training

Compute loss: $\Gamma_i = \text{CrossEntropy}(p, y)$

Compute gradients w.r.t. $G_i(\cdot)$ parameters: $\nabla \Gamma_i$

Update only parameters of $G_i(\cdot)$; all models M_i remain fixed

Return $G_i(\cdot)$

the critical energy-based confidence calibration, followed by the adaptive selection of experts, weight normalization, and generation of the fused prediction. The algorithm explicitly defines the forward propagation and parameter-update strategy, ensuring that only the gating network parameters are optimized while all other models remain fixed, thereby achieving efficient and robust model fusion with preserved generalization.

Algorithm 2 enables dynamic, uncertainty-aware selection and weighted fusion of candidate models on the client side. This design not only effectively balances personalization with the generalization provided by cross-client knowledge but also establishes a reliable dual-check mechanism. By integrating the energy confidence score, the gating network successfully filters out models that lack absolute predictive certainty for specific out-of-distribution (OOD) or high-noise inputs. As discussed in Section 3.4, the gating network directly affects the lower bound of accuracy, underscoring its theoretical feasibility in improving personalized federated learning. Furthermore, by adapting the number of models involved in the fusion process based on the candidate expert set size, and keeping model parameters fixed during its training phase, the energy-aware gating network achieves an optimal trade-off among classification performance, inference efficiency, and system reliability.

3.4. Theoretical analysis

3.4.1. Analysis of computation cost and communication overhead

Since SCMoE-PFL trains multiple models (shared model, private model, and gating network) on the client side and performs MCTC soft clustering and weighted aggregation on the server, its computational overhead is higher than that of FedAvg. This increased cost arises mainly from the parallel training of multiple models and the handling of clustering information. However, the communication overhead exhibits an asymmetric pattern. The uplink overhead remains comparable to that of FedAvg, as SCMoE-PFL only transmits the shared model information (gradients and parameters) while strictly retaining the private model locally. Regarding the downlink, although receiving multiple expert models inevitably introduces higher bandwidth requirements than FedAvg, the server only transmits a small, relevant subset of experts to each client, thereby keeping the communication load well within practical limits. Consequently, personalization is substantially enhanced while generalization is preserved. Although the computational and downlink overheads are slightly increased, the overall costs remain manageable, making SCMoE-PFL highly suitable for scenarios requiring robust personalization under heterogeneous data distributions.

In terms of computational complexity, the local training complexity for each client in SCMoE-PFL is $\mathcal{O}(E \cdot D \cdot M)$, where E, D, and M denote the local epochs, dataset size, and model parameter size, respectively. This shares the same asymptotic complexity as FedAvg. However, due to the concurrent forward passes of the selected experts and the gating network, the actual inference overhead (a constant factor) is moderately higher, which is bounded in practice by our Top-2 expert selection strategy.

3.4.2. Privacy protection analysis

In the SCMoE-PFL framework, only the shared model information (gradients and parameters) is uploaded to the server, while private models are trained locally on high-sensitivity data without transmitting any additional client information. This design significantly reduces the risk of sensitive data exposure to inference attacks. Client clustering is performed by the server based solely on dimensionality-reduced gradient information, and model aggregation relies only on model parameters, further minimizing privacy risks. Crucially, during inference, the extraction of unnormalized logits is executed entirely on the local client side. No confidence scores or predictive distributions are shared. Through these decentralized mechanisms, SCMoE-PFL improves model performance while reducing the exposure of sensitive client information.

3.4.3. Performance lower bound analysis and practical robustness

In SCMoE-PFL, the client performs inference by dynamically fusing the private model and selected expert models through an **energy-aware gating network**. Unlike standard static averaging or naive feature-matching routing, our energy-aware gating network assigns dynamic, uncertainty-calibrated weights ($\tilde{\alpha}_{ik}$) based on the joint evaluation of input features and absolute predictive confidence. To analyze the theoretical lower bound of the fusion accuracy, we first establish the ideal convergence state.

Assumption 1. **Ideal State Assumption:** 1) **Optimal Calibrated Selection:** The energy-aware gating network can perfectly quantify predictive uncertainty, identifying and assigning the maximum calibrated weight to the optimal expert model that yields the minimum loss for any given input sample x. 2) **Expert Complementarity:** The error distributions of the candidate models (expert models and private model) are diverse, meaning they specialize in different subsets of the overall data distribution.

Let $M_i^{selected}$ represent the set of models selected by client i, including the private model m_i^{PM} and the top-ranked expert models $m_{ik} \in M_i^{EM}$. For a specific input sample (x, y), let $\mathcal{L}(m_{ik}, x, y)$ represent the prediction

loss of model m_{ik}. Under the ideal state (Assumption 1), the energy-based confidence multiplier β_{ik} perfectly suppresses irrelevant experts. The network thus assigns a dominant calibrated weight to the optimal model m^* for the current sample x, such that:

$$\tilde{\alpha}_{m^*}(x) \to 1, \quad \text{where } m^* = \arg \min_{m_{ik} \in M_i^{selected}} \mathcal{L}(m_{ik}, x, y) \tag{6}$$

Consequently, the expected accuracy of the fused system, denoted as Acc_i^{final}, is strictly lower-bounded by the accuracy of the single best-performing model within the candidate set. Mathematically, this theoretical lower bound is expressed as:

$$Acc_i^{final} \geq \max\left(Acc(m_i^{PM}), \max_{m_{ik} \in M_i^{EM}} Acc(m_{ik}) \right) \tag{7}$$

Under the ideal routing assumption, this inequality indicates that instance-wise dynamic weighting can in principle achieve performance no worse than the best available candidate model. Furthermore, owing to the Expert Complementarity (Assumption 1), the ideal dynamic routing effectively integrates specialized knowledge across diverse subsets, enabling Acc_i^{final} to strictly exceed the global maximum of any single expert.

Robustness in Practical Scenarios (Relaxation of the Ideal State): While Eq. (7) provides an idealized theoretical bound, practical federated learning deployments are inherently perturbed by factors such as insufficient local training, massive data noise, and extreme label skewness. Under such non-ideal conditions, Assumption 1 is practically relaxed, meaning the gating network cannot always achieve perfect $\tilde{\alpha}_{m^*}(x) \to 1$ assignment.

Standard routing mechanisms suffer catastrophic routing errors under these perturbations, leading to severe deviations from the theoretical lower bound. However, our **energy-aware gating mechanism** explicitly mitigates this degradation. When data noise or insufficient training causes an expert to produce flat, unconfident logit distributions, the Free Energy-based posterior multiplier (β_{ik}) inherently penalizes this uncertainty. This **theoretically constrains** the routing error margin by exponentially decaying the weights of uncertain models, effectively preventing "guessing" experts from hijacking the fusion process. Therefore, even when the ideal assumptions are relaxed under pathological non-IID conditions, the energy-guided dual-check mechanism acts as a robust safeguard, ensuring that the empirical performance of SCMoE-PFL remains tightly coupled to the theoretical lower bound.

Remark 1. The distinctive novelty of this work lies in the joint design of a soft-clustering mechanism and an uncertainty-aware mixture-of-experts fusion strategy within a unified personalized federated learning framework. Specifically, the proposed MCTC enables overlapping client memberships, overcoming the limitations of traditional hard clustering and substantially improving data utilization under heterogeneous scenarios. Crucially, an energy-aware gating network is introduced to dynamically fuse the local private model with multiple cluster-level expert models. By calibrating feature-matching priors with Free Energy-based predictive confidence, this dual-check mechanism effectively penalizes spurious expert representations without increasing communication overhead. This mathematically grounded dual innovation provides a more expressive, reliable, and flexible personalization capability than existing PFL and MoE-based FL methods, ensuring enhanced stability, fairness, and robustness in highly non-IID environments.

4. Evaluation and experimental analysis

In this section, a series of systematic experiments are conducted on four widely adopted federated learning benchmark datasets to evaluate the effectiveness of SCMoE-PFL in non-IID scenarios. These datasets encompass character recognition and image classification tasks and, after configuration, exhibit substantial heterogeneity, thereby enabling a comprehensive assessment of the framework's adaptability to complex data environments. The experimental design is described in the following subsections.

4.1. Experimental setup

To provide a comprehensive evaluation of the personalization capability of the SCMoE-PFL in non-IID environments, four diverse federated datasets are used: FEMNIST [49], EMNIST [50], CIFAR-10 [51], and CIFAR-100 [51].

Hardware Platform: Intel Core i7-10875H processor, 16GB RAM, NVIDIA GeForce RTX 2060 GPU. **Software Platform:** Windows 10, Python 3.8, TensorFlow 2.8. **Experimental Settings:** 140 communication rounds, with each client performing 50 local training rounds per communication round. SGD is used as the optimizer, with a learning rate of 0.01 and momentum of 0.9. Client data are partitioned using a Dirichlet distribution to simulate moderate heterogeneity (significant distribution differences between clients) and extreme heterogeneity (highly skewed data and severe class imbalance).

Baseline Methods: FedAvg [15] is widely used for IID settings. FedProx [30] enhances stability in non-IID environments through a proximal term. Ditto [52] learns a personalized model for each device while preserving the consistency of the global model. PM-MoE [53] integrates a mixture-of-experts structure to enhance the reinforcement between local personalized modules and shared knowledge. PFMoE [54] balances personalization and generalization and is particularly effective in environments with substantial data heterogeneity.

Evaluation Metrics: *Average Test Accuracy* measures overall predictive performance across clients. *Average F1-score* evaluates classification stability, particularly under class imbalance. *Standard Deviation of Accuracy Between Clients* reflects fairness, with smaller values indicating more stable client-level performance. *Speed* denotes the average inference time (ms) per small batch, including energy-aware gating-network computation and private/expert model inference.

4.2. Experimental results

Tables 1 and 2 present the performance of different methods on four datasets under moderate heterogeneity ($\alpha = 1$) and extreme heterogeneity ($\alpha = 0.5$). SCMoE-PFL consistently achieves superior performance across all datasets and heterogeneity levels, demonstrating strong adaptability to diverse non-IID environments.

For $\alpha = 1$ (Moderate Heterogeneity), SCMoE-PFL achieves an average accuracy of 96.11% across four datasets, which represents an improvement of approximately 19.22 percentage points over the baseline method FedAvg, significantly outperforming FedProx and Ditto. Compared to FedAvg, the greatest improvement is observed on EMNIST, with an increase of 24.71 percentage points. Compared to the most advanced MoE-based baseline method PFMoE, SCMoE-PFL achieves accuracy gains of approximately 4.50, 2.52, and 3.48 percentage points on FEMNIST, EMNIST, and CIFAR-10, respectively, and remains comparable on CIFAR-100 (slightly better by 1.69 percentage points). These results highlight the benefits of the soft clustering algorithm and the energy-aware gating network. The improvements in F1-score further indicate enhanced classification stability. Inference speed is maintained between 0.17 and 0.21 ms per batch, comparable to PM-MoE and PFMoE and faster than certain competing methods such as Ditto, demonstrating that efficient inference is retained.

For $\alpha = 0.5$ (Extreme Heterogeneity), the heightened non-IID data distribution poses significant challenges for all evaluated methods; however, SCMoE-PFL consistently demonstrates superior performance. On the particularly challenging CIFAR-10 dataset, SCMoE-PFL achieves an impressive accuracy of 95.67%. This translates to a substantial advantage, yielding improvements of 19.80 and 7.30 percentage points over FedAvg and PFMoE, respectively. Furthermore, when compared to PFMoE, our method achieves absolute accuracy gains of approximately 5.85, 8.02, and 4.31 percentage points on FEMNIST, EMNIST,

Table 1
Algorithm performance comparison when $\alpha = 1$.

Method	FEMNIST			EMNIST			CIFAR10			CIFAR100		
	Acc	F1	Speed	Acc	F1	Speed	Acc	F1	Speed	Acc	F1	Speed
FedAvg	0.7354	0.7375	0.25	0.7156	0.7164	0.24	0.8388	0.8395	0.17	0.7856	0.7854	0.24
FedProx	0.8242	0.8231	0.26	0.8127	0.8134	0.23	0.8675	0.8697	**0.16**	0.8342	0.8341	0.27
Ditto	0.8305	0.8325	0.22	0.8139	0.8144	0.23	0.8924	0.8931	0.17	0.8147	0.8125	0.27
PM-MoE	0.8994	0.9012	**0.17**	0.8578	0.8556	0.22	0.8857	0.8876	**0.16**	0.8836	0.8823	0.22
PFMoE	0.9266	0.9265	0.18	0.9375	0.9364	0.19	0.9236	0.9247	0.18	0.9347	0.9329	**0.19**
SCMoE-PFL	**0.9716**	**0.9714**	**0.17**	**0.9627**	**0.9619**	**0.17**	**0.9584**	**0.9576**	0.17	**0.9516**	**0.9518**	0.21

Table 2
Algorithm performance comparison when $\alpha = 0.5$.

Method	FEMNIST			EMNIST			CIFAR10			CIFAR100		
	Acc	F1	Speed	Acc	F1	Speed	Acc	F1	Speed	Acc	F1	Speed
FedAvg	0.6978	0.6997	0.24	0.6774	0.6784	0.22	0.7587	0.7585	0.18	0.7448	0.7439	0.27
FedProx	0.7865	0.7855	0.24	0.7726	0.7729	0.21	0.7873	0.7871	0.16	0.7941	0.7939	0.26
Ditto	0.7945	0.7944	0.21	0.7837	0.7834	0.21	0.8123	0.8125	0.16	0.7745	0.7738	0.27
PM-MoE	0.8545	0.8543	0.18	0.8375	0.8366	0.22	0.8359	0.8366	0.17	0.8433	0.8432	**0.21**
PFMoE	0.8897	0.8905	**0.16**	0.8573	0.8564	**0.18**	0.8837	0.8847	0.15	0.8948	0.8956	0.22
SCMoE-PFL	**0.9482**	**0.9483**	**0.16**	**0.9375**	**0.9376**	**0.18**	**0.9567**	**0.9566**	**0.14**	**0.9379**	**0.9368**	**0.21**

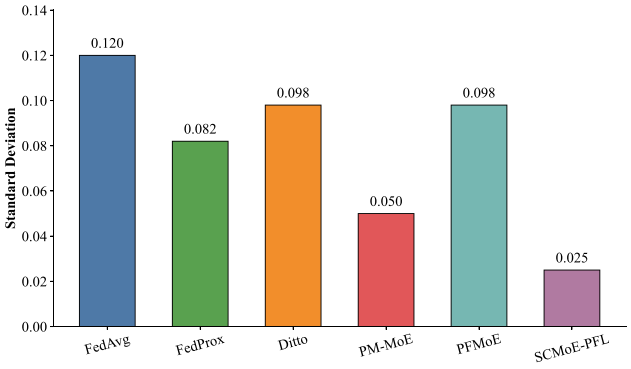


Fig. 3. Standard deviation of client accuracy under different methods when $\alpha = 0.5$.

and CIFAR-100, respectively. Most notably, the greatest performance leap relative to FedAvg is observed on EMNIST, exhibiting a 26.01 percentage-point increase. Finally, the inference speed also shows a slight improvement, highlighting the efficacy of the model's efficient expert selection mechanism.

Convergence Speed and Stability: Fig. 3 compares the standard deviation of the final test accuracy across all clients for different methods on CIFAR-100 ($\alpha = 0.5$). This metric measures the fairness and stability of the algorithm's performance across clients with diverse data distributions (smaller values are preferable). The results show that SCMoE-PFL significantly improves fairness, with
the lowest standard deviation in client accuracy, well below global model methods like FedAvg and FedProx, and also lower than Ditto and PM-MoE. This demonstrates that its personalization strategy effectively adapts to non-IID scenarios, preventing the dominance of a few data-rich or "mainstream" clients in certain expert models, which could lead to poor performance for other clients. Although PFMoE also uses MoE, the singularity of its global model restricts the ability of clients to leverage cross-cluster knowledge. When faced with highly personalized or boundary-blurred clients, its stability is slightly weaker than that of SCMoE-PFL, which supports multiple cluster memberships.

Fig. 4 shows the training loss curves of different schemes across four datasets as a function of communication rounds. SCMoE-PFL demonstrates the fastest and most stable convergence process. Compared to global model methods such as FedAvg and FedProx, its loss decreases more rapidly and ultimately converges to a lower level. In comparison to MoE methods (e.g. PFMoE), SCMoE-PFL's convergence curve is smoother, with less oscillation, reflecting the stability advantage afforded by the stable clustering structure provided by MCTC and the effective fusion enabled by the energy-aware gating network.

4.3. Ablation study

To investigate the contributions of the two core components, namely, MCTC soft clustering and the gating network, an ablation study is conducted on four datasets ($\alpha = 0.5$), with results shown in Fig. 5.

Removing MCTC Soft Clustering (w/o MCTC): When MCTC is replaced by traditional K-means [55] hard clustering (where each client belongs to only one cluster), the energy-aware gating network is retained, but clients only receive and fuse expert models and local private models from their assigned cluster. Consequently, performance declines by 8.7 percentage points, with the average accuracy falling to around 86%. This highlights the importance of the soft clustering method. The MCTC effectively identifies multiple similarities in client data, offering a richer set of expert model candidates for the gating network, thereby significantly enhancing fusion effectiveness and model expressiveness.

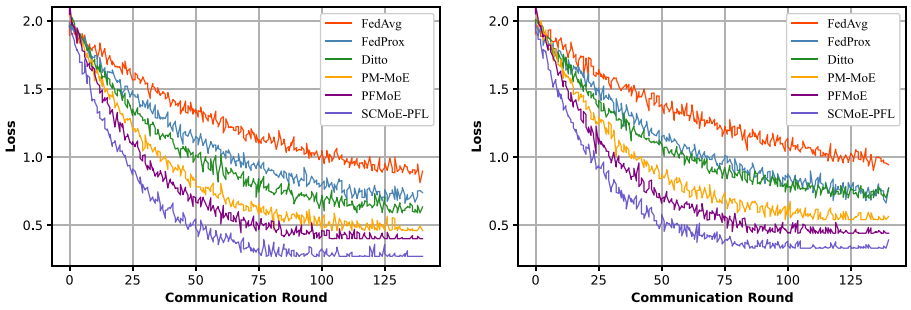
Removing the energy-aware gating network (w/o Gating Network): In this variant, only the MCTC soft clustering and intra-cluster average aggregation are retained. After receiving the expert models, clients perform average aggregation to create a single model for predictions. Consequently, performance drops significantly by approximately 18 percentage points, with the average accuracy decreasing from 94.51% to around 76.47%. This highlights the critical role of the gating network in dynamically fusing the local private model with multiple expert models to achieve deep personalization.

Complete SCMoE-PFL: By integrating MCTC soft clustering and gating fusion, the performance reaches its optimal level.

The ablation study demonstrates that both MCTC soft clustering and the energy-aware gating network are essential components of SCMoE-PFL. Removing either module leads to clear performance degradation. The results on CIFAR-100 confirm their complementary contributions and show that the combination of both modules significantly enhances accuracy and convergence stability.

4.4. Clustering quality assessment

To quantitatively evaluate the reliability and quality of our proposed MCTC mechanism, we introduced the Silhouette Score [56] as a rigorous



(a) Loss vs. Communication Round on FEMNIST

(b) Loss vs. Communication Round on EMNIST

(c) Loss vs. Communication Round on CIFAR10

(d) Loss vs. Communication Round on CIFAR100

Fig. 4. Loss vs. Communication Rounds comparison on multiple datasets when $\alpha = 0.5$.

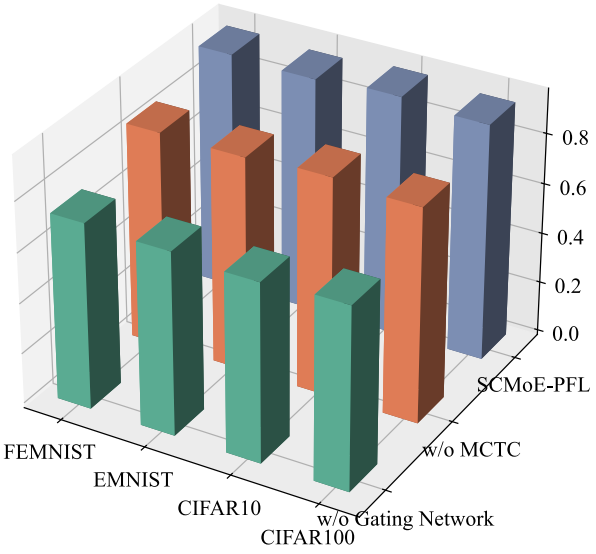


Fig. 5. Ablation study on SCMoE-PFL: impact of the energy-aware gating network and MCTC on performance across four datasets.

clustering evaluation metric. The Silhouette Score, ranging from -1 to 1, measures the cohesion within a cluster and the separation between different clusters. A higher score indicates a more reasonable and distinct cluster configuration. We compared the proposed MCTC method against two widely adopted baselines: the traditional centroid-based K-Means [55], and the connectivity-based Agglomerative Hierarchical Clustering (HC) [57], which is frequently utilized in standard clustered federated learning. As reported in Table 3, the clustering evaluation is conducted across the four datasets under extreme heterogeneity ($\alpha = 0.5$). The quantitative results demonstrate that MCTC achieves an average Silhouette Score of 0.65, substantially outperforming both K-Means (0.43) and HC (0.50).

Table 3
Quantitative evaluation of clustering quality using Silhouette Score across different datasets ($\alpha = 0.5$).

Clustering Method	FEMNIST	EMNIST	CIFAR-10	CIFAR-100
K-Means	0.45	0.48	0.41	0.38
HC	0.52	0.54	0.49	0.44
MCTC	**0.68**	**0.71**	**0.62**	**0.59**

These quantitative results provide clear evidence of the limitations inherent in traditional deterministic methods. By employing mutually exclusive cluster assignments, K-Means and HC assign clients with ambiguous data distributions into isolated groups, which tends to restrict intra-cluster cohesion. In contrast, the MCTC method, by enabling overlapping memberships and adaptively adjusting similarity thresholds, effectively captures the intricate similarity structures within the high-dimensional gradient space. Consequently, MCTC forms highly cohesive and reliable expert groups, providing an optimal model-fusion foundation for the subsequent energy-aware gating network.

4.5. Hyperparameter sensitivity and robustness

To comprehensively validate the stability of the proposed framework and its robustness against ultra-extreme data distributions, we perform rigorous sensitivity analysis on two core hyperparameters: the MCTC clustering threshold τ and the number of activated experts K, and further evaluate the system's performance under an ultra-extreme heterogeneous setting ($\alpha = 0.1$).

4.5.1. Sensitivity analysis of the threshold τ

The soft clustering threshold τ in the MCTC scheme governs the connectivity and overlap among clients. As illustrated in Fig. 6, we evaluated the testing accuracy across varying thresholds τ in [0.05, 0.55] under both moderate ($\alpha = 1.0$) and extreme ($\alpha = 0.5$) heterogeneity. When τ is excessively low (e.g., $\tau \leq 0.10$), the clustering conditions are too

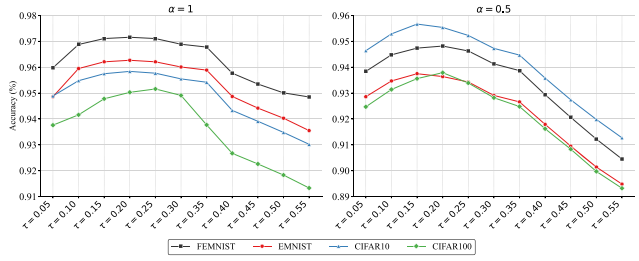


Fig. 6. Sensitivity analysis of the MCTC clustering threshold τ on model accuracy under different heterogeneity levels.

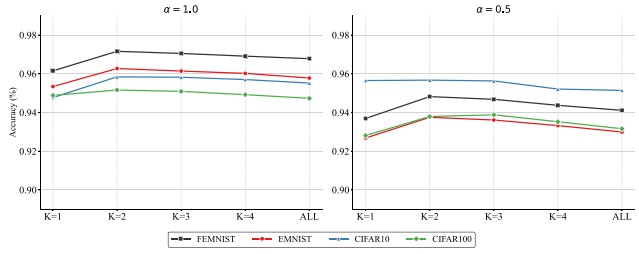


Fig. 7. Effect of the number of activated experts (K) on testing accuracy.

permissive, leading to the aggregation of clients with conflicting optimization directions, which induces negative transfer. Conversely, as τ increases beyond 0.30, the accuracy experiences a precipitous drop. This degradation occurs because an overly strict threshold isolates clients, degenerating the soft clustering mechanism into isolated local training and nullifying the benefits of cross-client knowledge sharing. Consequently, these empirical results highlight that there is no universal optimal threshold. Instead, the ideal τ is intrinsically tied to the dataset's complexity and the underlying data heterogeneity. Rather than relying on a static value, τ must be flexibly tuned according to the specific deployment scenario to secure the optimal trade-off between cross-client collaboration and local personalization.

4.5.2. Impact of the number of activated experts (K)

Fig. 7 presents the performance impact when scaling the number of activated experts, K, from 1 to the maximum available. Across all four datasets, selecting only the top-1 expert is insufficient to fully leverage cross-cluster diversity, resulting in sub-optimal accuracy. The performance achieves an optimal balance at $K = 2$. Crucially, incorporating additional experts ($K \geq 3$ or ALL) does not yield further improvements; rather, it not only inflates the computational overhead but also induces a slight performance degradation. Under highly heterogeneous settings, lower-ranked experts typically exhibit high predictive uncertainty and act as spurious noise sources for the current local input. By restricting the active subset to $K = 2$ and applying energy-based confidence calibration, the system effectively isolates and filters out this OOD noise. Consequently, the Top-2 strategy proves to be Pareto optimal, maximizing personalized accuracy while strictly bounding the computational inference overhead for our evaluated benchmarks. It is worth noting, however, that the ideal number of activated experts may naturally vary depending on the specific task complexity and the scale of data heterogeneity in other real-world deployments; thus, K can be flexibly adapted as a tunable hyperparameter in practice.

4.5.3. Robustness under ultra-extreme heterogeneity

To evaluate the robustness of the proposed framework under severe statistical heterogeneity, we conduct experiments in an ultra-extreme non-IID setting by reducing the Dirichlet parameter to $\alpha = 0.1$. In this highly skewed environment, each client's local dataset is typically dominated by only one or two classes, which induces drastic label distribution shifts and exposes the gating network to significant OOD uncertainty. As depicted in Fig. 8 for the challenging CIFAR-100 benchmark, transitioning the data distribution from moderate ($\alpha = 1.0$) to ultra-extreme ($\alpha = 0.1$) causes a precipitous performance drop in traditional aggregation methods like FedAvg and FedProx due to exacerbated client drift. Although MoE-based baselines (PM-MoE, PFMoE) partially mitigate this issue, they still suffer from steep degradation curves. Conversely, SCMoE-PFL exhibits remarkable resilience, sustaining an accuracy of nearly 90% even under $\alpha = 0.1$.

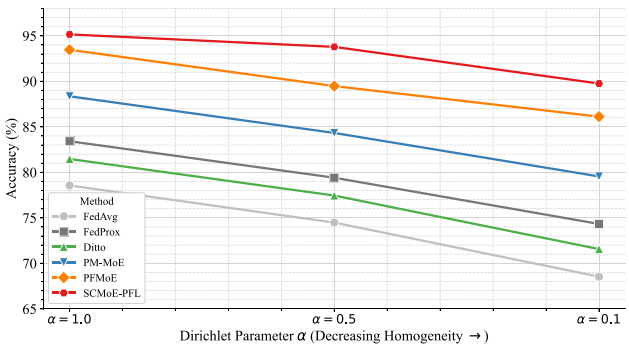


Fig. 8. Robustness evaluation under varying levels of data heterogeneity (α in {1.0, 0.5, 0.1}).

This robustness is inherently derived from our energy-guided fusion strategy detailed in Algorithm 2. In extremely heterogeneous environments, standard routing mechanisms easily misallocate high prior weights ($\alpha_{ik}^{\text{prior}}$) to irrelevant experts due to misleading feature overlaps. By introducing the Free Energy-based posterior confidence multiplier (β_{ik}), our energy-aware gating network dynamically calibrates these initial priors into uncertainty-aware weights ($\tilde{\alpha}_{ik}$). This mathematical dual-check mechanism effectively penalizes spurious expert representations that lack absolute predictive certainty. Coupled with the strict Top-2 expert selection, the system decisively truncates the impact of OOD noise, thereby guaranteeing highly reliable personalized predictions even under pathological data imbalances.

4.6. Comprehensive analysis

Extensive empirical evaluations across diverse image classification benchmarks and varying non-IID intensities demonstrate that SCMoE-PFL consistently establishes superior performance, thereby proving its effectiveness in addressing severe data heterogeneity. By leveraging the MCTC overlapping clustering scheme, the framework not only maximizes data utilization but also provides a rich, high-quality pool of candidate expert models to bridge the knowledge gap across disparate clients. Crucially, the intrinsic tension between personalization and generalization is alleviated by our energy-aware gating network. Rather than relying on naive feature matching, this mechanism introduces a Free Energy-based confidence calibration to dynamically quantify predictive uncertainty. By actively filtering out spurious experts and restricting fusion to the Top-2 most reliable models, it achieves a balance between high-fidelity personalization and inference efficiency.

Supported by the stable clustering topology of MCTC and this dual-check dynamic routing, SCMoE-PFL exhibits more consistent convergence trajectories and more balanced performance across the client cohort. Notably, under extreme statistical heterogeneity, the energy-guided mechanism effectively mitigates the impact of OOD noise. This translates into substantial system robustness in scenarios where traditional baselines suffer severe performance degradation. Ultimately, by synergizing soft clustering with uncertainty-aware MoE fusion, SCMoE-PFL provides a reliable, robust, and fair personalized federated learning paradigm, presenting a practical framework for complex real-world deployments with stringent privacy constraints.

5. Conclusion

This paper has presented a personalized federated learning framework, SCMoE-PFL, developed to address global model performance degradation and the difficulty of meeting personalization requirements under heterogeneous data distributions. By integrating MCTC soft clustering with energy-aware gating network-driven expert-model fusion, the framework has effectively balanced the generalization capability of shared models and the personalization needs of individual clients, while enhancing privacy protection. Experimental results have demonstrated that, compared with traditional methods such as FedAvg and FedProx, SCMoE-PFL has achieved substantial performance improvements across multiple datasets. In scenarios with extreme heterogeneity, our framework achieves a maximum accuracy improvement of 26.01 percentage points over the FedAvg baseline. Moreover, SCMoE-PFL has exhibited advantages in convergence speed, training stability, and fairness across clients.

Although notable progress has been achieved, further improvements remain possible in relation to computational overhead and applicability to resource-constrained devices. Future work will focus on enhancing computational efficiency to improve suitability for deployment on lightweight hardware. While this work focuses on clustering and MoE-based aggregation under strict privacy and zero-shot inference constraints, complementary paradigms such as knowledge distillation and meta-learning offer promising avenues for future integration within our framework. In conclusion, SCMoE-PFL has provided an efficient, adaptable, and privacy-preserving solution for personalized federated learning, particularly suited to practical environments characterized by severe data heterogeneity and stringent privacy requirements.

CRediT authorship contribution statement

Gongli Li: Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Xianzhong Jia:** Writing – original draft, Validation, Software, Methodology, Investigation, Data curation; **Weichen Liu:** Writing – original draft, Software, Investigation; **En Zhang:** Writing – original draft, Software, Investigation; **Zidong Wang:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Natural Science Foundation of Henan Province under Grant Number 252300421872.

References

[1] H. Cheng, Y. Qu, W. Liu, L. Gao, T. Zhu, Decentralized federated learning for private smart healthcare: a survey, Mathematics 13 (8) (2025). https://doi.org/10.3390/math13081296

[2] L. Kong, G. Zheng, A. Brintrup, A federated machine learning approach for order-level risk prediction in supply chain financing, Int. J. Prod. Econ. 268 (2024) 109095. https://www.sciencedirect.com/science/article/pii/S0925527323003274. https://doi.org/10.1016/j.ijpe.2023.109095

[3] W. Zhang, D. Deng, X. Wu, W. Zhao, Z. Liu, T. Zhang, J. Kang, D. Niyato, An adaptive asynchronous federated learning framework for heterogeneous internet of things, Inf. Sci. 689 (2025) 121458. https://www.sciencedirect.com/science/article/pii/S0020025524013720. https://doi.org/10.1016/j.ins.2024.121458

[4] M. Narula, J. Meena, D.K. Vishwakarma, A comprehensive review on federated learning for data-sensitive application: open issues & challenges, Eng. Appl. Artif. Intell. 133 (PB) (2024). https://doi.org/10.1016/j.engappai.2024.108128

[5] H. Zhu, J. Xu, S. Liu, Y. Jin, Federated learning on non-IID data: a survey, Neurocomputing 465 (2021) 371–390. https://www.sciencedirect.com/science/article/pii/S0925231221013254. https://doi.org/10.1016/j.neucom.2021.07.098

[6] F. Vieira, C.A.V. Campos, Reducing weight divergence impact using local learning normalization in federated learning for heterogeneous data distributions, Future Gener. Comput. Syst. 173 (2025) 107881. https://www.sciencedirect.com/science/article/pii/S0167739X25001761. https://doi.org/10.1016/j.future.2025.107881

[7] Q. Yin, Z. Feng, X. Li, S. Chen, H. Wu, G. Han, Tackling data-heterogeneity variations in federated learning via adaptive aggregate weights, Knowl.-Based Syst. 304 (2024) 112484. https://www.sciencedirect.com/science/article/pii/S0950705124011183. https://doi.org/10.1016/j.knosys.2024.112484

[8] H. Zhang, Q. Su, PJPFL: personalized federated learning with privacy preservation based on sample similarity, Inf. Fusion 122 (2025) 103221. https://www.sciencedirect.com/science/article/pii/S1566253525002945. https://doi.org/10.1016/j.inffus.2025.103221

[9] F. Sabah, Y. Chen, Z. Yang, M. Azam, N. Ahmad, R. Sarwar, Model optimization techniques in personalized federated learning: a survey, Expert Syst. Appl. 243 (2024) 122874. https://www.sciencedirect.com/science/article/pii/S0957417423033766. https://doi.org/10.1016/j.eswa.2023.122874

[10] M. Ren, Z. Wang, X. Yu, Personalized federated learning: a clustered distributed co-meta-learning approach, Inf. Sci. 647 (2023) 119499. https://www.sciencedirect.com/science/article/pii/S0020025523010848. https://doi.org/10.1016/j.ins.2023.119499

[11] Y. Cao, J. Ma, Z. He, Y. Li, AP-CFL: clustered federated learning through dynamic clustering and adaptive participation in heterogeneous IoT, IEEE Internet Things J. 12 (10) (2025) 13671–13682. https://doi.org/10.1109/JIOT.2025.3528624

[12] Q. Wang, Q. Li, B. Guo, J. Cui, Efficient federated learning with smooth aggregation for non-IID data from multiple edges, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 9006–9010. https://doi.org/10.1109/ICASSP48485.2024.10447506

[13] C. Liu, D.M. Alghazzawi, L. Cheng, G. Liu, C. Wang, C. Zeng, Y. Yang, Disentangling client contributions: improving federated learning accuracy in the presence of heterogeneous data, in: 2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 2023, pp. 381–387. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom59178.2023.00082

[14] J. Li, X. Liu, T. Mahmoodi, Federated learning in heterogeneous wireless networks with adaptive mixing aggregation and computation reduction, IEEE Open J. Commun. Soc. 5 (2024) 2164–2182. https://doi.org/10.1109/OJCOMS.2024.3381545

[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A.y. Arcas, Communication-efficient learning of deep networks from decentralized data, in: A. Singh, J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html.

[16] J. Wang, M.T. Quasim, B. Yi, Privacy-preserving heterogeneous multi-modal sensor data fusion via federated learning for smart healthcare, Inf. Fusion 120 (2025) 103084. https://www.sciencedirect.com/science/article/pii/S1566253525001575. https://doi.org/10.1016/j.inffus.2025.103084

[17] M. Li, P. Xu, J. Hu, Z. Tang, G. Yang, From challenges and pitfalls to recommendations and opportunities: implementing federated learning in healthcare, Med. Image Anal. 101 (2025) 103497. https://www.sciencedirect.com/science/article/pii/S1361841525000453. https://doi.org/10.1016/j.media.2025.103497

[18] Z.L. Teo, L. Jin, N. Liu, S. Li, D. Miao, X. Zhang, W.Y. Ng, T.F. Tan, D.M. Lee, K.J. Chua, J. Heng, Y. Liu, R.S.M. Goh, D.S.W. Ting, Federated machine learning in healthcare: a systematic review on clinical applications and technical architecture, Cell Rep. Med. 5 (2) (2024) 101419. https://www.sciencedirect.com/science/article/pii/S2666379124000429. https://doi.org/10.1016/j.xcrm.2024.101419

[19] Z. Xia, B. Zhong, T. Zhao, K. Li, S. Zhang, Federated learning system eliminating model drift in distributed edge computing: theoretical analytics and application on pit engineering state monitoring, Adv. Eng. Inform. 66 (2025) 103505. https://www.sciencedirect.com/science/article/pii/S1474034625003982. https://doi.org/10.1016/j.aei.2025.103505

[20] X. Chen, G. Xu, X. Xu, H. Jiang, Z. Tian, T. Ma, Multicenter hierarchical federated learning with fault-tolerance mechanisms for resilient edge computing networks, IEEE Trans. Neural Netw. Learn. Syst. 36 (1) (2025) 47–61. https://doi.org/10.1109/TNNLS.2024.3362974

[21] Y. Li, X. Wang, R. Zeng, P. Kumar Donta, I. Murturi, M. Huang, S. Dustdar, Federated domain generalization: a survey, Proc. IEEE 113 (4) (2025) 370–410. https://doi.org/10.1109/JPROC.2025.3596173

[22] Y. Deng, A. Wang, L. Zhang, Y. Lei, B. Li, Y. Li, FedRFC: federated learning with recursive fuzzy clustering for improved non-IID data training, Future Gener. Comput. Syst. 160 (2024) 835–843. https://www.sciencedirect.com/science/article/pii/S0167739X24003509. https://doi.org/10.1016/j.future.2024.06.049

[23] K. Hu, S. Gong, Q. Zhang, S. Chaowen, M. Xia, S. Jiang, An overview of implementing security and privacy in federated learning, Artif. Intell. Rev. 57 (8) (2024) 204. https://doi.org/10.1007/s10462-024-10846-8

[24] X. Wang, S. Garg, H. Lin, J. Hu, G. Kaddoum, M. Jalil Piran, M.S. Hossain, Toward accurate anomaly detection in industrial internet of things using hierarchical federated learning, IEEE Internet Things J. 9 (10) (2022) 7110–7119. https://doi.org/10.1109/JIOT.2021.3074382

[25] X. Wang, B. Ye, L. Xu, L. Wu, S.-Y. Hsieh, J. Wu, L. Lin, FedHAN: a cache-based semi-asynchronous federated learning framework defending against poisoning attacks in

heterogeneous clients, in: J. Kwok (Ed.), Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 3407–3416. Main Track, https://doi.org/10.24963/ijcai.2025/379

[26] Z. Qu, Y. Tang, G. Muhammad, P. Tiwari, Privacy protection in intelligent vehicle networking: a novel federated learning algorithm based on information fusion, Inf. Fusion 98 (2023) 101824. https://www.sciencedirect.com/science/article/pii/S1566253523001409. https://doi.org/10.1016/j.inffus.2023.101824

[27] X. Tang, Y. Wang, X. Liu, X. Yuan, C. Fan, Y. Hu, Q. Miao, Federated graph neural network for privacy-preserved supply chain data sharing, Appl. Soft Comput. 168 (C) (2025). https://doi.org/10.1016/j.asoc.2024.112475

[28] F. Sabah, Y. Chen, Z. Yang, A. Raheem, M. Azam, N. Ahmad, R. Sarwar, Communication optimization techniques in personalized federated learning: applications, challenges and future directions, Inf. Fusion 117 (2025) 102834. https://www.sciencedirect.com/science/article/pii/S1566253524006122. https://doi.org/10.1016/j.inffus.2024.102834

[29] W. Yang, Y. Yang, Y. Xi, H. Zhang, W. Xiang, FLCP: federated learning framework with communication-efficient and privacy-preserving, Appl. Intell. 54 (9–10) (2024) 6816–6835. https://doi.org/10.1007/s10489-024-05521-y

[30] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: I. Dhillon, D. Papailiopoulos, V. Sze (Eds.), Proceedings of Machine Learning and Systems, 2, 2020, pp. 429–450. https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf.

[31] X. Yang, W. Huang, M. Ye, FedAS: bridging inconsistency in personalized federated learning, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11986–11995. https://doi.org/10.1109/CVPR52733.2024.01139

[32] F. Sabah, Y. Chen, Z. Yang, A. Raheem, M. Azam, N. Ahmad, R. Sarwar, FairDPFL-SCS: fair dynamic personalized federated learning with strategic client selection for improved accuracy and fairness, Inf. Fusion 115 (2025) 102756. https://www.sciencedirect.com/science/article/pii/S1566253524005347. https://doi.org/10.1016/j.inffus.2024.102756

[33] Z. Lu, H. Pan, Y. Dai, X. Si, Y. Zhang, Federated learning with non-IID data: a survey, IEEE Internet Things J. 11 (11) (2024) 19188–19209. https://doi.org/10.1109/JIOT.2024.3376548

[34] X. Li, H. Zhao, W. Deng, IOFL: intelligent-optimization-based federated learning for non-IID data, IEEE Internet Things J. 11 (9) (2024) 16693–16699. https://doi.org/10.1109/JIOT.2024.3354942

[35] Y. Cong, Y. Zeng, J. Qiu, Z. Fang, L. Zhang, D. Cheng, J. Liu, Z. Tian, FedGA: a greedy approach to enhance federated learning with non-IID data, Knowl.-Based Syst. 301 (2024) 112201. https://www.sciencedirect.com/science/article/pii/S0950705124008359. https://doi.org/10.1016/j.knosys.2024.112201

[36] D. Zeng, X. Hu, S. Liu, Y. Yu, Q. Wang, Z. Xu, StoCFL: a stochastically clustered federated learning framework for non-IID data with dynamic client participation, Neural Netw. 187 (2025) 107278. https://www.sciencedirect.com/science/article/pii/S0893608025001571. https://doi.org/10.1016/j.neunet.2025.107278

[37] Z. Wei, J. Wang, Z. Zhao, K. Shi, Toward data efficient anomaly detection in heterogeneous edge–cloud environments using clustered federated learning, Future Gener. Comput. Syst. 164 (2025) 107559. https://www.sciencedirect.com/science/article/pii/S0167739X24005235. https://doi.org/10.1016/j.future.2024.107559

[38] L. Liu, J. Li, J. Wang, CFL-ICCV: clustered federated learning framework with an intra-cluster cross-validation mechanism for DER forecasting, Appl. Energy 377 (2025) 124699. https://www.sciencedirect.com/science/article/pii/S0306261924020828. https://doi.org/10.1016/j.apenergy.2024.124699

[39] R. Du, S. Xu, R. Zhang, L. Xu, H. Xia, A dynamic adaptive iterative clustered federated learning scheme, Knowl.-Based Syst. 276 (2023) 110741. https://www.sciencedirect.com/science/article/pii/S0950705123004914. https://doi.org/10.1016/j.knosys.2023.110741

[40] E. Yoo, H. Ko, S. Pack, Fuzzy clustered federated learning algorithm for solar power generation forecasting, IEEE Trans. Emerg. Top. Comput. 10 (4) (2022) 2092–2098. https://doi.org/10.1109/TETC.2022.3142886

[41] Y. Cheng, W. Zhang, Z. Zhang, J. Kang, Q. Xu, S. Wang, D. Niyato, SnapCFL: a pre-clustering-based clustered federated learning framework for data and system heterogeneities, IEEE Trans. Mob. Comput. 24 (6) (2025) 5214–5228. https://doi.org/10.1109/TMC.2025.3529487

[42] Y. Xu, Y. Tan, C. Zhang, P. Sun, Y. Zhang, J. Ren, H. Jiang, Y. Zhang, Towards privacy-enhanced and robust clustered federated learning, IEEE Trans. Mob. Comput. 24 (8) (2025) 7171–7188. https://doi.org/10.1109/TMC.2025.3547149

[43] P. Qi, D. Chiaro, A. Guzzo, M. Ianni, G. Fortino, F. Piccialli, Model aggregation techniques in federated learning: a comprehensive survey, Future Gener. Comput. Syst. 150 (2024) 272–293. https://www.sciencedirect.com/science/article/pii/S0167739X23003333. https://doi.org/10.1016/j.future.2023.09.008

[44] H. Imani, J. Anderson, S. Farid, A. Amirany, T. El-Ghazawi, RLFL: a reinforcement learning aggregation approach for hybrid federated learning systems using full and ternary precision, IEEE J. Emerg. Sel. Top. Circuits Syst. 14 (4) (2024) 673–687. https://doi.org/10.1109/JETCAS.2024.3483554

[45] K. Jin, L. Xu, X. Wang, S.-Y. Hsieh, J. Wu, L. Lin, FedCPD:personalized federated learning with prototype-enhanced representation and memory distillation, in: J. Kwok (Ed.), Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 5498–5507. Main Track, https://doi.org/10.24963/ijcai.2025/612

[46] C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: J. Van den Bussche, V. Vianu (Eds.), Database Theory — ICDT 2001, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 420–434.

[47] T. Kurita, Principal component analysis (PCA), in: Computer Vision: A Reference Guide, Springer, 2021, pp. 1013–1016.

[48] W. Liu, X. Wang, J.D. Owens, Y. Li, Energy-based out-of-distribution detection, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[49] S. Caldas, S.M.K. Duddu, P. Wu, T. Li, J. Konečnỳ, H.B. McMahan, V. Smith, A. Talwalkar, Leaf: A benchmark for federated settings, (2018). https://doi.org/10.48550/arXiv.1812.01097

[50] G. Cohen, S. Afshar, J. Tapson, A. van Schaik, EMNIST: extending MNIST to handwritten letters, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2921–2926. https://doi.org/10.1109/IJCNN.2017.7966217

[51] A. Krizhevsky, Learning multiple layers of features from tiny images, 2009. https://api.semanticscholar.org/CorpusID:18268744.

[52] T. Li, S. Hu, A. Beirami, V. Smith, Ditto: fair and robust federated learning through personalization, in: International Conference on Machine Learning, PMLR, 2021, pp. 6357–6368. https://doi.org/10.48550/arXiv.2012.04221

[53] Y. Feng, Y.-a. Geng, Y. Zhu, Z. Han, X. Yu, K. Xue, H. Luo, M. Sun, G. Zhang, M. Song, PM-MOE: mixture of experts on private model parameters for personalized federated learning, in: Proceedings of the ACM on Web Conference 2025, WWW '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 134–146. https://doi.org/10.1145/3696410.3714561

[54] Y. Zhuang, Y. Li, Y. Song, M. Qiu, Personalized federated learning for fault diagnosis with mixture of experts, Inf. Fusion 125 (2026) 103439. https://www.sciencedirect.com/science/article/pii/S1566253525005123. https://doi.org/10.1016/j.inffus.2025.103439

[55] J. Scott, C.H. Lampert, D. Saulpic, Differentially private federated K-means clustering with server-side data, in: Forty-second International Conference on Machine Learning, 2025. https://openreview.net/forum?id=EFLPHl5RGJ.

[56] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65. https://www.sciencedirect.com/science/article/pii/0377042787901257. https://doi.org/10.1016/0377-0427(87)90125-7

[57] B. Gong, T. Xing, Z. Liu, W. Xi, X. Chen, Towards hierarchical clustered federated learning with model stability on mobile devices, IEEE Trans. Mob. Comput. 23 (6) (2024) 7148–7164. https://doi.org/10.1109/TMC.2023.3332637