

Client Selection Approach in Support of Clustered Federated Learning over Wireless Edge Networks

Abdullatif Albaseer, Mohamed Abdallah, Ala Al-Fuqaha, and Aiman Erbad
Division of Information and Computing Technology, College of Science and Engineering,
Hamad Bin Khalifa University, Doha, Qatar
{amalbaseer, moabdallah, aalfuqaha, AERbad}@hbku.edu.qa

Abstract—Clustered Federated Multitask Learning (CFL) was introduced as an efficient scheme to obtain reliable specialized models when data is imbalanced and distributed in a non-i.i.d. (non-independent and identically distributed) fashion amongst clients. While a similarity measure metric, like the cosine similarity, can be used to endow groups of the client with a specialized model, this process can be arduous as the server should involve all clients in each of the federated learning rounds. Therefore, it is imperative that a subset of clients is selected periodically due to the limited bandwidth and latency constraints at the network edge. To this end, this paper proposes a new client selection algorithm that aims to accelerate the convergence rate for obtaining specialized machine learning models that achieve high test accuracies for all client groups. Specifically, we introduce a client selection approach that leverages the devices' heterogeneity to schedule the clients based on their round latency and exploits the bandwidth reuse for clients that consume more time to update the model. Then, the server performs model averaging and clusters the clients based on predefined thresholds. When a specific cluster reaches a stationary point, the proposed algorithm uses a greedy scheduling algorithm for that group by selecting the clients with less latency to update the model. Extensive experiments show that the proposed approach lowers the training time and accelerates the convergence rate by up to 50% while imbuing each client with a specialized model that is fit for its local data distribution.

Index Terms—Wireless Network Edge, Clustered Federated Edge learning (CFL), Clients Scheduling, Convergence Rate, None-i.i.d.

I. INTRODUCTION

IN wireless edge networks, Internet of Things (IoT) devices produce a massive volume of data that can be used to understand the state of the system or predict its future states. Mobile Edge Computing can utilize the unprecedented capabilities of edge devices and access points (APs) to perform complex tasks and enable intelligent services [1]. *Artificial Intelligence* (AI), and *Machine Learning* (ML) specifically, have witnessed high-speed developments and started to introduce intelligent services that can revolutionize our lives. Recently, several works have investigated the utilization of ML algorithms for future wireless edge applications, considering edge networks as an enabling technology for this vision [2], [3].

However, conventional ML techniques necessitate data to be offloaded and stored in a single central server which raises significant privacy concerns, increases the volume of data generated by edge devices, and the transmission delay of the limited communication resources when uploading data to the centralized server [4].

In response, Google proposed a promising decentralized machine learning technique termed Federated learning (FL)

that addresses these problems and keeps the data where it is produced. The learning algorithm is run within edge devices, and only model parameters are shared with the server [5] which initially broadcasts the latest version of the model for the update process. The server coordinates this process by aggregating and averaging all updates to form the global model. The server repeats these steps until the global model converges. Note that we refer to FL as Federated Edge Learning (FEEL) throughout this paper as our work focuses on deploying FL in wireless edge networks.

In fact, there are two major challenges, statistical heterogeneity and resource heterogeneity, when deploying FEEL. For the statistical heterogeneity, the data across the network is massively distributed in non-i.i.d and imbalanced fashion [3], [5]–[7]. As for the resource heterogeneity challenge, devices at the wireless edge are heterogeneous, and the bandwidth is limited; therefore, restricting the number of devices participating in a particular FEEL round. This heightens the need for efficient client selection strategies that contribute to a faster convergence rate, especially for large-scale edge networks.

To address the statistical challenges, a new approach called Clustered Federated Learning (CFL) has been proposed [8]. The multi-task clustering method utilizes the geometric characteristics of the FEEL loss surface to group the clients into clusters with federally trainable data distributions. It is worth noting that traditional FEEL assumes that only a single model can be trained on the consolidated data of all users. However, Sattler et al. [8] demonstrated that this assumption is frequently violated in practical FEEL applications. Concretely, it is shown that at every stationary point of the FEEL objective, the cosine similarity between the weight-updates of diverse clients can be used to infer the groups of clients having incongruent distributions. The main advantage of CFL, alongside multi-task FEEL [9] and Personalized Federated learning [10], is that CFL does not require any modifications of the core FEEL communications protocol as clustering is only performed after the FEEL model converges to a stationary point. CFL can be seen as a post-processing method that can improve FEEL performance by providing clients with more specialized models. Recently, the studies in [11]–[13] followed up on the CFL-based framework and confirmed that CFL is a more reliable choice to gain high accuracy than the conventional FEEL.

To address the resource heterogeneity, recent works [14]–[18] have attempted to address the problem of joint client

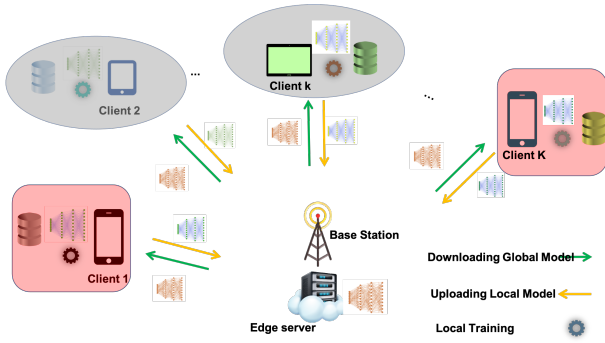


Figure 1: Clustered Multitask Federated Learning over wireless edge networks.

scheduling and resource allocation considering non-i.i.d data distribution. For instance, the authors in [14]–[18] proposed different algorithms to select the participants during FEEL rounds. The proposed selection algorithms choose the participants that either provide less updating time, have more powerful capabilities, or consume less energy, assuming that the number of data points is balanced for all devices. However, none of these works [14]–[18] consider the scheduling problem for the CFL approach, which initially needs to involve all active participating clients to perform local training updates and capture their local data distributions, making existing solutions inapplicable.

Spurred by the above observations, this work introduces a new client selection approach for CFL to minimize the training time, accelerate the convergence rate while tackling non-i.i.d and unbalanced data distribution problems. We propose an efficient client selection approach in which all clients have equal probability of participating in the training phase even if they have bad channels or low computational capabilities to obtain unbiased models. Specifically, we formulate an optimization problem and devise a client selection approach that achieves this objective. Experimental results verify that the proposed scheduling approach significantly minimizes the training time and accelerates the convergence rate.

The rest of the paper is structured as follows: we introduce the system, learning, computation, and communication models in Section II. The problem statement is formulated in Section III. Section IV presents the details of the proposed approach. Experimental results are presented in Section V. Finally, Section VI concludes this work and provides future directions.

II. PRELIMINARIES

A. System Model

As shown in fig. 1, we consider a FEEL across a set of K edge devices, $K = |\mathcal{K}|$, connected to one edge server which is equipped with a Base Station (BS). Each $k \in \mathcal{K}$ device holds a local dataset \mathcal{D}_k , with number of samples denoted by $D_k = |\mathcal{D}_k|$. Thus, the total data samples of all clients is $D = \sum_{k=1}^K D_k$. Each local dataset \mathcal{D}_k represents the collection of data samples with input-output pairs as $\{\mathbf{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^{D_k}$,

where $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ is an input vector with d features, and $y_i^{(k)} \in \mathbb{R}$ is the associated class-labeled output $\mathbf{x}_i^{(k)}$. Due to different device activity, the data distribution of each individual represents its local data, not the population; thus the \mathcal{D}_k 's is non-i.i.d and imbalanced. To train the model, each k -th participant runs its local solver, such as stochastic gradient descent (SGD), locally to minimize the loss function for several local epochs denoted by E . For example, if a minibatch SGD is used, the total number of local updates is defined as: $\mathbf{n} = E \frac{D_k}{b}$ where b is a batch-size determining a subset of the training set required for one local update.

B. FEEL model

The edge server aims to learn a shared global model that can perform well across the network by leveraging the data of the K connected devices. To perform this task, FEEL carries out the following steps every r -th round: 1) The server initializes random model parameters w^0 at the beginning of the training process and selects a subset \mathcal{S}^r of the clients to take part in the global model update by conducting local training, 2) The server broadcasts the latest global FEEL model parameters w^r to all selected participants, 3) The selected participants in \mathcal{S}^r upload their updated models. Last, the edge server aggregates and fuses all updates to form a new version of the global model. Formally, the edge server coordinates to fit the global model parameters $w^r \in \mathbb{R}^d$ to capture the associated output y_i , by iteratively minimizing a local loss function $f_i(w^r) = \ell(\mathbf{x}_i^{(k)}, y_i^{(k)}; w^r)$ in the r -th communication round. The local loss function is defined as: $F_k(w^r) := \frac{1}{D_k} \sum_{i \in \mathcal{D}_k} f_i(w^r)$. Accordingly, given datasets D_1, \dots, D_K distributed over K different clients, the goal in FEEL is to minimize the joint value of the loss function overall client's data $\mathcal{D} = \cup_k \mathcal{D}_k$ according to:

$$\min_w F(w, \mathcal{D}) := \frac{1}{D} \sum_{(x,y) \in \mathcal{D}} l(F_w(x), y) = \sum_{k=1}^K \frac{D_k}{D} \underbrace{F_k(w^r, \mathcal{D}_k)}_{\text{data local sub-problem}} \quad (1)$$

C. Computation and Communication Models

Due to limited bandwidth in FEEL, the BS cannot accept all existing clients as the number of sub-channels is limited. Thus, in conventional FEEL, only a subset of these devices indicated as elected set \mathcal{S}_r is selected every r -th round. Particularly, $\mathcal{S}_r = \{k \mid S_k^r = 1, k = 1, 2, \dots, K\}$, where $S_k^r = 1$ means that device k is selected and added to \mathcal{S}_r , otherwise $S_k^r = 0$. The round latency of each device $k \in \mathcal{S}_r$ consists of two parts, i.e., communication latency and computation latency. In general, all devices have the same model structure initialized by the coordinating server, with fixed-size w^r denoted by ζ . Without the loss of generality, we assume that the orthogonal frequency division multiple access (OFDMA) scheme is employed and in every r -th round, each k -th device is assigned an allocated bandwidth $\lambda_k^r B$ with one sub-channel where the system bandwidth is divided into N sub-channels equals to the number of selected clients. Thus, the achievable data rate for each k can be defined as: $r_k^r = \lambda_k^r B \ln \left(1 + \frac{P_k^r h_k^2}{N_0} \right)$, where h_k^r denotes the channel gain between device k and the BS, P_k^r denotes the k -th device's transmit power, and N_0 denotes

spectral power density (i.e., Gaussian noise). Therefore, the transmit latency can be defined as: $T_k^{trans} = \frac{\zeta}{r_k}$. On the other hand, the local computation latency of every device k to train its local model can be expressed as: $T_k^{cmp} = E \frac{\phi_k D_k}{f_k}$ where ϕ_k denotes the number of required CPU cycles to process one sample; $\phi_k D_k$ denotes the number of required CPU cycles to process all samples in r -th round; and f_k is the CPU frequency. Therefore, the total latency for every k -th device in the r -th round can be written: $T_k^{total} = T_k^{trans} + T_k^{cmp}$. In practice, the FEEL round latency, i.e., round deadline constraint, is determined by the latency of the slowest selected device $k \in \mathcal{S}_r$. Mathematically, we can define the round latency as: $T_r = \max\{T_k^{trans} + T_k^{cmp}\}$.

D. CFL Mechanism

As opposed to the standard FEEL framework where all clients are treated equally to train only one global model, the goal of CFL is to provide every client with a model that optimally fits its local data distribution. To consolidate the earlier exhibited problems with non-i.i.d and imbalanced data distributions, CFL [8] aims to generalize the conventional FEEL assumption to a set of clients having the same data distribution:

Assumption 1. ("CFL") [8]: *There exists a partitioning $\mathcal{M} = \{c_1, \dots, c_m\}$, $\bigcup_{i=1}^M c_i = \{1, \dots, K\}$ of the client population, in such a way that each subset of clients $c \in \mathcal{M}$ satisfies the standard FEEL assumption.*

where \mathcal{M} is a set of clusters in which each cluster includes the clients having the same data distribution (i.e., congruent data distribution) and $M = |\mathcal{M}|$ is the number of clusters. In CFL, the FEEL system can be seen as an exemplary parameter tree. At the root node resides the standard FEEL model, obtained by converging to a stationary point w^* of the FEEL objective. According to their cosine similarities, the client population is split up into two groups in the next layer, and each subgroup repeatedly converges to a stationary point w_0^* and w_1^* , respectively. Branching lasts recursively till no stationary solution meets the partitioning criteria. This is worth noting that the *cosine similarity*, *sim*, between the updates of any two clients k and k' is obtained by:

$$\begin{aligned} \text{sim}_{k,k'} &:= \text{sim}(\nabla F_k(w^*), \nabla F_{k'}(w^*)) := \frac{\langle \nabla F_k(w^*), \nabla F_{k'}(w^*) \rangle}{\|\nabla F_k(w^*)\| \|\nabla F_{k'}(w^*)\|} \\ &= \begin{cases} 1 & \text{if } I(k) = I(k') \\ -1 & \text{if } I(k) \neq I(k') \end{cases} \end{aligned} \quad (2)$$

where $I(k)$ and $I(k')$ are the local data distribution of k and k' , respectively. Therefore, an accurate bi-partitioning is given by: $c_1 = \{k | \text{sim}_{k,0} = 1\}$, $c_2 = \{k | \text{sim}_{k,0} = -1\}$. According to [8], the split should only perform if the following conditions are satisfied [8]:

$$0 \leq \left\| \sum_{k=1, \dots, K} \frac{D_k}{D_m} \nabla_w F_k(w^*) \right\| < \varepsilon_1 \quad (3)$$

Equ. (3) holds that the solution is closer to a stationary point of the FEEL objective. Conversely, the participating clients are

away from a stationary point of their local loss if the following condition is satisfied:

$$\max_{i=k, \dots, K} \|\nabla_w F_k(w^*)\| > \varepsilon_2 > 0 \quad (4)$$

where ε_1 , and ε_2 are hyperparameters to control the clustering task.

III. PROBLEM FORMULATION

In CFL, it is assumed that either all clients are selected from the beginning of the training as in [8] or only a random subset is selected at each round as in [19]. The former is impossible due to the limited resources at the wireless network edge. This latter will not capture all incongruent data distribution among clients, inducing a problem of efficiently designing a client selection approach while satisfying the edge wireless network constraints and achieving desired performance. Specifically, to adequately theorize the classical FEEL setting, it is crucial to split up clients with incongruent data distributions at the earliest to accelerate the convergence rate and reduce the communication costs considering the limited bandwidth and the data and device heterogeneity. Hence, we formulate an optimization problem to find the optimal scheduling for a set of clients (i.e., \mathcal{S}_r) that accelerates the convergence rate and minimizes the training time for all trained models (i.e., M groups models and one FEEL global model). Concretely, we use R to indicate the whole rounds during the training time budget T^{total} . We aim to find the optimal set of M models $w_m, m = 1, 2, \dots, M$ within T^{total} that minimize the global empirical risk function for each cluster as follows:

$$w_m \triangleq \arg \min_{w \in \{w_r^{\mathcal{S}_r} : r=1, 2, \dots, R\}} F(w_m). \quad (5)$$

To ease the presentation, we use $[R]$ to denote $\{1, 2, \dots, R\}$. The optimization problem for all clusters can be written as follows:

$$\min_{R, \mathcal{S}_{[R]}, T_{[R]}} T^{total} \quad (P1)$$

$$\text{s.t.} \quad F(w_m) - F(w_m^*) \leq \epsilon, \forall m \in \mathcal{M}, \quad (P1.0)$$

$$\sum_{r=1}^R T_r(\mathcal{S}_r) \leq T, \quad (P1.1)$$

$$T_k^{total} \leq T_r(\mathcal{S}_r), \quad (P1.2)$$

$$\mathcal{S}_r \subset \mathcal{K}, \forall r \in [R], \forall m \in \mathcal{M} \quad (P1.3)$$

$$|\mathcal{S}_r^m| \leq N, \forall r \in [R], \quad (P1.4)$$

where $\mathcal{S}_{[R]} = [\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_R]$ is the selected scheduling set for all rounds and $T_{[R]} = [T_1, T_2, \dots, T_R]$ is the maximum latency for each round (i.e., deadline constraint). Constraint (P1.0) guarantees that the convergence of each group model is satisfied. Constraint (P1.1) is set to assure that the total training time of all rounds doesn't exceed the time budget. At the same time, constraint (P1.2) ensures that the computation and communication time for every participating client is aligned with the deadline constraint. Constraint (P1.4) stipulates that the allocated bandwidth for all selected clients can not exceed the system bandwidth. To find the solution of

P1, we first need to know the effects of R and $\mathcal{S}_{[R]}$ on the loss function of each cluster model, i.e., $F(w_m)$. It is worth noting that finding an exact expression of $F(w_m)$ considering R and $\mathcal{S}_{[R]}$ is impossible. Thus, we tend to solve $F(w_m)$ iteratively in terms of R and $\mathcal{S}_{[R]}$. In addition, due to the variation of computation delay, T_k^{cmp} and wireless channel states h_k^r through the rounds R , attaining the optimal client selection approach is challenging and could be non-stationary. In reality, the FEEL process is iterative in nature and depends on the scheduled clients in the preceding rounds, making the bounding of $F(w_m)$ also very hard. To this end, we solve P1 in the next sections as follows. First, we extend the algorithm [8] considering the computation and communication resources. Then, we propose a fair client scheduling and resource allocation approach to solve the issues that emanate from having to schedule all clients in every round to avoid obtaining biased models.

IV. PROPOSED SCHEDULING ALGORITHM

In our proposed scheduling algorithm, all active and available clients during the training phase have equal priority to participate in the training. Thus, the server will collect preliminary information from these devices, such as data size, local CPU speed, and channel state; then, the server evaluates the expected latency of all clients to schedule the model uploading based on their finishing time. To aggregate the updates in an efficient manner, the server sorts the clients based on their expected latency in ascending order. At this stage, The number of aggregations that the server needs to collect all updates in one round can be defined as:

$$ng = \frac{|\mathcal{S}_{[r]}|}{N} \quad (6)$$

As some participating clients complete updating and uploading tasks, other clients can perform their updates while the former clients are in the upload phase. Therefore, the total time budget is efficiently utilized as the server can handle many updates subsequently. The aggregation set can be defined as:

$$\mathcal{G} = \{\{1 + (N + (j - 1)), \dots, N + (N + (j - 1))\}, j = 1, 2, \dots, ng\} \quad (7)$$

It is worth noting that the server is assumed to have far greater computational resources than the participating clients; thus, the cost related to this task will typically be negligible. Our scheduling problem has a trade-off between the number of clients selected in a specific round and the total number of rounds handled in a particular duration, which can be controlled by T_r . If we initiated long T_r , the number of participating clients in $\mathcal{S}_{[r]}$ can increase, which is desired for CFL at the round and can expedite the splitting and improve the performance of the group models. Moreover, longer T_r decreases the number of rounds performed in a particular duration, which can catch the incongruent degree of data distribution. Thus, we should carefully select T_r . At each round, the server uses (2), (3), and (4) to partition and cluster the clients. In such a case, we have the similarities between

Algorithm 1 Client Selection Strategy for CFL

Input: K (number of clients), initial parameters w_0 , controlling parameters $\varepsilon_1, \varepsilon_2 > 0$, number of local epochs E , batch size b
Output: M reliable models, one FEEL global model
Initialization: set initial clusters $\mathcal{M} = \{\{1, \dots, K\}\}$, initial models $w_k \leftarrow 0 \forall k$, set $r = 1$

```

1: while Not converge do
    // Server Task Preprocessing
2:   Server collects prior information from all active and available clients
3:   if  $M > 1$  then
4:     Server selects the best clients for all  $M - 1$  clusters while selects
       all client in last cluster,  $M$ , from the cluster set  $\mathcal{M}$ 
5:   else
6:     Server selects all active and available clients
7:   end if
8:   Server estimates the latency of all clients and sort them in ascending
       order to schedule uploading the updates
9:   Server finds the aggregation set using (6) and (7)
10:  Server broadcasts  $w_{r-1}$  to all clients
    // Participating Clients Task in parallel
11:  for  $k = 1$  to  $|\mathcal{S}_{[r]}|$  do
12:    Receive  $w_{r-1}$  from the server
13:    Perform local updates  $w_k = w^{r-1} - \eta \sum_{i=1}^n \nabla F_k^r(w_k(i))$ 
14:  end for
15:  Server aggregates all updates as ordered in  $\mathcal{G}$  (i.e., Equ (7))
    // Server Task Post-processing
16:  Server sets  $\mathcal{M}_{tmp} \leftarrow \mathcal{M}$ 
17:  Server computes  $F(w^r) = \sum_{k=1}^K \frac{D_k}{D} F_k(w^r)$  and  $w^r = \sum_{k=1}^K \frac{D_k}{D} w^r$ 
18:  for  $c \in \mathcal{M}$  do
19:    • Server finds  $\Delta w_m \leftarrow \frac{1}{|c|} \sum_{k \in c} \Delta w_k$ 
20:    if  $\|\Delta w_c\| < \varepsilon_1$  and  $\max_{k \in c} \|\Delta w_k\| > \varepsilon_2$  then
21:      •  $sim_{k,k'} \leftarrow \frac{\langle \Delta w_k, \Delta w_{k'} \rangle}{\|\Delta w_k\| \|\Delta w_{k'}\|}$ 
22:      •  $c_1, c_2 \leftarrow \arg \min_{c_1 \cup c_2 = c} (\max_{k \in c_1, k' \in c_2} sim_{k,k'})$ 
23:      •  $sim_{cross}^{max} \leftarrow \max_{k \in c_1, k' \in c_2} sim_{k,k'}$ 
24:      •  $\gamma_k := \frac{\|\nabla F_{I(k)}(w^*) - \nabla F_k(w^*)\|}{\|\nabla F_{I(k)}(w^*)\|}$ 
25:      if  $max(\gamma_k) < \sqrt{\frac{1 - sim_{cross}^{max}}{2}}$  then
26:        •  $\mathcal{M}_{tmp} \leftarrow (\mathcal{M}_{tmp} \setminus c) \cup c_1 \cup c_2$ 
27:      end if
28:    end if
29:  end for
30:  •  $\mathcal{M} \leftarrow \mathcal{M}_{tmp}$ 
31:   $r = r + 1$ 
32: end while
33: Server Return  $M$  reliable models, one FEEL global model

```

participating clients k and k' within a cluster which can be bounded as:

$$sim_{within}^{min} = \min_{\substack{k, k' \\ I(k)=I(k')}} sim(\nabla w r_k(w^*), \nabla w r_{k'}(w^*)) \quad (8)$$

While the similarities between participating clients k and k' across different clusters (i.e., different data distributions) can be bounded as:

$$sim_{cross}^{max} = \max_{k \in c_1^*, k' \in c_2^*} sim(\nabla w r_k(w^*), \nabla w r_{k'}(w^*)) \quad (9)$$

We should emphasize that (8) and (9) are used to estimate the separation gap as follow:

$$g(sim) := sim_{intra}^{min} - sim_{cross}^{max} \quad (10)$$

The steps of the proposed client selection approach are listed in Algorithm 1. In Alg. 1, we can see that once a particular cluster reaches the stationary point of FEEL objective (lines 2-6), the server uses greedy scheduling for such clusters to

choose clients providing less training time to perform further updates. This stems from that the clustered clients, in this case, have congruent distribution.

V. NUMERICAL EXPERIMENTS

In this section, we present the experimental setup, the simulated scenarios, and the numerical results. We consider a total bandwidth of $B = 10$ MHz, and the bandwidth of each sub-channel is 1 MHz. We model a random wireless channel gain h_k^r for each participant with a path loss ($\mu = g_0(\frac{d_0}{d})^4$) where $g_0 = -35$ dB and the reference distance $d_0 = 2$ m. The distances between edge devices and the edge server are distributed uniformly between 20 and 100 m. Also, Additive White Gaussian Noise (AWGN) power is set to $N_0 = 10^{-6}$ watt. The transmit power P_k^r is randomly distributed between $p_{min} = -10$ dBm and $p_{max} = 20$ dBm. The CPU frequency f_k is randomly distributed between 1 GHz and 9 GHz, and the number of cycles ϕ required to process one sample is set to 20. We use a federated dataset called, **FEMNIST** [20], for classifications tasks. FEMNIST is used for the classification task of both letters and digits (A-Z, a-z, and 0-9), and it has 244,154 images for training and 61500 for testing. FEMNIST dataset [20] was specifically designed to model the challenges of realistic FEEL problems. We use this dataset under non-i.i.d data distribution, where the dataset is first split into 62 partitions, and then each user is assigned batches of two classes only. We use the convolutional neural networks (CNN) classifier to train the models.

To evaluate the performance of the proposed scheduling algorithm, we compared the results of the proposed algorithm against the random scheduling CFL algorithm (baseline algorithm as in [8], [19]). To ensure a fair comparison, we use a similar setting in terms of several local updates, learning rate, model structure, data distribution (i.e., non-i.i.d and imbalanced data distribution) for both scheduling algorithms. We set R , the communication rounds, to 200 and the number of epochs E to 10 for all clients. We run the experiments to training different models using both algorithms, and we train the models over 100 users and test the resulting models using the test data among 15 clients.

Fig. 2 shows the performance of the proposed approach, and the baseline algorithms in terms of accuracy bounded with the confidence interval at every round (Figs. 2a and 2b, left) and the gradient norms to exhibit the convergence rate (Figs. 2a and 2b, right). One can see that the proposed algorithm gains much better accuracy with a faster convergence rate as the splitting started at round 37 while the split using the baseline algorithm started at round 83, accelerating the convergence rate by more than 50%. It is observed that, in the proposed algorithm, the partitioning is performed again for all clients with incongruent distributions at rounds 45 and 63 till all clusters were split. From Fig. 2a, right, we can note that all trained models reached the stationary point at round 190, indicating that clustering is finalized. At the same time, From Fig. 2b, right, the baseline algorithm, we can clearly observe that the FEEL still needs more rounds to converge to the

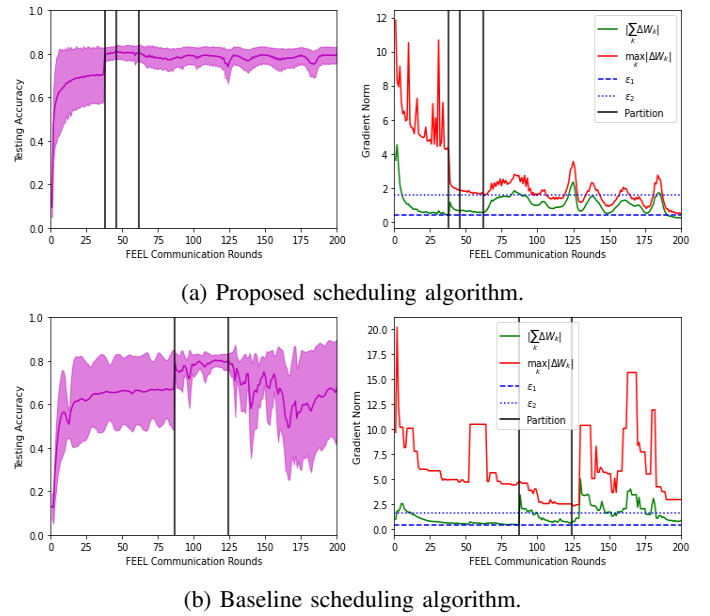


Figure 2: Average accuracy of the resulting model(s) using our proposed scheduling algorithm, top: (a: left), and using the baseline scheduling algorithm, Bottom: (b:left). In (a: right) the Gradient norm of global and local risk functions using the proposed algorithms and (b:right) using the baseline algorithm. Results are collected and averaged over five trials.

stationary point at round 200. This stems from the behavior of random client selection, which might either select clients that have been clustered in former rounds or select the clients that exceed the deadline.

From the presented results in Table I, one can observe that the training process using both scheduling algorithms has resulted in multiple models for every client: A FEEL base model as well as more specialized models for the different clusters. Note that the proposed algorithm produces three more specialized models. From table Ia, one can notice that, in the results of the proposed algorithm, all clients attained satisfying accuracy (highlighted in green), and the performance gap is about 10% only. In contrast, in table Ib, about $\frac{1}{3}$ of the tested clients (clients 1, 2, 3, 4, and 5) attained bad accuracy (highlighted in red). The performance gap reaches up to 30.4%, indicating that some models are still biased to those frequently selected clients. Overall, our results demonstrate that the proposed scheduling algorithm outperforms the baseline scheduling algorithm method in terms of clustering quality not only with more specialized models but also on accelerating the convergence rate and reducing the costs.

VI. CONCLUSION

This work introduces a novel client selection approach for CFL to minimize the training time and accelerate the convergence rate. We first formulate an optimization problem to find the optimal client selection approach that reduces the training time and accelerates the convergence rate considering non-i.i.d

Table I: TESTING ACCURACY OF ALL RESULTING MODELS: A FEEL BASE MODEL AS WELL AS MORE SPECIALIZED MODELS FOR THE DIFFERENT CLUSTERS, C*=CLIENT.

(a) Proposed scheduling algorithm.

	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10	C 11	C 12	C 13	C 14	C 15
FEEL Model	46	37.9	45.2	44.8	37.3	75.5	80.7	83.2	79	79	77.3	76.6	77.9	76.4	77.5
Model 1	0	0	0	0	0	77.1	83.3	86	81.6	82	81.8	81.5	82.5	79.6	80.9
Model 2	0	0	81.6	76	77.6	0	83.8	85.5	0	82.7	80.1	0	84.5	82	81.8
Model 3	83.3	0	77	67	77.9	76.8	0	0	0	82.7	0	0	0	0	0
Model 4	0	74.5	0	0	0	0	82.5	83	75.2	0	0	0	0	0	0
Model 5	83.3	78.6	0	0	0	76.8	0	0	75.2	0	0	76.8	0	0	0
Model 6	83.3	74.5	76.9	67.2	77.9	0	0	0	0	0	0	0	0	0	0
Max Acc	83.3	78.6	81.6	76	77.9	77.1	83.8	86	81.6	82.7	81.8	81.5	84.5	82	81.8

(b) Baseline scheduling algorithm.

	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10	C 11	C 12	C 13	C 14	C 15
FEEL Model	50.7	41.6	49.2	46.4	41.5	73.1	80.1	82.1	76.5	77.8	77	77.3	76.7	76.1	77.7
Model 1	0	0	0	0	0	78.5	84.5	86.5	82.1	81.4	81.5	79.9	81	79.9	81.1
Model 2	0	0	0	52.4	56.5	0	78.3	75.8	77.9	72.7	0	0	78.7	78.2	67.4
Model 3	59.7	61.1	62.5	0	0	82	0	0	0	0	82.8	67.9	0	0	0
Model 4	59.7	61.1	62.5	52.4	56.5	0	0	0	0	0	0	0	0	0	0
Max Acc	59.7	61.1	62.5	52.4	56.5	82	84.5	86.5	82.1	81.4	82.8	79.9	81	79.9	81.1

and imbalanced data distribution, device heterogeneity, and limited bandwidth. The proposed client selection approach is fair to all clients, i.e., all clients have an equal probability of participating in the training phase to obtain unbiased models even if they have bad channels or few data points. We carry out extensive simulation experiments using a realistic federated dataset. The results verify that the proposed solution effectively accelerates the convergence rate by up to 50% and minimizes the training time while attaining a more satisfying performance for each client with an optimal fit model for its local data distribution. Finding the optimal threshold values for splitting the clusters will be considered in future studies.

ACKNOWLEDGEMENT

This publication was made possible by NPRP-Standard (NPRP-S) Thirteen (13th) Cycle grant # NPRP13S-0201-200219 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work and are solely the responsibility of the authors.

REFERENCES

- [1] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.
- [2] C. Lu, W. Xu, H. Shen, J. Zhu, and K. Wang, "MIMO channel information feedback using deep recurrent network," *IEEE Communications Letters*, vol. 23, no. 1, pp. 188–191, 2018.
- [3] A. M. Albaser, M. Abdallah, A. Al-Fuqaha, and A. Erbad, "Fine-grained data selection for improved energy efficiency of federated edge learning," *IEEE Transactions on Network Science and Engineering*, 2021.
- [4] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.
- [6] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.
- [7] A. Albaser, M. Abdallah, A. Al-Fuqaha, and A. Erbad, "Threshold-based data exclusion approach for energy-efficient federated edge learning," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2021.
- [8] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [9] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Advances in Neural Information Processing Systems*, pp. 4424–4434, 2017.
- [10] M. Zhang, K. Sapra, S. Fidler, S. Yeung, and J. M. Alvarez, "Personalized federated learning with first order model optimization," *arXiv preprint arXiv:2012.08565*, 2020.
- [11] M. Xie, G. Long, T. Shen, T. Zhou, X. Wang, and J. Jiang, "Multi-center federated learning," *arXiv preprint arXiv:2005.01026*, 2020.
- [12] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 586–19 597. Curran Associates, Inc., 2020.
- [13] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2020.
- [14] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1188–1200, 2020.
- [15] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 453–467, 2020.
- [16] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Transactions on Communications*, 2019.
- [17] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1–6. IEEE, 2020.
- [18] M. M. Wadu, S. Samarakoon, and M. Bennis, "Federated learning under channel uncertainty: Joint client scheduling and resource allocation," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6. IEEE, 2020.
- [19] M. Duan, D. Liu, X. Ji, R. Liu, L. Liang, X. Chen, and Y. Tan, "Fedgroup: Ternary cosine similarity-based clustered federated learning framework toward high accuracy in heterogeneous data," *CoRR*, vol. abs/2010.06870, 2020. [Online]. Available: <https://arxiv.org/abs/2010.06870>
- [20] S. Caldas, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.