Ensemble Federated Learning With Non-IID Data in Wireless Networks

Zhongyuan Zhao, *Member, IEEE*, Jingyi Wang, Wei Hong, Tony Q. S. Quek, *Fellow, IEEE*, Zhiguo Ding, *Fellow, IEEE*, and Mugen Peng, *Fellow, IEEE*

***Abstract*— Federated learning is a promising technique to implement network intelligence for the sixth generation (6G) communication systems. However, the collected data in wireless networks is non-independent and identically distributed (non-IID), which leads to severe deterioration of model performance. Although various enhanced schemes are proposed, it is still challenging to balance the communication cost and the model performance, due to the scarcity of radio resource for model update in wireless networks. In this paper, an ensemble federated learning paradigm is proposed for handling non-IID data, which is also optimized for its deployment in wireless networks in a cost efficient way. First, the framework of ensemble federated learning is designed. By formulating individual user clusters, intra-cluster federated learning models can be generated to reduce the impact of non-IID data, which can be integrated to adapt to various learning data via model ensemble. Second, the optimization of user cluster formation is studied to improve the performance of ensemble federated learning, which is modeled as a coalition formation game to design a Nash-stable algorithm. Finally, the simulation results on the public data sets are provided to verify the performance gains of our proposed schemes for deploying federated learning with non-IID data in wireless networks.**

***Index Terms*— Network intelligence, federated learning, model ensemble, non-IID data, coalition formation game.**

Manuscript received 6 August 2022; revised 12 March 2023 and 20 June 2023; accepted 21 August 2023. Date of publication 5 September 2023; date of current version 11 April 2024. The work of Zhongyuan Zhao was supported in part by the National Natural Science Foundation under Grant 61971061, in part by the Beijing Natural Science Foundation under Grant L223026, and in part by the 5G Evolution Wireless Air Interface Intelligent Research and Development and Verification Public Platform Project under Grant 2022-229-220. The work of Tony Q. S. Quek was supported in part by the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research and Development Program. The associate editor coordinating the review of this article and approving it for publication was C. Huang. *(Zhongyuan Zhao and Jingyi Wang are co-first authors.) (Corresponding author: Zhongyuan Zhao.)*

Zhongyuan Zhao and Mugen Peng are with the State Key Laboratory of Networking and Switching Technology, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zyzhao@bupt.edu.cn; pmg@bupt.edu.cn).

Jingyi Wang is with the China Telecom Research Institute, Beijing 102209, China (e-mail: wangjy74@chinatelecom.cn).

Wei Hong is with Beijing Xiaomi Mobile Software, Beijing 102628, China (e-mail: hongwei@xiaomi.com).

Tony Q. S. Quek is with the Department of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372 (e-mail: tonyquek@sutd.edu.sg).

Zhiguo Ding is with the Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates (e-mail: zhiguo.ding@manchester.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2023.3309376.

Digital Object Identifier 10.1109/TWC.2023.3309376

I. Introduction

ARTIFICIAL intelligence (AI) has been considered as an attracting feature of the future sixth generation (6G) systems, which is one of the most important enabling techniques, and can widely extend its application scenarios [1], [2]. Therefore, the concept of network intelligence is emerging, which has become a research hotspot in the fields of academia and industry [3], [4]. It requires to fully integrate AI techniques into wireless signal processing and network management of 6G, rather than acquire learning models and inference results externally via extra interactions. Recently, network intelligentization is extending from the core networks to the network edge devices [5], [6], [7], which can support flexible deployment and orchestration of deep learning/machine learning models to satisfy ultimate and diverse quality-of-service (QoS) requirements of future 6G applications.

Unlike the conventional research areas with respect to AI, the data resource and the computation capability are widely dispersed in wireless networks, and thus the distributed learning paradigms should be employed to fully explore the potential of 6G systems. In [8], [9], and [10], federated learning has been proposed as a novel model-level collaborated learning paradigm. By aggregating the learning results of all the participators, federated learning can generate high quality global learning models, which allows each client to keep its own data privately. Due to the advantages of deployment flexibility and user privacy protection, federated learning provides a promising solution of implementing network edge intelligence for the future 6G systems [11]. An intelligent fog computing-based radio access network has been designed in [12] by supporting hierarchical federated learning among the users, the fog nodes, and the cloud computing centers, which can provide a potential evolution path of future 6G network architectures.

However, the deployment of federated learning in wireless networks still faces its unique challenges [13], [14], [15]. In particular, the collected network data samples with respect to network status and user behavior, are usually non-independent and identically distributed (non-IID) and unbalanced. In [16], the experiment results show that the accuracy performance of federated learning can be decreased even more than 11% when the non-IID data is employed for simple image classification tasks. Moreover, the exchanging of model update results occupies a huge amount of radio resource, which puts a heavy burden on wireless networks.

Therefore, it is critical to mitigate the impact of non-IID data in a communication-efficient way.

A. Related Works With Respect to Federated Learning With Non-IID Data

The performance loss of federated learning caused by non-IID data can be alleviated by sophisticated design of data set and learning model enhancement. In particular, the data enhancement schemes aim to reduce the distribution divergence by manipulating the training data samples. Moreover, the model enhancement schemes cope with non-IID data by rethinking the paradigms of federated learning, which can be jointly employed with the data enhanced schemes without any modification. To provide a comprehensive literature review, the related works of both the data and the model enhanced schemes are introduced as follows.

1) Data Enhanced Schemes: A straightforward strategy is to share a data set among all the clients. In [17], a common data set is shared by the server before the training procedure of federated learning to harmonize the data distribution divergence. In [18], a global IID data set construction scheme has been proposed by collecting non-IID data samples from a limited number of clients. It takes a great risk of privacy leakage for sharing client data samples, especially in wireless circumstances. To solve this problem, data augmentation has been considered as a promising alternative method. Its key idea is to imitate data based on the real data by using generative learning models. In [19] and [20], data augmentation for federated learning has been studied. In [21], a data distribution information-based augmentation scheme without seed sample collection has been proposed.

2) Model Enhanced Schemes: Based on the global averaging results of federated learning, personalized learning models can be generated by employing model enhanced schemes, which aim to adapt to diverse non-IID learning tasks of clients. Inspired by multi-task learning, an enhanced federated learning paradigm, named FedPer, has been proposed in [22], in which the federated learning model can be adapted to the individual learning task of each client via local adaption. In [23], a similar scheme has been proposed to improve the performance by designing the local personalized layers from a representation learning perspective.

Moreover, the impact of non-IID data can be mitigated by sophisticatedly managing the participators of federated learning. In [24], a cluster-based federated learning paradigm has been proposed for multi-task learning scenarios, which can reduce the performance loss caused by the data distribution divergence. In [25], a multi-center federated learning scheme, named FeSEM, has been proposed to generate multiple global models for handling diverse learning tasks of the clients. Similarly, a group-based federated learning (FedGroup) scheme has been designed in [26], where the clients with statistically correlated data are divided into a common group to implement federated learning. In [29], the paradigm of cluster federated learning is deployed in wireless networks via over-the-air computation, which provides a promising method to make full use of disperse computation resource at the edge of networks. In [30], a dynamic clustering scheme with power control is proposed to improve the performance of cluster federated learning with restricted radio resource in wireless networks.

Another method of federated learning personalization is to update the global model parameters by using the local data sets directly at the clients, which is similar to the fine-tune procedure of meta-learning. In [31], a federated meta-learning (FedMeta) framework has been designed. In particular, the non-IID data set of each client is treated as a separated learning task, which can be used for generating a personalized learning model based on the federated meta-learner. To quickly adapt to various non-IID data of clients, a new model training and aggregation scheme, Per-FedAvg, has been proposed in [32], where the loss function is redesigned to capture the data distribution differences of clients.

Besides the studies of framework design for federated learning, the gradient descent-based algorithms are also enhanced to mitigate the impact of non-IID data. In [33], the loss function is modified by using a proximal term to restrict the inexactness caused by non-IID data, and an algorithm named FedProx has been proposed. By adding a correlation term into the loss function of FedProx, an enhanced scheme, named FedDane, has been proposed in [34] to speed up the convergence of federated learning with non-IID data.

B. Motivations and Contributions of This Paper

Although the existing works can mitigate the impact of non-IID data on federated learning, there still face several critical issues, especially in wireless networks: *First*, the data enhanced schemes require the exchange of data samples/distribution information between the server and the clients via wireless channels. It violates the privacy protection principle of federated learning, and thus cannot be applied in wireless networks. *Second*, the model enhanced schemes require sophisticated design of model structures/training strategies, which may occupy extra communication and computation resources. *Finally*, as the scale and the distribution divergence of non-IID data increase, the communication and computation costs of existing schemes also increase, while the performance cannot be guaranteed.

Motivated by achieving better tradeoff between the performance and the cost of federated learning with non-IID data in wireless networks, a model ensemble-enabled paradigm, named ensemble federated learning, is proposed in this paper, and our main contributions can be summarized as follows:

- *First*, a new paradigm, named ensemble federated learning, is proposed to combat with non-IID data. In particular, individual clusters are firstly formulated by all the participators to reduce the distribution divergence of non-IID data, and thus intra-cluster federated learning can be deployed among the users with similar data distributions. Then, a final global model can be generated by integrating all the intra-cluster federated learning models via model ensemble, which can be adaptive to various learning data. Except the operations of formulating user clusters at the base station, our proposed scheme does not cause other extra communication or computation cost.

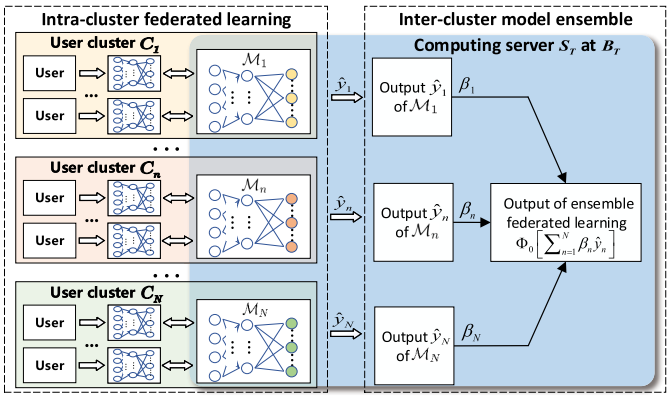


Fig. 1. The framework of ensemble federated learning scheme.

- *Second*, the user cluster formation is studied to fully explore the potential of ensemble federated learning. It requires to keep a balance between the intra-cluster and the inter-cluster data distribution divergence, so that both the convergence of intra-cluster federated learning models and the adaptability of ensemble model can be guaranteed. Based on coalition formation game theory, a switching-based coalition formation algorithm for user clustering is designed to ensure that our proposed ensemble federated learning paradigm can achieve considerable performance with low cost and high transmission reliability in wireless networks.
- *Finally*, our proposed schemes are evaluated by employing classic image classification tasks on both the MNIST and FEMNIST data sets. The simulation results show that our proposed scheme can achieve better performance than several typical schemes with non-IID data cases.

The rest of this paper is organized as follows. Section II introduces the paradigm of ensemble federated learning. In Section III, a coalition formation game-based user cluster formation algorithm is proposed, and the proof of algorithm stability is also provided. The simulation results are shown in Section IV, followed by the conclusions in Section V. A summary of the important notions used throughout this paper is provided in Table I for better understand.

II. The Framework of Ensemble Federated Learning

Consider the deployment of federated learning between a base station B_T and M wireless users $U_1, \ldots, U_M$. A computing server S_T is equipped with B_T, while a computing unit S_i and a local data set $\mathcal{D}_i$ are with U_i to enable local data processing and model training. In this paper, we assume that the local data sets are non-IID, i.e., $\mathcal{D}_i$ can be generated by uniformly sampling based on a given distribution $P(\mathbf{X}_i)$, $P(\mathbf{X}_i) \neq P(\mathbf{X}_j)$, $i \neq j$.

In the classic federated learning schemes, it requires to obtain a global model without model personalization, which has to support all the potential target inferring data sets for model inference. Therefore, unlike the individual learning mentioned by the reviewer, it cannot provide personalized learning model for each user for inferring its own data. Moreover, the performance loss caused by distribution divergence of non-IID user data is a unique feature of federated learning [15], since the user data cannot be directly aggregated, and the distribution divergence cannot be harmonized by linear federated averaging. The main motivation of this paper is to design an effective federated learning paradigm via model ensemble, and mitigate the performance loss caused by non-IID data.

To mitigate the impact of non-IID data, an ensemble federated learning paradigm is proposed. The key idea is to divide all the participants into individual user clusters based on their data distributions, which encourages the users with similar training data set to generate a common model collaboratively via intra-cluster federated learning. Due to the distribution divergence, the achieved learning model of each cluster is not adaptive to all the categories of learning data. Therefore, to improve the generalization performance, all the generated learning models are integrated by model ensemble.

A. User Cluster Formation for Ensemble Federated Learning

To mitigate the impact of non-IID data, individual user clusters are formulated, which can reduce the data distribution divergence by managing the participation scale and manipulating the aggregated distribution of each federated learning cluster. To guarantee the convergence and the stability of the framework, the cluster formations are firstly fixed at the beginning of ensemble federated learning, which are not changed during the following parameter update procedure.

In this paper, the local model parameters are updated by employing stochastic gradient descent-based (SGD) method with the mini-batch strategy. In particular, the update result of U_i for the t-th round of iteration can be obtained based on the gradients of E local batches, which can be expressed by following [51]:

$$\nabla \bar{F}(\mathbf{w}_t^i; \mathcal{D}_i) = \sum_{e=0}^{E-1} \nabla \hat{F}(\mathbf{w}_{t,e}^i; \mathcal{S}_{t,e}^i), \ t = 0, 1, 2, \ldots, \quad (1)$$

where $\nabla \hat{F}(\mathbf{w}_{t,e}^i; \mathcal{S}_{t,e}^i)$ denotes the gradients of empirical risk with respect to the e-th data batch $\mathcal{S}_{t,e}^i$ during the t-th round, and $\mathbf{w}_{t,e}^i$ is defined similarly for its model parameters.

The detailed design of cluster formation for ensemble federated learning is studied in Section III, which aims to partition the users based on the data distribution similarity with considering the energy consumption and the unreliability of communications. To provide some useful insights, the characterization of data distribution similarity is first studied in this part.

A Data Distribution Similarity Metric Based on Cosine Similarity: As shown in Algorithm 1, the user clusters are formulated in the first round of iteration. The gradients, i.e., $\nabla \bar{F}(\mathbf{w}_0; \mathcal{D}_i)$ by setting $t = 0$ and $\mathbf{w}_0^i = \mathbf{w}_0$ in (1), can be employed to capture the key features of data distribution, since they are mainly determined by the local training data. As introduced in [26], the cosine similarity of different users can be employed as an effective metric to characterize the similarity of data distributions, which is defined as cosines

TABLE I
SUMMARY OF NOTATIONS

Notations	Explanations
$U_1, \ldots, U_M$	M users in the system, indexed by i and i_n
$C_1, \ldots, C_N$	N user clusters formulated, indexed by n
$\mathcal{D}_i$	local training data set of U_i
$\mathbf{w}_0$	initialized model parameters for all the users
$\mathcal{S}_{t,e}^i$	e-th training data batch of $\mathcal{D}_i$ during the t-th round
$\mathbf{w}_{t,e}^i$	local model parameters of U_i w.r.t. the e-th training data batch of $\mathcal{D}_i$ during the t-th round
$\mathbf{w}_t^i$	initial local model parameters of U_i during the t-th round, $\mathbf{w}_t^i = \mathbf{w}_{t,0}^i$
$\nabla \bar{F}(\mathbf{w}_t^i; \mathcal{D}_i)$	update result of U_i for the t-th round based on the gradients of E local batches
$\lambda_{i,j}$	cosine similarity w.r.t. gradients between U_i and U_j
$\bar{\mathbf{w}}_{n,t}$	global model parameters of cluster C_n in the t-th round
$\mathbf{x}$, $\mathbf{x}_T$	data sample for model training, a set of data samples for model inference
$\mathcal{M}_n$, $\bar{\mathcal{M}}$	intra-cluster federated learning model of C_n, ensemble model
y_n, $\bar{y}$	output result of $\mathcal{M}_n$, output result of $\bar{\mathcal{M}}$
θ_n	weight of $\mathcal{M}_n$ for model ensemble
$\varphi_i(C_n, \Omega_l)$	utility function of U_i in C_n with a partition result Ω_l

with respect to their initial gradients, i.e., the cosine similarity of U_i and U_j can be expressed as follows in this paper:

$$\lambda_{i,j} = \frac{\nabla \bar{F}\left(\mathbf{w}_0; \mathcal{D}_i\right)^\top \nabla \bar{F}\left(\mathbf{w}_0; \mathcal{D}_j\right)}{\left\|\nabla \bar{F}\left(\mathbf{w}_0; \mathcal{D}_i\right)\right\| \left\|\nabla \bar{F}\left(\mathbf{w}_0; \mathcal{D}_j\right)\right\|}, \tag{2}$$

where $\nabla \bar{F}\left(\mathbf{w}_0; \mathcal{D}_i\right)$ and $\nabla \bar{F}\left(\mathbf{w}_0; \mathcal{D}_j\right)$ can be derived by following (1). As shown in (2), the data distributions of U_i and U_j are similar with each other when $\lambda_{i,j}$ is large. In particular, their data distributions are identical when $\lambda_{i,j} = 1$.

In our proposed scheme, the learning results of users in the same cluster are updated towards a common target with similar descent rates, which can guarantee that the aggregated intra-cluster model can decrease consistently. Therefore, by following [24], the model initialization has to be identical to align the starting points of users, i.e., $\mathbf{w}_{0,0}^i = \mathbf{w}_0$, $i = 1, \cdots, M$ in Algorithm 1.

B. Intra-Cluster Federated Learning

After user cluster formation, the users with similar data distributions can train learning models collaboratively via intra-cluster federated learning. Without loss of generality, we focus on a specific user cluster C_n. During the t-th round of model parameter update, the update result of U_{i_n}, which can be denoted as $\nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$, is generated based on (1), $U_{i_n} \in C_n$. Then, each user transmits its updated gradient $\nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$ to the computing server at the base station via the wireless channel. Due to the employment of orthogonal frequency division multiple access (OFDMA) technique, all the users communicate with the base station by using orthogonal channels. Please note that the required data rate and bandwidth of our studied scenario are quite lower than the multimedia data services, which do not exceed the performance capability of existing wireless networks. The orthogonal channel allocation assumption can be satisfied.To satisfy the transmit power constraint [27], [28], $\nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$ is scale normalized, and the observation at B_T can be expressed as

$$\mathbf{z}_{i_n,t} = h_{i_n,t} d_{i_n}^{-\alpha/2} \sqrt{P} \nabla \tilde{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n}) + \mathbf{n}_{i_n}, \tag{3}$$

where $\mathbf{z}_{i_n,t}$ denotes the received gradient at B_T, $h_{i_n,t}$ captures the flat channel fading of the wireless link between U_{i_n} and B_T, d_{i_n} denotes the distance between U_{i_n} and B_T, α is the path loss exponent, P denotes the transmit power of U_{i_n}, $\nabla \tilde{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$ denotes $\nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$ with scale normalization, i.e., $\nabla \tilde{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n}) = \nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n}) / \|\nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})\|$, and $\mathbf{n}_{i_n}$ is the additive white Gaussian noise. Then, $\nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$ can be recovered at B_T based on $\nabla \tilde{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$ and its scale. To guarantee the transmission reliability of $\nabla \tilde{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n})$, the retransmission can be required if it is not successfully recovered at B_T.

After recovering the feedback gradients of the users, the global model of C_n can be updated by employing intra-cluster federated learning at B_T, i.e., the global model parameters of C_n for the t-th round can be expressed as follows:

$$\begin{aligned} \bar{\mathbf{w}}_{n,t+1} &= \bar{\mathbf{w}}_{n,t} - \eta \nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_t\right) \\ &= \bar{\mathbf{w}}_{n,t} - \eta \sum_{U_{i_n} \in C_n} p_{i_n} \nabla \bar{F}(\mathbf{w}_t^{i_n}; \mathcal{D}_{i_n}), \end{aligned} \tag{4}$$

where $\bar{\mathbf{w}}_{n,t}$ and $\bar{\mathbf{w}}_{n,t+1}$ denote the global model parameters of C_n for the t-th and the $(t+1)$-th rounds, respectively, $\nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_t\right)$ is the gradient of global averaged empirical risks of C_n for the t-th round, and p_{i_n} denotes the weight of U_{i_n} with respect to data volume, i.e., $p_{i_n} = \frac{L_{i_n}}{\sum_{U_{i_n} \in C_n} L_{i_n}}$, in which L_{i_n} denotes the number of training data samples in $\mathcal{D}_{i_n}$ [8]. $\bar{\mathbf{w}}_{n,t+1}$ will be broadcast back to all the users in C_n for the next round of model update i.e., $\mathbf{w}_{t+1,0}^{i_n} = \bar{\mathbf{w}}_{n,t+1}$.

The global model of C_n is finally generated via iterative global averaging based on (4) until it converges, i.e., $\|\nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_t\right) - \nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_{t-1}\right)\| < \epsilon$, and ϵ is a fixed threshold. Therefore, N sets of individual model parameters can be obtained after intra-cluster federated learning, where N denotes the number of user clusters.

C. Inter-Cluster Model Ensemble for Model Inference

Due to the existence of distribution divergence with the global data distribution, the accuracy performance of each intra-cluster federated learning model with respect to some

specific categories of learning data cannot be guaranteed. To provide a generalized global model for all the potential target inferring data sets, model ensemble is an efficient strategy, which can reduce the performance loss caused by data distribution divergence in federated learning. The key idea of model ensemble is to add several different models using weights and create a combined model that is better than all of its constituents individually.

In this paper, the neural network-based learning models can be employed, which consist of an input layer, multiple hidden layers, and an output layer. Without loss of generality, we focus on the generated learning model of cluster C_n, which can be denoted as $\mathcal{M}_n$. For a specific target inferring data $\mathbf{x}_T$, the output result of $\mathcal{M}_n$ can be expressed as

$$y_n = \Phi_{\mathrm{o}}\left(\hat{\mathbf{y}}_n\right) = \Phi_{\mathrm{o}}\left[\Phi_{\mathrm{h}}\left(\mathbf{x}_T; \bar{\mathbf{w}}_n\right)\right], \tag{5}$$

where y_n denotes the final output result of $\mathcal{M}_n$ and is employed to evaluate the model inference performance, $\Phi_{\mathrm{o}}(\cdot)$ denotes the activation function of our employed learning model in the output layer, x_T denotes a set of data samples for model inference, which is not a part of training data, and cannot be known during the model training phase, $\hat{\mathbf{y}}_n = \Phi_{\mathrm{h}}(\mathbf{x}_T; \bar{\mathbf{w}}_n)$ denotes the input of output layer, $\bar{\mathbf{w}}_n$ denotes the global model parameters of $\mathcal{M}_n$, and $\Phi_{\mathrm{h}}(\cdot;\cdot)$ characterizes the mapping relationship between the output of hidden layers and the input data samples.

As illustrated in Fig. 1, an ensemble model $\bar{\mathcal{M}}$ can be constructed by combining the output layers of all the generated federated learning models $\mathcal{M}_1, \ldots, \mathcal{M}_N$, and the output result of $\bar{\mathcal{M}}$ can be expressed as follows:

$$\bar{y} = \Phi_{\mathrm{o}}\left[\sum_{n=1}^{N} \beta_n \hat{\mathbf{y}}_n\right], \tag{6}$$

where $\hat{\mathbf{y}}_n$ follows the notations given by (5), and β_n denotes the weight of $\hat{\mathbf{y}}_n$. It indicates that the output result of $\bar{\mathcal{M}}$ is based on the inferring results of $\mathcal{M}_1, \ldots, \mathcal{M}_N$, and thus the model accuracy can be improved. It should be noted that $\hat{\mathbf{y}}_n$ is employed to ensure that the output of model ensemble is feasible.

Recalling (6), the contribution of $\hat{\mathbf{y}}_n$ on the output result of $\bar{\mathcal{M}}$ is determined by its weight β_n. To guarantee the accuracy performance, all the intra-cluster learning models should be assembled in a data distribution adaptive approach with the target inferring data set. The value of β_n should be large when $\mathcal{M}_n$ has experienced similar training data set with the target inferring data set, and vice versa. Therefore, it is important to capture the similarity of target inferring data $\mathbf{x}_T$ and experienced training data of $\mathcal{M}_n$. Similar to (2), the corresponding cosine similarity can be expressed as

$$\theta_n = \frac{\nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_0\right)^{\top} \nabla \bar{F}\left(\mathbf{w}_0; \mathbf{x}_T\right)}{\left\|\nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_0\right)\right\| \left\|\nabla \bar{F}\left(\mathbf{w}_0; \mathbf{x}_T\right)\right\|}, \tag{7}$$

where $\nabla \bar{F}\left(\mathbf{w}_0; \mathbf{x}_T\right)$ denotes the gradient with respect to $\mathbf{w}_0$ for $\mathbf{x}_T$ in (1), and $\nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_0\right)$ denotes the initial gradient of global averaged empirical risks with respect to C_n, and it can be defined identically with $\nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_t\right)$ in (4) by setting $t = 0$. All the users start from a common initialized parameter $\mathbf{w}_0$ and thus θ_n in (7) can well characterize the similarity of training data set and target inferring data set. Since the value of cosine similarity is in the range of $[-1, 1]$, while the weight β_n in (6) should be a non-negative coefficient. Therefore, based on (7), β_n in (6) can be expressed as

$$\beta_n = \frac{\theta_n + 1}{\sum_{i=1}^{N}(\theta_i + 1)}. \tag{8}$$

Similar to the local training data set, the corresponding data distribution of the target inferring data set also diverged due to the limitation of data volumes. Therefore, our ensemble model can achieve considerable accuracy performance on the studied target inferring data set when the intra-cluster federated learning models can be combined in a data distribution adaptive approach, and a feasible scheme can be designed based on (8).

By formulating individual clusters, our proposed scheme encourages the users with similar data distribution to learn collaboratively by employing intra-cluster federated learning, which can speed up the convergence. Moreover, to guarantee the test accuracy of all categories of learning data, multiple intra-cluster federated learning models are combined by using model ensemble, and thus the performance loss caused by data distribution divergence can be mitigated. Therefore, sophisticated design of user cluster formation is necessary to improve both the model accuracy and training efficiency, which will be studied in the next section.

Algorithm 1 A Framework of Ensemble Federated Learning

1: **Initialization**: Each user U_i initializes the model parameter $\mathbf{w}_{0,0}^i = \mathbf{w}_0$, $i = 1, \cdots, M$.
2: *Model Training Phase*: For each round of interaction $t = 0, 1, \cdots$
- Each user U_i updates the gradient $\nabla \bar{F}(\mathbf{w}_t^i; \mathcal{D}_i)$ based on (1) and transmits to B_T.
- User cluster formation when $t = 0$: N individual user clusters $C_1, \cdots, C_N$ are formulated at the base station by following Algorithm 2 in Section III.
- Intra-cluster federated learning: For each user cluster C_n, $n = 1, \cdots, N$
  - C_n updates $\nabla \bar{F}_{C_n}\left(\bar{\mathbf{w}}_t\right)$ and $\bar{\mathbf{w}}_{n,t+1}$ based on (4) at B_T.
  - $\bar{\mathbf{w}}_{n,t+1}$ is sent back to all the users for the next round of model update.

Until: All the cluster models $\mathcal{M}_n$ converges, i.e., $n = 1, \cdots, N$.
3: *Model Inference Phase*: Inter-cluster model ensemble for a specific target inferring data $\mathbf{x}_T$
- Each cluster model $\mathcal{M}_n$ obtains its own output $\hat{\mathbf{y}}_n$ and weight β_n based on (5) and (8).
- The final output result of the ensemble model $\bar{\mathcal{M}}$ is provided based on (6).

III. A Coalition Formation Game-Based Cluster Formation Algorithm

To ensure the performance of our proposed scheme, the cluster partition should be adaptive to the data distributions

and the channel conditions of users. Since the number of formulated clusters is unknown, it is challenging to obtain the global optimal cluster partition results in a computation-efficient way. To balance the tradeoff between the model performance and the computational complexity, our studied user cluster formation problem can be formulated as a coalition formation game in this section, and then a switching-based coalition formation algorithm for user clustering is designed.

As introduced in [49], the target of a coalition formation game is to divide all the players into a group of disjoint coalitions. In this paper, our studied federated learning cluster formation problem can be modeled as a coalition formation game $\mathcal{G}$. In particular, $\mathcal{U} = \{U_1, \ldots, U_M\}$ denotes a set of all the users, which can act as the players in $\mathcal{G}$. A cluster C_n can be treated as a coalition formed by a group of users, and $\Omega_l = \{C_1, \ldots, C_N\}$ denotes a partition result of coalition formation, $C_m \cap C_n = \varnothing$, $m \neq n$, $\cup_{n=1}^{N} C_n = \mathcal{U}$. Please note that the total number of coalitions is not fixed, i.e., $N = 1, \ldots, M$. $\mathcal{B}$ denotes the set that consists of all the possible partition results, $\Omega_l \in \mathcal{B}$. Moreover, a utility function is defined for each user to characterize its achieved benefit of participating in a specific federated learning cluster, i.e., $\varphi_i(C_n, \Omega_l)$ denotes the utility function of U_i when it joins a cluster C_n with a partition result Ω_l.

In a coalition formation game, each player decides to join/leave a coalition to improve its own utility without invading the benefit of other players. By following a series of operations, the partition results can be finally stabilized. To guarantee the performance of our proposed scheme, the utility function of each user should be sophisticatedly designed. Then, a switching-based coalition formation algorithm for user clustering can be provided, which can be proved that can it converge to a Nash-stable partition result. The details can be introduced as follows.

A. Utility Function Formulation

As introduced in [35], the utility of each user is defined as the payoff achieved by the user subtracting the cost in the coalition. Without loss of generality, we focus on a specific user U_i in our studied coalition formation game, and its utility function can be expressed as

$$\varphi_i(C_n, \Omega_l) = \varphi_{i,\mathrm{P}}(C_n, \Omega_l) - \gamma \varphi_{i,\mathrm{C}}(C_n, \Omega_l), \tag{9}$$

where $\varphi_{i,\mathrm{P}}(C_n, \Omega_l)$ denotes the payoff function of U_i with respect to its coalition C_n in partition Ω_l, $\varphi_{i,\mathrm{C}}(C_n, \Omega_l)$ denotes the corresponding cost function, and γ is set as a weight parameter to manipulate the impacts of $\varphi_{i,\mathrm{P}}(C_n, \Omega_l)$ and $\varphi_{i,\mathrm{C}}(C_n, \Omega_l)$ on $\varphi_i(C_n, \Omega_l)$.

In this paper, the objective of user cluster formation is to improve the performance and efficiency of intra-cluster federated learning models, which are assembled to deal with various target inferring data sets. Therefore, the payoff and the cost should be modeled as follows.

1) The Payoff Function $\varphi_{i,\mathrm{P}}(C_n, \Omega_l)$ in (9): The following two aspects is considered for the design of the payoff function: a) The accuracy performance and the convergence rate of each intra-cluster federated learning model. b) The generalization performance of ensemble model $\bar{\mathcal{M}}$.

For a federated learning cluster C_n, the performance of accuracy and convergence with respect to its intra-cluster federated learning model $\mathcal{M}_n$ is only determined by its own members. As introduced in [16] and [24], the convergence and accuracy performance loss of the intra-cluster federated learning model is caused by the distribution divergence of non-IID data, which can be effectively characterized by the cosine similarity with respect to the local gradients. Therefore, the intra-cluster payoff of user U_i can be defined as the averaged cosine similarity between user U_i and all other users in C_n to capture the performance gains achieved by user clustering:

$$\psi_{i,\mathrm{intra}}(C_n, \Omega_l) = \frac{1}{D_n} \sum_{\substack{U_j \in C_n, i \neq j, \\ C_n \in \Omega_l}} (\lambda_{i,j} + 1), \tag{10}$$

where $\lambda_{i,j}$ follows the definition given by (2), and D_n denotes the total number of users in C_n.

Recalling (10), it indicates that $\psi_{i,\mathrm{intra}}(C_n, \Omega_l)$ simply focuses on the accuracy and the convergence of $\mathcal{M}_n$ without considering the global ensemble model $\bar{\mathcal{M}}$. Due to the existence of data distribution divergence, the performance of $\bar{\mathcal{M}}$ cannot be guaranteed. In particular, some specific categories of learning data may be overlooked by all the intra-cluster federated learning models, which will cause performance loss that cannot be improved by model ensemble. Therefore, it requires that the cosine similarities of all the intra-cluster federated learning models should be reduced to cover all the categories of learning data.

As introduced in [36], the generalization performance on the potential target inferring data set is related with the inter-cluster difference. Similar to (10), the cosine similarity of U_i and any other intra-cluster federated learning model $\mathcal{M}_m$ can be derived based on the gradients of the empirical risks with respect to all the users in C_m. However, the computational complexity is extremely high, since it requires to exhaustively calculate the cosine similarity between U_i and each user in the cluster. To solve this problem, the gradient of global averaged empirical risks with respect to each intra-cluster federated learning model is employed, i.e., the cosine similarity of U_i and $\mathcal{M}_m$ can be expressed as

$$\bar{\lambda}_{i,C_m} = \frac{\nabla \bar{F}(\mathbf{w}_0; \mathcal{D}_i)^{\top} \nabla \bar{F}_{C_m}(\bar{\mathbf{w}}_0)}{\left\| \nabla \bar{F}(\mathbf{w}_0; \mathcal{D}_i) \right\| \left\| \nabla \bar{F}_{C_m}(\bar{\mathbf{w}}_0) \right\|}, \tag{11}$$

where $\nabla \bar{F}(\mathbf{w}_0; \mathcal{D}_i)$ follows the notations given by (1) which denotes the initial gradient of U_i, and $\nabla \bar{F}_{C_m}(\bar{\mathbf{w}}_0)$ follows the notation given by (4) which denotes the initial gradient of global averaged empirical risks with respect to $\mathcal{M}_m$. Based on (11), the averaged inter-cluster cosine similarity of U_i can be written as follows, which can characterize the impact of distribution divergence between U_i and other intra-cluster federated learning models:

$$\psi_{i,\mathrm{inter}}(C_n, \Omega_l) = \frac{1}{N-1} \sum_{\substack{U_i \in C_n, m \neq n, \\ C_m, C_n \in \Omega_l}} (\bar{\lambda}_{i,C_m} + 1). \tag{12}$$

Recalling (12), the data distribution of U_i is quite different with the users in other clusters when the value of $\psi_{i,\mathrm{inter}}(C_n, \Omega_l)$ is small. Therefore, it indicates that the

intra-cluster federated learning model $\mathcal{M}_n$ is important for the global ensemble model $\bar{\mathcal{M}}$, since $\mathcal{M}_n$ can handle some categories of learning data that are overlooked by other intra-cluster federated learning models.

The intra-cluster cosine similarity should be enhanced to guarantee the convergence rate and the accuracy of federated learning models [13]. The inter-cluster cosine similarity should be lowered to cover all categories of learning data and improve the model generalization performance. By following the method in [37], we define the payoff function as the ratio of the intra-cluster and the inter-cluster cosine similarities, which can be expressed as

$$\varphi_{i,\mathrm{P}}(C_n,\Omega_l)=\frac{\psi_{i,\mathrm{intra}}\left(C_n,\Omega_l\right)}{\psi_{i,\mathrm{inter}}\left(C_n,\Omega_l\right)}. \tag{13}$$

As shown in (13), when the value of $\varphi_{i,\mathrm{P}}(C_n,\Omega_l)$ is large, it indicates that the intra-cluster cosine similarity $\psi_{i,\mathrm{intra}}\left(C_n,\Omega_l\right)$ is large, while the inter-cluster cosine similarity $\psi_{i,\mathrm{inter}}\left(C_n,\Omega_l\right)$ is small. It ensures that $\mathcal{M}_n$ can be generated with fast convergence rate, which focuses on different categories of learning data with other intra-cluster federated learning models.

2) The Cost Function $\varphi_{i,\mathrm{C}}(C_n,\Omega_l)$ in (9): As introduced in Section II, the users are required to implement local model training and feedback to deploy federated learning in C_n, and thus the cost function can be defined as follows:

$$\varphi_{i,\mathrm{C}}(C_n,\Omega_l)=\phi_{i,\mathrm{train}}\left(C_n,\Omega_l\right)+\phi_{i,\mathrm{trans}}\left(C_n,\Omega_l\right), \tag{14}$$

where $\phi_{i,\mathrm{train}}\left(C_n,\Omega_l\right)$ denotes the expected energy consumption caused by local model training of user U_i with respect to its coalition C_n in partition Ω_l, and $\phi_{i,\mathrm{trans}}\left(C_n,\Omega_l\right)$ denotes the expected energy consumption caused by transmitting the model update results to the base station of user U_i with respect to its coalition C_n in partition Ω_l.

$\phi_{i_n,\mathrm{train}}\left(C_n,\Omega_l\right)$ can be modeled as a summation of the energy consumption for all T rounds in the intra-cluster federated learning, which can be expressed as

$$\phi_{i_n,\mathrm{train}}\left(C_n,\Omega_l\right)=TE\kappa_{i_n}L_{i_n}B_{i_n}f_{i_n}^2, \tag{15}$$

where f_{i_n} is the CPU/GPU frequency of U_{i_n}, L_{i_n} denotes the size of data set at U_{i_n}, B_{i_n} denotes the number of CPU/GPU cycles that are required to process one data sample at U_{i_n}, and κ_{i_n} is the effective capacitance of U_{i_n} [38].

To derive $\phi_{i,\mathrm{trans}}\left(C_n,\Omega_l\right)$ in (14), we first focus on the expected energy consumption of U_{i_n} for transmitting its normalized gradient update result $\nabla\tilde{F}(\mathbf{w}_t^{i_n};\mathcal{D}_{i_n})$. As introduced in Section II, $\nabla\tilde{F}(\mathbf{w}_t^{i_n};\mathcal{D}_{i_n})$ is transmitted repeatedly until it is successfully detected by the base station, and the corresponding expected energy consumption can be expressed as

$$\bar{\phi}_{i_n,\mathrm{trans}}\left(C_n,\Omega_l\right)=\sum_{\tau=1}^{K}\Pr\left\{K_{i_n}=\tau\right\}\tau E_{i_n}, \tag{16}$$

where τ denotes the transmission times, K is the maximum transmission times, $\Pr\{K_{i_n}=\tau\}$ is the probability that $\nabla\tilde{F}(\mathbf{w}_t^{i_n};\mathcal{D}_{i_n})$ is transmitted by τ times, and E_{i_n} denotes the energy consumption of transmitting $\nabla\tilde{F}(\mathbf{w}_t^{i_n};\mathcal{D}_{i_n})$ for a single time. Recalling (3), a closed-form expression of $\bar{\phi}_{i_n,\mathrm{trans}}\left(C_n,\Omega_l\right)$ can be provided by the following lemma.

Lemma *1: The expected energy consumption of transmitting $\nabla\tilde{F}(\mathbf{w}_t^{i_n};\mathcal{D}_{i_n})$ is expressed as (17), shown at the bottom of the next page, where W denotes the bandwidth, r_{i_n} is the target data rate of U_{i_n}, and B_{grad} denotes the data volume of the gradient update result in bits. In particular, when the maximum transmission time is large, i.e., $K\to\infty$, $\bar{\phi}_{i_n,\mathrm{trans}}\left(C_n,\Omega_l\right)$ in (17) can be further derived as*

$$\bar{\phi}_{i_n,\mathrm{trans}}\left(C_n,\Omega_l\right)=e^{\frac{\sigma^2}{Pd_{i_n}^{-\alpha}}\left(2^{\frac{r_{i_n}}{W}}-1\right)}\frac{B_{\mathrm{grad}}}{r_{i_n}}P. \tag{18}$$

Proof: Please refer to Appendix. ∎

Based on Lemma 1, the total expected energy consumption caused by feedback model update results of user U_{i_n} can be written as

$$\phi_{i_n,\mathrm{trans}}(C_n,\Omega_l)=T\bar{\phi}_{i_n,\mathrm{trans}}\left(C_n,\Omega_l\right). \tag{19}$$

B. A Switching-Based Coalition Formation Algorithm for User Clustering

As introduced in [39], our studied problem can be modeled as a coalition formation game in the partition form. Moreover, the utility of C_n can be modeled as a summation of the utility of its participants. As shown in (9), the utility of each user is deterministic for a given partition result. Therefore, our studied problem is with non-transferable utility, where the utility achieved by C_n cannot be arbitrarily allocated or transferred among its participants.

To solve our studied problem efficiently, a switching-based coalition formation algorithm for user clustering is designed, and the key idea is to formulate a group of disjoint coalitions via a series of negotiations between the users and the coalitions. During each round of negotiations, each user can individually decide to join/leave a coalition to improve its own utility. To guarantee the stability of partition results, the preference relation of users should be first defined.

1) Preference Relation of U_i: Without loss of generality, we focus on a specific user U_i. Since its utility is jointly determined by the partition results of all the coalitions, we focus on two different coalition-partition pairs, i.e., (C_m,Ω_k) and (C_n,Ω_l), which denote two coalitions C_m and C_n in its given partitions Ω_k and Ω_l, respectively. A preference relation of U_i with respect to (C_m,Ω_k) and (C_n,Ω_l) can be defined as follows:

$$(C_m,\Omega_k)\succ_i\left(C_n,\Omega_l\right),\ m\neq n,\ \text{or}\ k\neq l, \tag{20}$$

which indicates that U_i prefers to join C_m with a partition form Ω_k, rather than C_n with a partition form Ω_l. In particular, (20) can be employed to characterize the preference relations of the following three cases: 1) When $m\neq n$ and $k=l$, it captures the preference relation with respect to two different coalitions C_m and C_n in the same partition result. 2) When $m=n$ and $k\neq l$, it captures the preference relation with respect to a common coalition in different partitions Ω_k and Ω_l. 3) When $m\neq n$ and $k\neq l$, it captures the preference relation with respect to two different coalitions C_m and C_n in different partitions Ω_k and Ω_l. Please note that the preference relation $\succeq_i$ is similar to $\succ_i$, which indicates that U_i prefers to join one

coalition-partition pair rather than another, or it has the same preference for both coalition-partition pairs.

During the coalition formation procedure, U_i proposes to join another coalition C_q to improve its own utility. Meanwhile, its switching operation should not violate the benefits of other users in C_q, otherwise the proposal of U_i would be refused by C_q. Therefore, U_i is allowed to join C_q in partition Ω_p if and only if the following constraint can be satisfied:

$$\varphi_j\left(C_q, \Omega_p\right) \geqslant \varphi_j\left(C_q \setminus \{U_i\}, \Omega_p\right), \ \forall U_j \in C_q \setminus \{U_i\}, \quad (21)$$

where $\varphi_j(C_q, \Omega_p)$ and $\varphi_j(C_q \setminus \{U_i\}, \Omega_p)$ denote the utility functions of an original user U_j in C_q with and without U_i joining C_q in a given partition Ω_p, respectively. It indicates that the participation of U_i should not cause decrement of each user's utility in C_q.

Besides joining/leaving the existing coalitions, U_i is also allowed to formulate a singleton coalition C_q by itself, i.e.,

$$C_q = \{U_i\}. \quad (22)$$

Please note that the singleton coalition C_q in (22) can still be formed by the switching operation of U_i, i.e., U_i leaves its current coalition, and joins an empty coalition $\varnothing$.

In this paper, the preference function is defined to capture the preference relation, which can be designed based on the utility function to characterize the benefit achieved by a user joining/leaving a specific coalition. And then the following lemma can be provided.

Lemma 2: *The preference relation given by (20) can be established if and only if the following inequality with respect to the corresponding preference functions can be satisfied, i.e.,*

$$\left(C_m, \Omega_k\right) \succ_i \left(C_n, \Omega_l\right) \Leftrightarrow \delta_i\left(C_m, \Omega_k\right) > \delta_i\left(C_n, \Omega_l\right). \quad (23)$$

In particular, the preference function of U_i in C_q with a given partition Ω_p can be defined as (24), shown at the bottom of the next page, where $\varphi_i(C_q, \Omega_p)$ follows the definition given by (9), and Γ_i denotes a set consisting of all the coalition-partition pairs that U_i previously joined during the coalition formation procedure.

Recalling Lemma 2, the preference function is defined to guarantee the convergence and the performance of our studied coalition formation game: First, U_i is not allowed to join a coalition C_n in partition Ω_l if (C_n, Ω_l) is recorded in the historical coalition-partition set Γ_i. It can be avoided that U_i switches between two coalition-partition pairs repeatedly. Second, formulating a singleton coalition is an option for U_i during the coalition formation procedure, which should be considered to ensure that a stable partition result can be achieved. Moreover, when the constraints given in (24) cannot be satisfied, the preference function of U_i is set as zero since its switching operation is not allowed.

2) A Switching-Based Coalition Formation Algorithm: Based on Lemma 2, each user can switch from its current coalition to another coalition if (23) can be satisfied, and the partition form changes accordingly. Therefore, the coalitions can be formulated by a series of switching operations of users, and the switching rules can be defined as follows.

Definition 1: (Switching Rule) For a specific user U_i in a coalition C_n with a partition Ω_l, it can join another coalition C_m, and form a new partition Ω_k by switching operation, which can be expressed as the following transformation of coalition-partition pairs:

$$\left(C_n, \Omega_l\right) \rightarrow \left(C_m \cup \{U_i\}, \Omega_k\right), \quad (25)$$

where $C_m \in \Omega_l \cup \{\varnothing\}$, and $\Omega_k = \{\Omega_l \setminus \{C_m, C_n\}\} \cup \{C_m \cup \{U_i\}, C_n \setminus \{U_i\}\}$. In particular, the switching operation of U_i can be allowed if and only if

$$\left(C_m \cup \{U_i\}, \Omega_k\right) \succ_i \left(C_n, \Omega_l\right) \Leftrightarrow \ (25). \quad (26)$$

Algorithm 2 A Switching-Based Coalition Formation Algorithm for User Clustering

1: **Initialization**: $U_1, \ldots, U_M$ are randomly divided into N_0 clusters. The partition result can be denoted as $\Omega_0 = \{C_1, \ldots, C_{N_0}\}$, and the historical set given in (24) is initialized as $\Gamma_i = \varnothing$, $i = 1, \ldots, M$.
2: **Repeat**: For $U_i \in C_n$ in a given partition Ω_l
- U_i negotiates with another cluster C_m, $C_m \in \{\Omega_l \setminus \{C_n\}\} \cup \{\varnothing\}$.
- The values of preference functions $\delta_i(C_m \cup \{U_i\}, \Omega_k)$ and $\delta_i(C_n, \Omega_l)$ are calculated based on (24), where Ω_k follows the notations given in (25).
- **If** $\delta_i(C_m \cup \{U_i\}, \Omega_k) > \delta_i(C_n, \Omega_l)$
  - The preference relation in (26) can be established, and the switching operation given by (25) is executed.
  - A new coalition-partition pair $(C_m \cup \{U_i\}, \Omega_k)$ is formed, which is recorded in the historical set Γ_i.
- **Else** U_i stays in C_n without switching.

3: **Termination**: No user wants to leave its current cluster, and the partition result never changes.
4: **Return**: The final partition result Ω_{final}.

As shown in Algorithm 2, each user iteratively negotiates with another cluster, or tries to formulate a singleton cluster during the cluster formation procedure. To evaluate the benefit of joining its target cluster, the preference functions are first calculated based on (24) to formulate the preference relation between its current coalition-partition and target coalition-partition pairs. As shown in (26), it would switch to the target cluster if and only if its preference function can be improved. A new coalition-partition pair would be formed if it decides to join the target cluster, which should be kept in its historical

$$\bar{\phi}_{i_n,\text{trans}}\left(C_n, \Omega_l\right) = \sum_{\tau=1}^{K}\left[1 - e^{-\frac{\sigma^2}{Pd_{i_n}^{-\alpha}}\left(2^{\frac{r_{i_n}}{W}}-1\right)}\right]^{\tau-1} e^{-\frac{\sigma^2}{Pd_{i_n}^{-\alpha}}\left(2^{\frac{r_{i_n}}{W}}-1\right)} \frac{B_{\text{grad}}}{r_{i_n}} P\tau, \quad (17)$$

set. The final cluster formation result can be obtained until the partition result stays unchanged.

C. The Stability Analysis of Algorithm 2

As introduced in [40], the definition of Nash-stable for coalition formation games is first provided as follows.

Definition 2: (Nash-stable, [40]) A partition result Ω_l is Nash-stable, if and only if no user has the incentive to switch to another coalition from its current coalition, i.e., (27), shown at the bottom of the next page.

Based on Definition 2, it can be proved that Algorithm 2 can converge to a Nash-stable partition result, and the following theorem can be provided.

Theorem 1: Algorithm 2 can always converge to a Nash-stable partition result from any given initialized partition with finite steps of switching operations.

Proof: The stability of Algorithm 2 can be proved if the following two conditions can be satisfied: *1)* It can converge to an ultimate partition result via finite steps of switching operations. *2)* The achieved partition result is Nash-stable.

First, condition *1)* should be verified. Recalling Lemma 2, U_i has to stay in C_n when all the available coalition-partition pairs are recorded in its historical set Γ_i. Please note that the number of users is limited, which can formulate finite coalition-partition pairs. For a given partition form $\Omega_l = \{C_1, \ldots, C_N\}$, it can formulate $|\Omega_l|$ different coalition-partition pairs, i.e., $|\Omega_l| = N$. Therefore, $D = \sum_{l=1}^{B_M} |\Omega_l|$ coalition-partition pairs can be formed for our studied coalition formation game, where B_M denotes the M-th order Bell number [50] that can count up all the possible partition results. It indicates that at most D coalition-partition pairs can be added to Γ_i with the switching operations executed by U_i, $i = 1, \ldots, M$, and thus Algorithm 2 can converge to a final partition result Ω_{final} within $M \times D$ steps of switching operations.

Next, the stability of Ω_{final} is proved by contradiction to verify condition *2)*. In particular, we assume that the converged partition result Ω_{final} is not Nash-stable. Based on Definition 2, there exists at least one user U_i in a coalition C_n that intends to switch to another coalition C_m, $C_m, C_n \in \Omega_{\text{final}}$, $m \neq n$. Then, the following preference relation is established based on (27):

$$(C_m \cup \{U_i\}, \Omega_k) \succeq_i (C_n, \Omega_{\text{final}}), \tag{28}$$

where $\Omega_k = \{\Omega_{\text{final}} \setminus \{C_n, C_m\}\} \cup \{C_n \setminus \{U_i\}, C_m \cup \{U_i\}\}$.

Recalling (28), it shows that U_i prefers to switch to another coalition C_m, rather than stay in the current coalition C_n. The partition result also changes from Ω_{final} to Ω_k. It is contradictory with the fact that Ω_{final} is a converged partition result. Therefore, the proposed assumption is not correct, and Ω_{final} is a Nash-stable partition result.

Since both of conditions *1)* and *2)* are verified, the proof of Theorem 1 has been finished. ■

D. Convergence Discussions of Our Proposed Ensemble Federated Learning Scheme

As shown in Algorithm 1, our proposed scheme consists of two phases, which are the model training and the model inference phases, respectively.

During the model training phase, the convergence performance is determined by the final update results of intra-cluster federated learning, which is identical with the traditional federated learning with non-IID data. In Corollary 1 given by [16], the convergence of federated learning with non-IID data has been proved, which can guarantee the convergence of intra-cluster federated learning in our proposed scheme straightforwardly. Moreover, it indicates that the performance loss caused by non-IID data is determined by the summation of distribution divergence of all the users, which depends on the number of participators of federated learning. Since the number of users in each cluster is reduced after user clustering, the aggregated distribution divergence can be lowered and the convergence rate can be improved compared with other federated learning schemes without clustering.

During the model inference phase, the convergence can be guaranteed as long as a stable user clustering result can be obtained, which can be achieved by solving a coalition formation game based on Algorithm 2. In Theorem 1, it has been proved that Algorithm 2 can always get a Nash-stable partition result, which can guarantee the convergence of our proposed scheme in the model inference phase.

Since the convergence of both the model training and the model inference phases can be guaranteed, our proposed scheme can always converge.

E. Further Discussions on the Communication and the Computation Costs

First, please note that our proposed user clustering algorithm is executed simply at the centralized server of base station, but not among the users distributively. The motivation of employing coalition formation game is to provide an efficient method to formulate user clusters without knowing the number of clusters, which cannot be solved efficiently by the conventional optimization theory [45]. Since all the required information of our proposed algorithm, such as model update results and the network status parameters, is all available at the base station, it does not cause extra communication cost.

Second, the user clusters are formulated only once during the initial round of model update, and the following training procedure is identical with the conventional federated learning without other extra computation cost. In [35], it has been proved that the computation complexity of coalition formation

$$\delta_i(C_q, \Omega_p) = \begin{cases} \varphi_i(C_q, \Omega_p), & \text{if (21) and } (C_q, \Omega_p) \notin \Gamma_i, \text{ or (22)} \\ 0, & \text{else} \end{cases}, \tag{24}$$

TABLE II
EXPERIMENT DATA SETS AND EMPLOYED LEARNING MODELS

Task	Data set	Model	Model design details
Image classification tasks	MNIST [41], FEMNIST [42]	MLR	A fully connected layer with 10 / 26 units, followed by a softmax activation
		MLP	A hidden layer with 128 / 512 units, followed by a ReLU activation
	CIFAR-10 [43]	VGG [46]	4 convolutional layers with 3×3 filters, followed by two fully connected layers with 512 and 10 units
		ResNet [47]	A 7×7 convolutional layer with 64 channels is employed, followed by 16 residual blocks and a fully connected layer with 10 units
Regression tasks	California Housing Prices [48]	MLR	A fully connected layer with a unit, followed by a linear activation
		MLP	A hidden layer with 32 units, followed by a ReLU activation

grows polynomially with respect to the number of users, i.e., $\mathcal{O}(M^2)$, which is not high compared with the computation cost of model training. Therefore, except solving a coalition formation game only once with low computation complexity, our proposed scheme does not cause other computation cost during the model training phase.

IV. SIMULATION RESULTS

In this section, the simulation results are provided to evaluate the performance of our proposed ensemble federated learning scheme. We focus on a single base station scenario, and its coverage region C_T is modeled as a disc with a fixed radius $R = 500$ m. The allocated bandwidth of each user is set as $W = 10$ MHz. The transmit power is set as $P = 20$ dBm, and the spectral density of noise is set as -174 dBm/Hz. The path loss exponent is set as $\alpha = 4$. Moreover, the number of CPU/GPU cycles required to process one data sample is $B_{i_n} = 737.5$ cycles/bit, and the CPU/GPU frequency of each user is $f_{i_n} = 1$ GHz. The effective capacitance is set as $\kappa_{i_n} = 1 \times 10^{-29}$. The detailed settings of experiment data sets and employed learning models are provided in Table II. To characterize the features of non-IID data, the training data sets of all the users are divided by following a non-IID division method in [33].

A. Accuracy Performance of Ensemble Federated Learning

In Fig. 2 and Fig. 3, the test accuracy is plotted to evaluate the performance of our proposed ensemble federated learning scheme. The classic global federated learning scheme in [8] is selected as a baseline. The accuracy performance is obtained by evaluating the test accuracy of global model on the target inferring data set for inference.

In Fig. 2, our proposed scheme can converge more stably and faster than the baseline scheme. Moreover, it can always achieve better test accuracy performance by integrating multiple intra-cluster federated learning models via model ensemble. As depicted in Fig. 2b, the MLP model with MNIST data set converges with 38 rounds by employing our proposed scheme when $M = 60$, while the baseline scheme needs 70 rounds. Moreover, it shows that the test accuracy can be improved from 87.3% to 91.0%. The performance gains can be enlarged with CIFAR data set, i.e., the test accuracy can be improved from 49.5% to 73.9% for the VGG model, and from 54.2% to 76.2% for the ResNet model.

In Fig. 3, the performance of our proposed scheme is provided by employing the California Housing Prices data set. As shown in the figure, the root mean squared error (RMSE) is plotted, which decreases stably and finally converges. Moreover, our proposed scheme can achieve higher prediction accuracy, compared with traditional global federated learning scheme. In particular, the RMSE can be reduced from 0.5564 to 0.4899 for the MLR model, and the performance gains can be enlarged when the MLP model is employed, i.e., from 0.5279 to 0.4527.

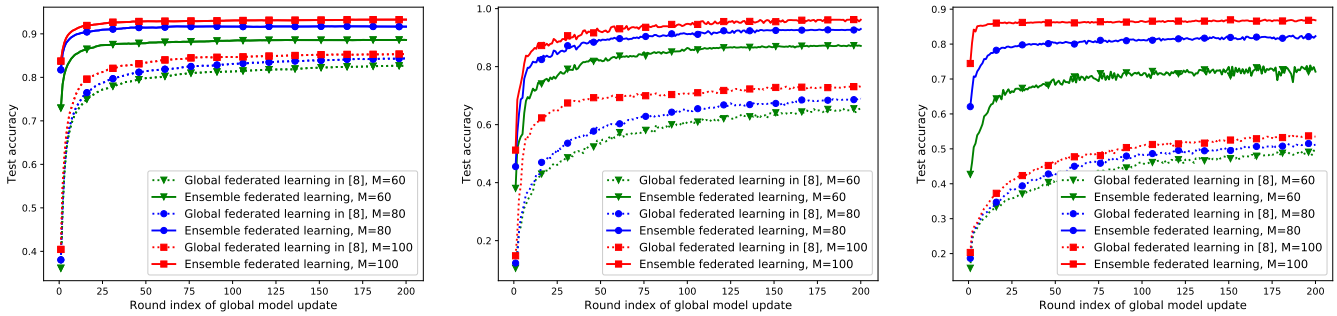
B. Convergence Performance of Algorithm 2 in Section III

In Fig. 4, the convergence performance of Algorithm 2 is evaluated. Both the total utility of all the users and the number of switching operations are plotted with respect to the negotiation round of users. As shown in the figures, the total utility keeps increasing as the index of negotiation rounds increases, and finally saturates at a fixed value. Moreover, during each round, the number of switching operations taken by the users keeps decreasing as the index increases, and converges to zero. It can verify the convergence of Algorithm 2. In particular, the simulation results show that Algorithm 2 can converge within 8 rounds when the number of users is set as $M = 60$ and 100, which indicate that it can achieve Nash-stable partition results with fast convergence rate.

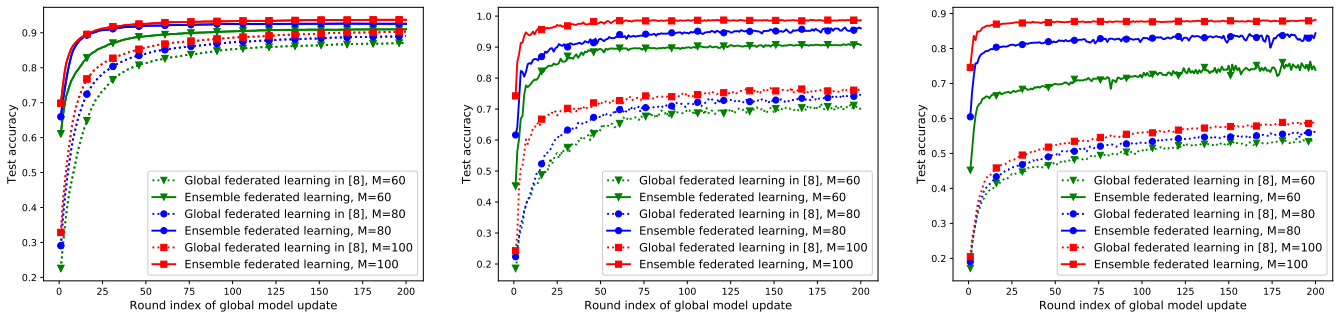
C. Performance Comparisons With Different User Clustering Algorithms

In Fig. 5 and Table III, the test accuracy performance and the cost of Algorithm 2 are evaluated by comparing with existing user clustering strategies for federated learning. In particular, the following algorithms are selected as benchmarks:1) Benchmark 1: Global federated learning in [8] without clustering. 2) Benchmark 2: Ensemble federated learning with

$$\Omega_l \text{ is Nash} - \text{stable} \Leftrightarrow (C_n, \Omega_l) \succeq_i (C_m \cup \{U_i\}, \Omega_k),\ i = 1, \ldots, M,\ U_i \in C_n,$$
$$\text{for } \forall C_m \in \Omega_l \cup \{\varnothing\},\ m \neq n,\ \text{and } \Omega_k = \{\Omega_l \setminus \{C_n, C_m\}\} \cup \{C_n \setminus \{U_i\}, C_m \cup \{U_i\}\}. \quad (27)$$

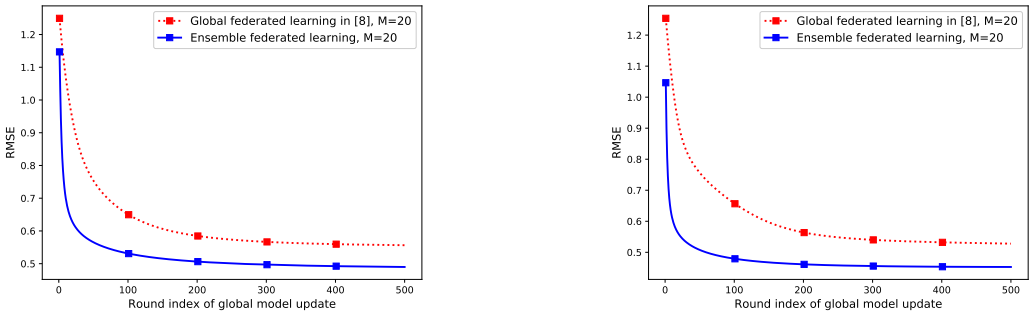


(a) MNIST data set with MLR. (c) FEMNIST data set with MLR. (e) CIFAR data set with VGG.



(b) MNIST data set with MLP. (d) FEMNIST data set with MLP. (f) CIFAR data set with ResNet.

Fig. 2. Accuracy performance of the image classification tasks.



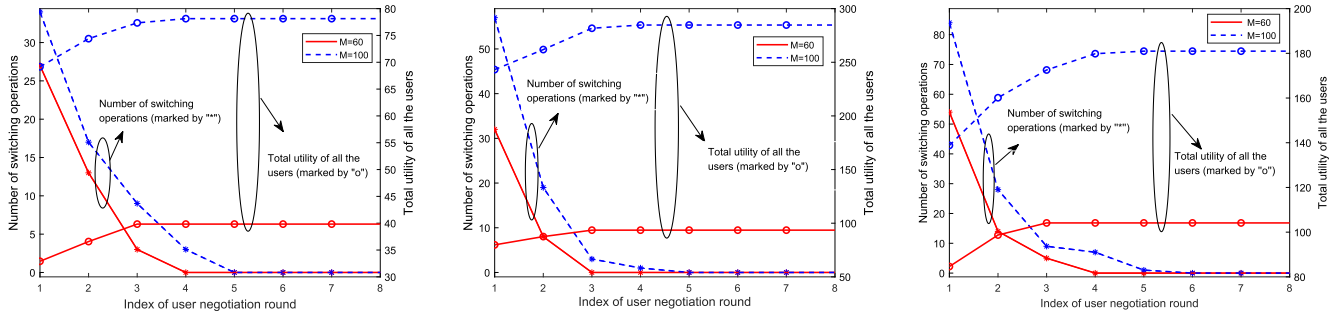
(a) With the MLR model. (b) With the MLP model.

Fig. 3. Accuracy performance of regression tasks on the California Housing Prices data set.

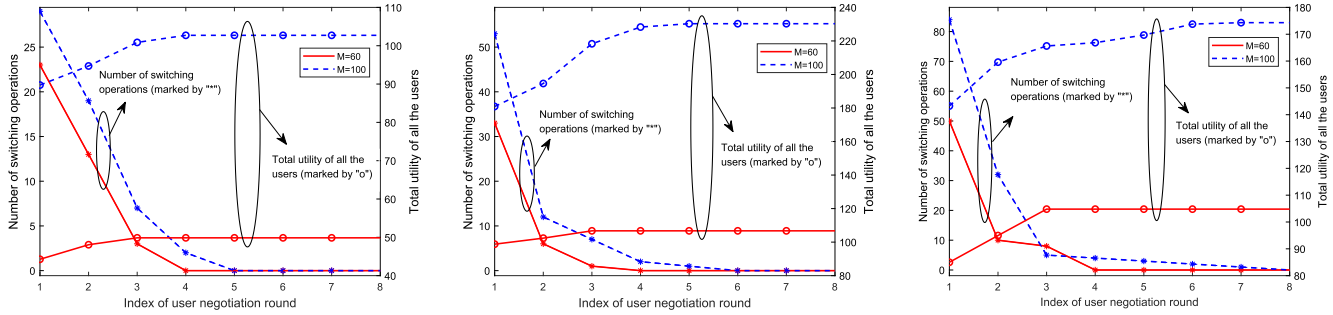
random user clustering. 2) Benchmark 3: Ensemble federated learning with K-Means user clustering in [44] based on the cosine similarity.

As shown in Fig. 5, Algorithm 2 can always achieve the best test accuracy performance. In particular, the simulation results show that the accuracy performance of the ensemble federated learning schemes, such as Algorithm 2 and Benchmark 2 and 3 in the figures, is better than the conventional global federated learning paradigm, which verifies the necessity of employing ensemble federated learning. By employing Algorithm 2 and Benchmark 3, the performance of ensemble federated learning can be significantly improved via sophisticated design of user cluster formation. Moreover, the cluster partition results of Algorithm 2 are adaptive to the data distribution divergence, i.e., the number of user clusters is flexible, while Benchmark 3 requires to fix the number of user clusters. Therefore, Algorithm 2 can achieve better accuracy performance than Benchmark 3. As illustrated in Fig. 5a, compared with Benchmark 3, the test accuracy can be improved from 90.8% to 91.7% when the MLR model is applied on MNIST data set with $M = 80$, and the performance gains are almost the same for the MLP model. When the VGG and ResNet models are used on CIFAR data set with $M = 60$, the performance can be improved from 72.2% to 73.9%, and from 75.0% to 76.2%, respectively.

The evaluation results of the communication and the computation costs for our proposed scheme are provided in Table III. As shown in the table, the communication cost of our proposed scheme is identical with both benchmarks, which are caused by the feedback of local model update results only. Moreover, the computation cost at the users is also the same. Although the clustering procedure of our proposed scheme and benchmark 3 require extra computation operations at the base station, the corresponding computation cost is quite low, while the improvement of accuracy performance is significant. In particular, compared with local model training cost, our proposed scheme causes only 0.127 % extra floating-point operations at the base station when the VGG model is employed on CIFAR

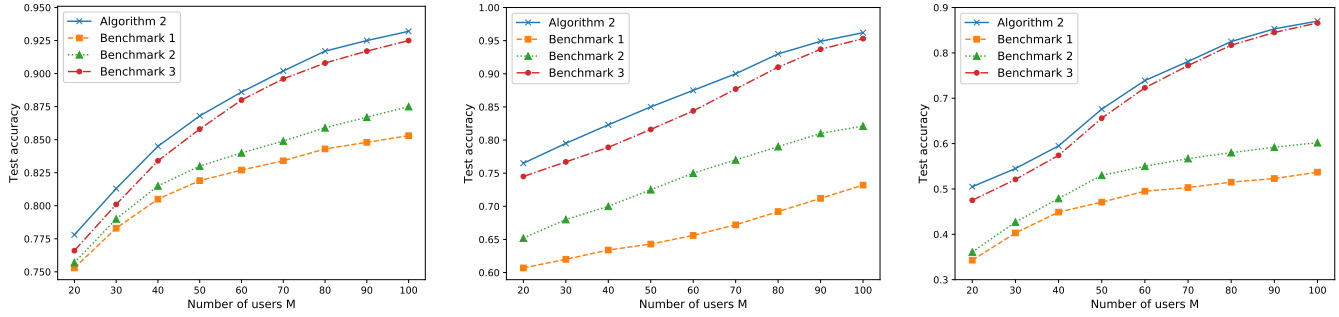


(a) MNIST data set with MLR. (c) FEMNIST data set with MLR. (e) CIFAR data set with VGG.

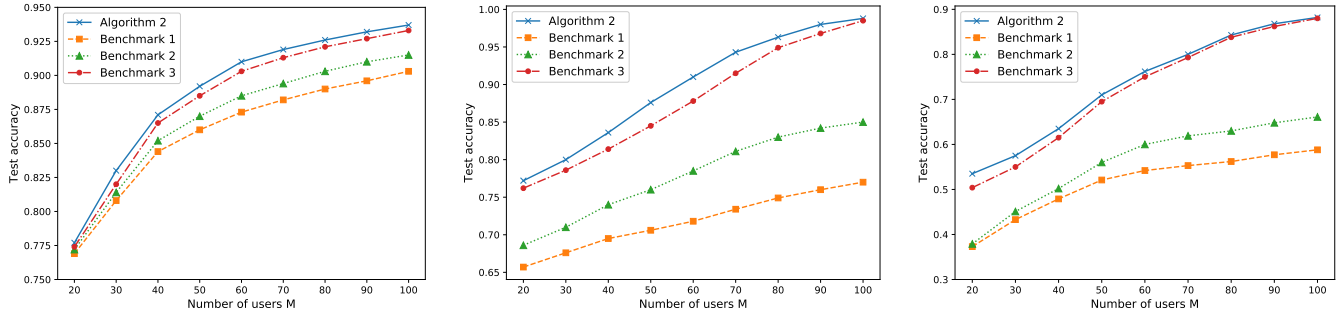


(b) MNIST data set with MLP. (d) FEMNIST data set with MLP. (f) CIFAR data set with ResNet.

Fig. 4. Convergence performance of Algorithm 2 in Section III.



(a) MNIST data set with MLR. (c) FEMNIST data set with MLR. (e) CIFAR data set with VGG.



(b) MNIST data set with MLP. (d) FEMNIST data set with MLP. (f) CIFAR data set with ResNet.

Fig. 5. Test accuracy of ensemble federated learning with different user clustering algorithms ($K = 20$).

data set, and the accuracy performance can be improved from 49.5 % to 73.9 %.

D. Performance Comparisons With Existing Federated Learning Schemes for Non-IID Data

In Table IV, it shows that our proposed ensemble federated learning scheme can always achieve the best test accuracy performance. When we focus on the MNIST data set, Per-FedAvg in [32] can achieve considerable accuracy performance, i.e., 91.0% and 91.9% for the MLR and the MLP models, respectively. By employing ensemble federated learning, the accuracy performance still can be improved to 91.7% and 92.6%, respectively. Moreover, the performance of Per-FedAvg scheme is deteriorated on the FEMNIST data

TABLE III
COMPUTATION AND COMMUNICATION COSTS COMPARISONS WITH DIFFERENT USER CLUSTERING ALGORITHMS

Data set-Model	Algorithm	Computation cost		Communication cost
		Local model update at the users (MFLOPs)	User cluster formation at the base station (MFLOPs)	(MBytes)
MNIST-MLR	Benchmarks 1 and 2	100.74	0	0.03
	Benchmark 3	100.74	0.79	0.03
	Algorithm 2	100.74	1.24	0.03
MNIST-MLP	Benchmarks 1 and 2	1.30×10^4	0	0.39
	Benchmark 3	1.30×10^4	10.18	0.39
	Algorithm 2	1.30×10^4	10.65	0.39
FEMNIST-MLR	Benchmarks 1 and 2	261.91	0	0.08
	Benchmark 3	261.91	2.04	0.08
	Algorithm 2	261.91	2.33	0.08
FEMNIST-MLP	Benchmarks 1 and 2	5.31×10^4	0	1.58
	Benchmark 3	5.31×10^4	41.53	1.58
	Algorithm 2	5.31×10^4	41.86	1.58
CIFAR-VGG	Benchmarks 1 and 2	3.41×10^5	0	16.51
	Benchmark 3	3.41×10^5	432.87	16.51
	Algorithm 2	3.41×10^5	433.27	16.51
CIFAR-ResNet	Benchmarks 1 and 2	9.05×10^5	0	79.33
	Benchmark 3	9.05×10^5	2129.65	79.33
	Algorithm 2	9.05×10^5	2130.03	79.33

TABLE IV
TEST ACCURACY PERFORMANCE COMPARISONS WITH THE EXISTING ENHANCED SCHEMES FOR NON-IID DATA

Schemes	MNIST-MLR	MNIST-MLP	FEMNIST-MLR	FEMNIST-MLP	CIFAR-VGG	CIFAR-ResNet
FedAvg in [8]	84.3%	89.0%	69.2%	74.9%	51.5%	56.2%
FedProx in [33]	85.0%	89.6%	71.3%	76.6%	52.2%	56.5%
FedDane in [34]	83.3%	86.7%	64.0%	70.2%	46.3%	49.9%
FedPer in [22][1]	–	90.6%	–	77.4%	73.7%	75.7%
FeSEM in [25]	82.2%	87.3%	64.6%	71.7%	52.5%	58.3%
FedGroup in [26]	90.6%	91.1%	83.6%	86.7%	75.3%	77.5%
Per-FedAvg in [32]	91.0%	91.9%	70.7%	72.3%	74.4%	76.8%
Ensemble FL	**91.7%**	**92.6%**	**93.0%**	**96.3%**	**82.5%**	**84.3%**

set, since it is difficult to personalize the learning models by employing only a few data samples for local fine-tune. In this case, the performance can be significantly improved by employing ensemble federated learning. In particular, the test accuracy performance can be improved to 93.0% and 96.3% for the MLR and the MLP models, respectively. Therefore, ensemble federated learning can achieve the best accuracy performance. The performance can be improved to 82.5% and 84.3% for the VGG and the ResNet models on the CIFAR-10 data set, respectively.

V. CONCLUSION

In this paper, an ensemble federated learning paradigm has been proposed in wireless networks. By intra-cluster federated learning, our proposed scheme can improve the convergence rate and the accuracy performance of federated learning with non-IID data for some specific learning data, and then an integrated model can be obtained via inter-cluster model ensemble to make full use of all the intra-cluster federated learning models for generalization. Next, a coalition formation-based algorithm for user clustering has been designed with considering the unreliability of feedback model update results for ensemble federated learning in wireless networks, which can be proved to be Nash-stable. Finally, the simulation results are provided to verify the performance gains of our proposed scheme with non-IID data in wireless networks.

APPENDIX
PROOF OF LEMMA 1

(16) can be expressed as follows when $h_{i_n,t}$ in (3) is IID flat Rayleigh fading:

$$\Pr\{K_{i_n} = \tau\} = \varrho_{i_n}^{\tau-1}\left(1-\varrho_{i_n}\right), \tag{29}$$

where ϱ_{i_n} denotes the outage probability of transmitting $\nabla\tilde{F}(\mathbf{w}_{i_n,t};\mathcal{S}_{i_n,t})$ for a single time. ϱ_{i_n} can be written as follows:

$$\begin{aligned}\varrho_{i_n} &= \Pr\left\{c_{i_n} < r_{i_n}\right\} = \Pr\left\{\rho_{i_n} = \frac{|h_{i_n}|^2 d_{i_n}^{-\alpha} P}{\sigma^2} < 2^{\frac{r_{i_n}}{W}} - 1\right\} \\ &= 1 - e^{-\frac{\sigma^2}{P d_{i_n}^{-\alpha}}\left(2^{\frac{r_{i_n}}{W}}-1\right)},\end{aligned} \tag{30}$$

where ρ_{i_n} is the received SNR of U_{i_n}. It is derived based on the fact that $|h_{i_n}|^2$ follows exponential distribution, i.e., $|h_{i_n}|^2 \sim \exp(1)$, and $E_{i_n} = B_{\text{grad}}/r_{i_n}P$. Then (17) can be proved. Moreover, $\bar{\phi}_{i_n,\text{trans}}\left(C_n,\Omega_l\right)$ can be rewritten as follows when $K \to \infty$:

$$\bar{\phi}_{i_n,\text{trans}}\left(C_n,\Omega_l\right) = \sum_{\tau=1}^{\infty}\varrho_{i_n}^{\tau-1}\left(1-\varrho_{i_n}\right)\tau E_{i_n} \overset{(a)}{=} \frac{E_{i_n}}{1-\varrho_{i_n}}, \tag{31}$$

where (a) in (31) follows the fact that the series converges to a constant, i.e., $\sum_{\tau=1}^{\infty} \tau \varrho_{i_n}^{\tau-1} = \frac{1}{(1-\varrho_{i_n})^2}$. Then (18) can be proved. And the proof has been finished.

REFERENCES

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[3] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, and K. Wu, "Artificial-intelligence-enabled intelligent 6G networks," *IEEE Netw.*, vol. 34, no. 6, pp. 272–280, Nov. 2020.

[4] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.

[5] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[6] L. Lovén et al., "EdgeAI: A vision for distributed, edge-native artificial intelligence in future 6G networks," in *Proc. 6G Wireless Summit*, Levi, Finland, Mar. 2019, pp. 1–2.

[7] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[8] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, Dec. 2017, pp. 1273–1282.

[9] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.

[10] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[11] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[12] Z. Zhao, C. Feng, H. H. Yang, and X. Luo, "Federated-learning-enabled intelligent fog radio access networks: Fundamental theory, key techniques, and future trends," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 22–28, Apr. 2020.

[13] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.I.d. data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[14] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-IID data," in *Proc. Int. Joint Conf. Neural Netw.*, Glasgow, U.K., Jul. 2020, pp. 1–9.

[15] L. Li et al., "FedSAE: A novel self-adaptive federated learning framework in heterogeneous systems," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Shenzhen, China, Jul. 2021, pp. 1–10.

[16] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*.

[17] Z. Zhao et al., "Federated learning with non-IID data in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1927–1942, Mar. 2022.

[18] N. Yoshida, T. Nishio, M. Morikura, K. Yamamoto, and R. Yonetani, "Hybrid-FL for wireless networks: Cooperative learning mechanism using non-IID data," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–7.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[20] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 2672–2680.

[21] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *J. Amer. Stat. Assoc.*, vol. 82, no. 398, pp. 528–540, Jun. 1987.

[22] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," 2019, *arXiv:1912.00818*.

[23] P. Pu Liang et al., "Think locally, act globally: Federated learning with local and global representations," 2020, *arXiv:2001.01523*.

[24] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3710–3722, Aug. 2021.

[25] G. Long et al., "Multi-center federated learning: Clients clustering for better personalization," 2021, *arXiv:2108.08647*.

[26] M. Duan et al., "FedGroup: Efficient clustered federated learning via decomposed data-driven measure," 2020, *arXiv:2010.06870*.

[27] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

[28] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[29] R. Jiang and S. Zhou, "Cluster-based cooperative digital over-the-air aggregation for wireless federated edge learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2020, pp. 887–892.

[30] W. Guo, C. Huang, X. Qin, L. Yang, and W. Zhang, "Dynamic clustering and power control for two-tier wireless federated learning," 2022, *arXiv:2205.09316*.

[31] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," 2018, *arXiv:1802.07876*.

[32] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 3557–3568, Dec. 2020.

[33] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst. (MLSys)*, Austin, TX, USA, Mar. 2020, pp. 429–450.

[34] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smithy, "FedDANE: A federated Newton-type method," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 1227–1231.

[35] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.

[36] Y. Yan and M. Kamel, "Clustering ensemble using swarm intelligence," in *Proc. IEEE Swarm Intell. Symp.*, Indianapolis, IN, USA, Mar. 2003, pp. 65–71.

[37] S. E. Kim and I. W. Seo, "Artificial neural network ensemble modeling with conjunctive data clustering for water quality prediction in rivers," *J. Hydro-Environ. Res.*, vol. 9, no. 3, pp. 325–339, Sep. 2015.

[38] E. García-Martín, C. F. Rodrigues, G. Riley, and H. Grahn, "Estimation of energy consumption in machine learning," *J. Parallel Distrib. Comput.*, vol. 134, pp. 75–88, Dec. 2019.

[39] W. Saad, Z. Han, R. Zheng, A. Hjorungnes, T. Basar, and H. V. Poor, "Coalitional games in partition form for joint spectrum sensing and access in cognitive radio networks," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 2, pp. 195–209, Apr. 2012.

[40] W. Saad, Z. Han, T. Basar, M. Debbah, and A. Hjorungnes, "A coalition formation game in partition form for peer-to-peer file sharing networks," in *Proc. IEEE Global Telecommun. Conf.*, Miami, FL, USA, Dec. 2010, pp. 1–5.

[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[42] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2921–2926.

[43] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[44] D. Arthur and S. Vassilvitskii, "K-Means++: The advantages of careful seeding," in *Proc. ACM-SIAM Symp. Discrete Algorithms*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[45] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, "Coalitional game theory for communication networks," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 77–97, Sep. 2009.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–12.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[48] California Census. (Nov. 2017). *California Housing Prices*. [Online]. Available: https://www.kaggle.com/camnugent/california-housingprices
[49] Z. Han, D. Niyato, W. Saad, T. Baar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, Oct. 2011.
[50] D. Ray, *A Game-Theoretic Perspective on Coalition Formation*. New York, NY, USA: Oxford Univ. Press, Jan. 2007.
[51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

Zhongyuan Zhao (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2009 and 2014, respectively.

He is currently an Associate Professor with BUPT. His research interests include fog computing/edge computing, content caching, and edge intelligence in wireless networks. He was a recipient of the Best Paper Awards from the IEEE CIT 2014 and the WASA 2015. He received the Exemplary Editors Award twice in 2017 and 2018. Since 2016, he has been serving as an Editor for IEEE COMMUNICATIONS LETTERS. He was a Guest Editor of IEEE ACCESS.

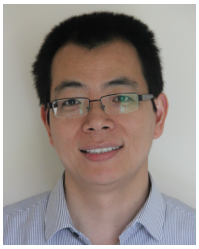
Jingyi Wang received the M.E. degree in communication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China. He is currently devoted to standardization research on B5G/6G technologies with the China Telecom Research Institute, Beijing. His research interests include wireless communication, machine learning, and federated learning.

Wei Hong received the M.S. degree from the Beijing University of Posts and Telecommunications, (BUPT), Beijing, China. He has more than 15 years of standardization and industry experience. He currently leads Xiaomis standardization activities at several international and domestic standardization bodies, including IEEE, 3GPP, GSMA, and WFA. He has broad knowledge on developments in emerging technologies and strategies. He has rich experience on how to realize the new technologies in the industry and also transfer industry's real requirements and practical technologies into standards. His expertise areas include 5G/B5G/6G, AI, smart home, and the IoT. Besides, he holds several positions in different working groups of different standardization bodies, including the Vice Chair of IEEE 1939.1, the Deputy Chair of GSMA TSGAI, the Vice Chair of CCSA TC11 WG1 Mobile Internet and Car Group, a board member of SAC/TC426, Co-rapporteur of 3GPP Rel-14 enhanced LTE-WLAN Aggregation. He was an active industry player, he has successfully led several series of national and industry standards.

Tony Q. S. Quek (Fellow, IEEE) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, in 2008.

He is currently the Cheng Tsang Man Chair Professor with the Singapore University of Technology and Design (SUTD) and a ST Engineering Distinguished Professor. He serves as the Director of the Future Communications Research and Development Program, the Head of ISTD Pillar, and the Deputy Director of the SUTD-ZJU IDEA. His current research interests include wireless communications and networking, network intelligence, non-terrestrial networks, open radio access networks, and 6G.

Dr. Quek is a fellow of the Academy of Engineering Singapore. He was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards–Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, the 2020 IEEE Communications Society Young Author Best Paper Award, the 2020 IEEE Stephen O. Rice Prize, the 2020 Nokia Visiting Professor, and the 2022 IEEE Signal Processing Society Best Paper Award. He has been actively involved in organizing and chairing sessions and served as a member of the Technical Program Committee and the symposium chairs in a number of international conferences. He is serving as an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.



Zhiguo Ding (Fellow, IEEE) received the B.Eng. degree from the Beijing University of Posts and Telecommunications in 2000 and the Ph.D. degree from Imperial College London in 2005.

He is currently a Professor of communications with Khalifa University. His research interests include 6G networks, cooperative and energy harvesting networks, and statistical signal processing.

Dr. Ding recently received the EU Marie Curie Fellowship 2012–2014, the Top IEEE TVT Editor 2017, the IEEE Heinrich Hertz Award 2018, the IEEE Jack Neubauer Memorial Award 2018, the IEEE Best Signal Processing Letter Award 2018, the Friedrich Wilhelm Bessel Research Award 2020, and the IEEE SPCC Technical Recognition Award 2021. He is a Distinguished Lecturer of IEEE ComSoc and a Web of Science Highly Cited Researcher in two categories in 2022. He was an Editor of IEEE WIRELESS COMMUNICATION LETTERS, IEEE TRANSACTIONS ON COMMUNICATIONS, and IEEE COMMUNICATIONS LETTERS. He is serving as an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, an Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *Science China Information Sciences*, and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.

Mugen Peng (Fellow, IEEE) received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. Afterward, he joined BUPT, where he has been a Full Professor with the School of Information and Communication Engineering, since 2012. In 2014, he was an Academic Visiting Fellow with Princeton University, Princeton, NJ, USA. He leads a Research Group focusing on wireless transmission and networking technologies with the State Key Laboratory of Networking and Switching Technology, BUPT. He has authored/coauthored more than 100 refereed IEEE journal articles and more than 300 conference proceeding papers. He was a recipient of the 2018 Heinrich Hertz Prize Paper Award, the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the Best Paper Award in the JCN 2016 and the IEEE WCNC 2015. He is on the Editorial/Associate Editorial Board of the *IEEE Communications Magazine*, the IEEE INTERNET OF THINGS JOURNAL, and IEEE ACCESS.