IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 June 2024, accepted 8 July 2024, date of publication 15 July 2024, date of current version 23 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3427694

RESEARCH ARTICLE

Multi-Head DNN-Based Federated Learning for RSRP Prediction in 6G Wireless Communication

MENGHAN YU, XIONG XIONG, ZHEN LI, AND XU XIA, (Member, IEEE)
6G Research Center, China Telecom Research Institute, Beijing 102209, China

Corresponding author: Menghan Yu (yumh1@chinatelecom.cn)

This work was supported by the 2020 National Key Research and Development Program "Broadband Communication and New Network" Special "6G Network Architecture and Key Technologies" under Grant 2020YFB1806700.

ABSTRACT In the realm of wireless communications, accurate Radio Signal Received Power (RSRP) prediction serves as the foundation for improving user experience and optimizing network efficiency and reliability. With the deep integration of Artificial Intelligence (AI) technology and the wireless communication network, Federated Learning (FL) is considered as a promising approach for enhancing RSRP prediction while protecting user data privacy in the upcoming of 6G network. However, in practice, the heterogeneity of User Equipment (UE) environments and the limitations of UE communication bandwidth and computational capabilities can lead to poor model performance and inefficient model interactions in FL. To address these challenges, this paper proposes a Multi-head DNN based FL algorithm for RSRP prediction. The experimental results show that the proposed algorithm can enhance both RSRP prediction performance and communication efficiency.

INDEX TERMS Wireless communication, RSRP prediction, federated learning, 6G network.

I. INTRODUCTION

The rapid evolution of communication network has witnessed remarkable achievements with the advent of Fifth-generation (5G) and Fifth-generation Advanced (5GA) [1]. As the 5G landscape continues to flourish, the industry has already embarked on the journey towards the Sixth-generation (6G) [2]. As an important metric in cellular networks, RSRP prediction also plays a crucial role in ensuring system robustness and efficiency in the future 6G network. Due to network densification, environment complexity and UE mobility, RSRP typically exhibits a high degree of randomness. The accurate estimation and prediction of RSRP have far-reaching implications, encompassing network optimization, resource allocation, handover management and overall network performance enhancement [3]. The lack of accurate RSRP prediction may lead to issues such as network performance degradation, resource wastage, and deterioration of user experience. Thus, RSRP estimation and prediction is a fundamental and typical use case in wireless communication networks.

To meet the requirements brought by diversified emerging network services, the 6G wireless network is envisioned to be integrated communication with AI. The computational capabilities will be ubiquitous throughout the network including end devices and the edge side of wireless communication network [4]. By leveraging the capabilities of AI, the accuracy and efficiency of RSRP prediction can be further improved. In recent years, in order to facilitate the intelligent evolution and the deployment of AI algorithms in the radio access network (RAN), the open-radio access network (O-RAN) alliance designed the Radio Intelligent Controller (RIC) platform, a key element in RAN architecture to achieve monitoring and control of wireless network [5]. RIC platform can be classified into non-real-time RIC (non-RT RIC) and near-real-time RIC (near-RT RIC) based on the processing latency characteristics. The near-RT RIC uses E2 interface to communicate with Base Stations (BSs) and can apply AI technology to use cases through centralized data processing and model training and inference [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Casoni.

The future trend in network deployment involves utilizing AI technology for RSRP prediction, which is essential in 6G network. Many studies have focused on the AI based RSRP estimation and prediction [7], [8], which predominantly rely on centralized AI methods. However, the conventional centralized data collection and model training methods struggle to fulfill the stringent demands of ultra-low latency and consumer data privacy. Similar to the conventional centralized training method, centralized RSRP prediction also faces challenges such as high latency and poor privacy protection. Hence, the distributed machine learning approaches are poised to find extensive utilizations in 6G network where computational capabilities will be widely dispersed among UEs and network and heightened focus will be placed on user data privacy. As a prominent representative of distributed machine learning approach, federated learning has garnered extensive attention in academia and field of communications due to its notable advantages [9]:

- Privacy preservation: During the process of FL, the clients update the model using their local data and upload the updated model instead of their private raw data.
- Low latency: Since both model training and inference decisions take place at the edge side in FL, latency can be significantly reduced to some extent.
- High resource utilization: Unlike centralized training, FL efficiently leverages the computational, storage, and communication resources of UE and network edges.

Recently, there has also been researches proposing the use of federated learning to train wireless signal strength including RSRP prediction models [10], [11], which primarily emphasize the privacy preserving characteristic of FL. However, in addition to user privacy preserving, it is crucial to focus on addressing the performance degradation of federated models due to disparate data distributions among different users and the feasibility of deploying complex models in real-world network. In mobile communication network, one of the primary reasons for varying user data distributions is the differing geographical locations of users. This can potentially result in non-independent and identically distributed (non-IID) data among member UEs. A simple solution is to use multiple DNNs, each corresponding to a geographic area. However, using multiple models in FL scenario demands considerably higher computational sources. The practical deployment feasibility of FL algorithms should primarily consider the overall communication bandwidth and the complexity of machine learning (ML) model because of the limited communication and computing resources of UEs.

This paper takes into consideration the capabilities of near-RT RIC, including AI model inference, decision-making, QoS management, and its low processing latency. Consequently, we proposed the FL framework based on near-RT RIC in wireless communication leveraging the near-RT RIC as the central server in the context of FL to aggregate the local updates uploaded by FL member UEs. And to tackle the performance degradation problems while minimizing communication and computing costs as much as possible, we propose a novel FL approach within the framework of FL called Multi-head DNN based federated learning (abbreviated as Multi-head FL). The main contributions of this paper are summarized as follows:

- We design a federated learning framework based on the O-RAN architecture, which leverages near-RT RIC with robust AI capabilities and low latency to enhance the efficiency of federated learning.
- We propose a novel FL approach, called Multi-head DNN based federated learning (Multi-head FL), not only to address the global model performance degradation in RSRP prediction caused by variations in geographical locations and environmental differences among different user data, but also to reduce the overall communication costs.
- We compare the RSRP prediction performance and the overall communication costs of the proposed Multi-head FL algorithm with the federated averaging (FedAVG) algorithm [12]. The results indicate that the Multi-head FL algorithm can reduce the global test loss by up to 38.6%, and can reduce communication costs by up to 62.7% compared to FedAVG.

The rest of the paper is organized as follows. Section II presents some existing research on RSRP prediction and FL. We describe the RSRP prediction scenario in Section III. Section IV recalls the principles of FedAVG firstly, and then illustrates the details of our proposed Multi-head FL approach. The experimental results and comparative analysis are presented in Section V. In Section VI, we conclude this paper.

II. RELATED WORKS

Early RSRP prediction methods involved channel modeling based on theories such as electromagnetic propagation, geometric optics, and uniform diffraction. These methods rely on extensive empirical data to fit the relevant parameters. The most representative methods include Cost 231-Hata, Okumura, Volcano, etc. [13], [14], [15] But the traditional methods are constrained by complex propagation environment and require a substantial amount of additional measured data to correct the propagation model. In contrast, employing ML methods for RSRP prediction can reduce system complexity and unearth hidden features within the measured data. Many studies have been dedicated to using data-driven ML model for RSRP prediction [16], [17], [18].

However, all of the above RSRP prediction methods are based on the centralized intelligence. In the recent past, given the concern over user privacy, few studies have leveraged federated learning technology into RSRP prediction. The author in [19] utilizes UEs' location information in the context of FL to predict RSRP and brings a privacy-preserving approach with differential privacy (DP) against possible privacy attacks. Reference [20] compares several so-called

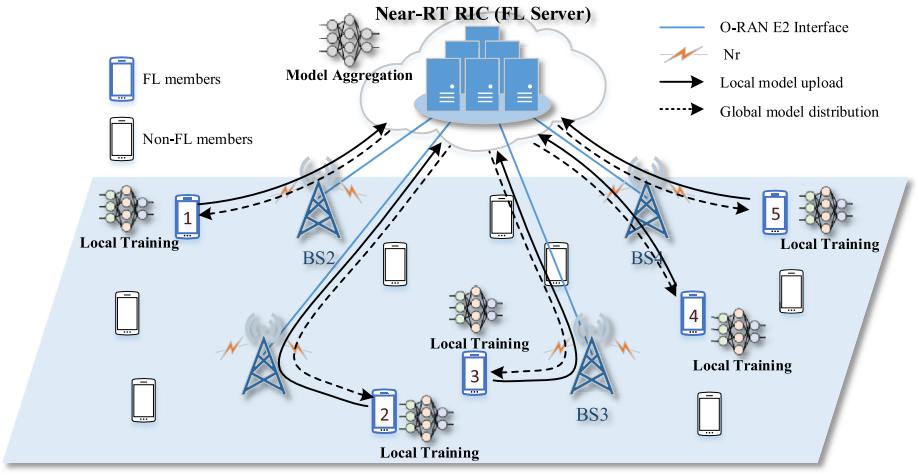


FIGURE 1. Reference system architecture of FL in wireless network with RIC platform as central server.

FL-native mechanisms including FedAVG and Federated SGD (FedSGD) and additional privacy preserving techniques such as DP and Secure Aggregation on two different datasets for RSRP prediction. However, while privacy-preserving is important in future wireless network FL, it is also crucial to address the performance degradation of FL model caused by the differences in user data characteristics, such as geographical location and environment factors in practical RSRP prediction scenarios. Additionally, challenges related to the feasibility of algorithms due to limited capabilities in terms of UEs' communication and computing should be given due consideration.

Due to the broad applicability of FL, there is a significant level of interest and rapid research progress in the academic field. Reference [12] introduced FedAVG, one of the most commonly used algorithms in federated learning, and it serves as the primary object of comparison in this paper. In this context, we also present some other relevant and outstanding FL algorithms. Reference [21] proposed Local Global Federated Averaging (LG-FedAVG) that combines local models with a global model, which not only reduces communication costs but also enhances flexibility in handling heterogeneous data. But this method requires each FL member to maintain an additional local model, which may have limited practicality for UE. Reference [22] proposed Multi-model FedAVG (Multi-FedAVG) and the experiment proves that the performance is not worse than that of a single model FedAVG. But for multi-model settings, the slow global convergence speed may increase the overall communication costs.

Unlike multiple DNNs, multi-head DNN is a single DNN augmented by multiple head models. The multi-head DNN can adapt multi-feature learning tasks, mitigate overfitting, and enhance overall model performance. Moreover, compared to multiple DNNs, multi-head DNN can significantly save computational resources [23], [24], [25].

Based on an analysis of many existing FL algorithms and drawing inspiration from the multi-head DNN, we believe that the integration of FL with Multi-head DNN can effectively address all of the aforementioned challenges. In this paper, we propose a Multi-head FL algorithm and validate the effectiveness through experimental results. The results show that the Multi-head FL algorithm can reduce the global test loss by up to 38.6%, and can reduce communication costs by up to 62.7% compared to FedAVG.

III. SYSTEM ARCHITECTURE AND PROBLEM FORMULATION

This section introduces the reference system architecture of FL in this paper and provides a detailed description of the RSRP prediction problem addressed in this paper.

A. SYSTEM ARCHITECTURE

Currently, most FL architecture in wireless communication system are implemented with the base stations serving as the FL servers [9]. However, FL typically involves a wide scope and a large number of users where different users may be connected to different base stations. If we select one of the connected base stations serving as FL server, it would involve frequent interaction between multiple base stations to transfer model weights for every epoch of FL, leading to inefficient communication in the entire FL procedure. Therefore, this paper proposes using near-RT RIC as the FL server to improve the communication efficiency. During the FL, all member users directly upload their locally updated model weights to near-RT RIC through the E2 interface between the connected base station and near-RT RIC, eliminating the need for model weights interaction between different base stations.

As illustrated in Figure 1, an O-RAN architecture based FL framework is considered, which consists of a near-RT RIC, users and base stations. With the capabilities of AI model inference, decision-making, QoS management, and

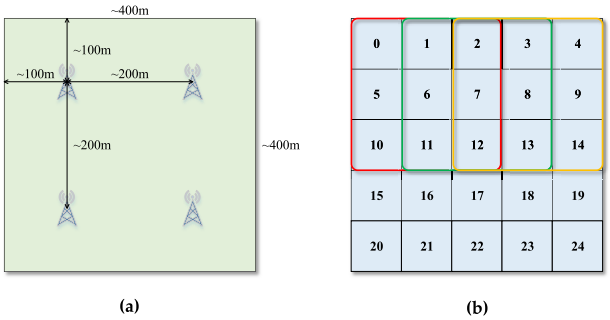


FIGURE 2. (a) Schematic map of the experimental area. (b) Schematic diagrams of sub-areas and grouped areas.

low-latency processing, the near-RT RIC plays the role as the central server in the context of FL to aggregate the local updates uploaded by FL member UEs and distribute the aggregated global model to the members of next round. The base stations connect to the near-RT RIC via the O-RAN E2 Interfaces, and provides the wireless access capability for UEs. As the FL members, some UEs are selected to perform local training in each round of FL training, and send their updated local model weights to the near-RT RIC through base station forwarding.

B. PROBLEM FORMULATION

There are total $M = 90$ users and $B = 4$ base stations in a specific 400m*400m area, as shown in Figure 2a. Each user's raw training dataset is sourced from Huawei ModelArts platform, and shares the same features and labels. The raw data of i-th user can be represented as follow:

$$D_i = \{(x_i^k, y_i^k; r_i^{k,1}, r_i^{k,2}, r_i^{k,3}, r_i^{k,4}), k = 1, \ldots, K_i\} \quad (1)$$

where k denotes the k-th sample in a local training dataset, and K_i denotes the total number of samples of the i-th user's local training dataset. Each raw data sample consists of a 6 dimensional array. The input features $\mathbf{l}_i^k = [x_i^k, y_i^k]$ consist of two dimensions, representing the horizontal and vertical coordinates of the measurement point within the area. The labels $\mathbf{r}_i^k = [r_i^{k,1}, r_i^{k,2}, r_i^{k,3}, r_i^{k,4}]$ include four dimensions, representing the RSRP values between the user and the four base stations at the coordinate.

In order to better capture the distributed features and the spatial heterogeneity within mobile communication networks, the specific area is divided into 25 sub-area, each identified by numbers 0 through 24. The 90 users are divided into 9 different groups. To simplify the problem and emphasize the regional characteristics of user data, it is assumed that the 10 users in the same group will only move within a fixed subset of the sub-areas. Each group of users is confined to a square moving area of 240m*240m, which includes 9 sub-areas. The user groups with adjacent group number have their moving areas arranged sequentially.

For example, the first group of users will only move within the red area in Figure 2b (the sub-area numbers 0, 1, 2, 5, 6, 7, 10, 11, 12) while the second group of users will only move within the green area in Figure 2b (the sub-area numbers 1, 2, 3, 6, 7, 8, 11, 12, 13), and so forth. Due to the variations of the moving area, the number of data samples for each user slightly differs.

TABLE 1. Detailed user grouping and the corresponding sub-area numbers for each user group.

User Group ID	User ID	Moving sub-areas
0	0~9	0,1,2,5,6,7,10,11,12
1	10~19	1,2,3,6,7,8,11,12,13
2	20~29	2,3,4,7,8,9,12,13,14
3	30~39	5,6,7,10,11,12,15,16,17
4	40~49	6,7,8,11,12,13,16,17,18
5	50~59	7,8,9,12,13,14,17,18,19
6	60~69	10,11,12,15,16,17,20,21,22
7	70~79	11,12,13,16,17,18,21,22,23
8	80~89	12,13,14,17,18,19,22,23,24

The detailed grouping and sub-area information is provided in Table 1. For each user, the input coordinates of data sampling points are obtained by randomly scattering points within the moving sub-areas. Due to environment factors, users cannot guarantee reachability with every base station. Each user has at least one reachable base station, and they may have multiple (up to 4) reachable base stations. For non-reachable base stations, the corresponding RSRP is set to the integer 0.

According to the local datasets of UEs, the RSRP distribution map of all user samples for the 4 base stations are depicted in Figure 3 respectively, where the white areas represent the area where connectivity to the corresponding base station is not available and the red spots are considered as outliers or noisy data points which will be removed during the subsequent data processing. Moreover, by preliminary analysis of the datasets, there are 28 numbered grey boxes considered as buildings in the area, where no samples are located at these positions. From Figure 3, it can be observed that the majority of buildings in the area are concentrated in the central region. There are few buildings at the periphery, leading to higher and more evenly distributed RSRP strengths. In particular, the central sub-area number 12 and the sub-area number 22 exhibit the most complex RSRP distributions. Therefore, we focused on these two sub-areas in the subsequent experiments.

IV. MULTI-HEAD FL ALGORITHM-BASED SOLUTION

In this section, we introduce the details of the deployment of FedAVG and Multi-head FL algorithm within the proposed framework.

A. FEDAVG ALGORITHM

In FedAVG, the local loss function of the i-th user is as follow:

$$F_i(\boldsymbol{\omega}) = \frac{1}{d_i} \sum_{k \in D_i} f(\boldsymbol{\omega}_i, \mathbf{l}_i^k, \mathbf{r}_i^k) \quad (2)$$

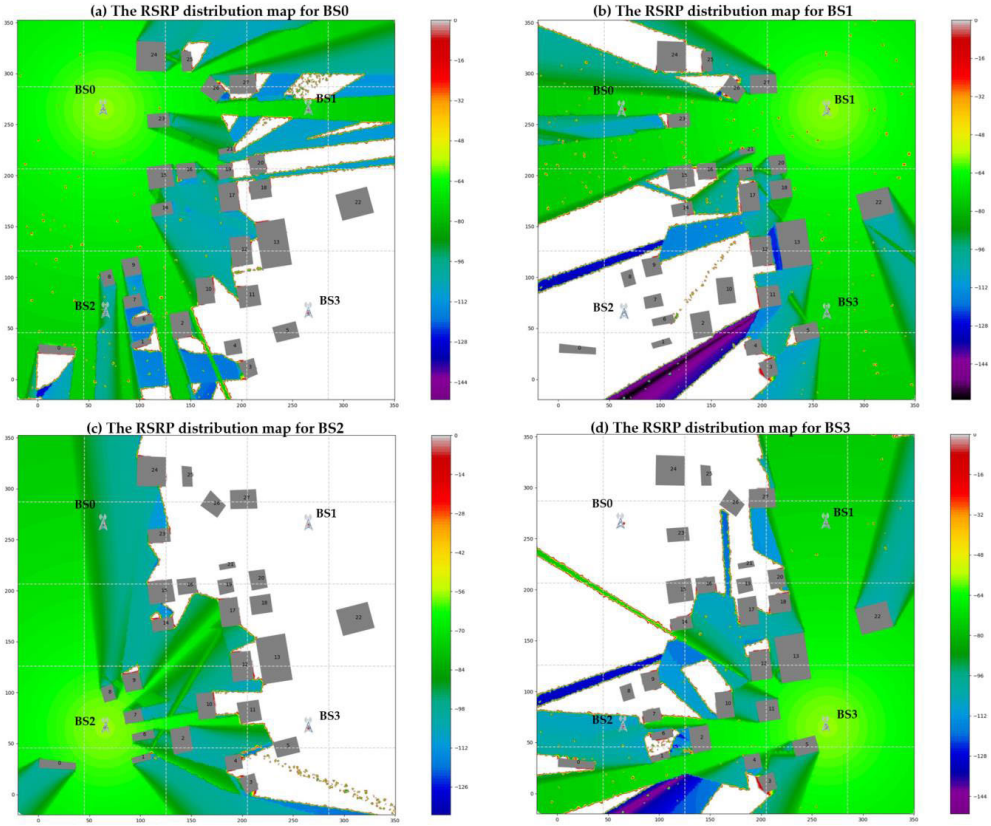


FIGURE 3. Schematic representation of the RSRP distribution map in the training datasets for (a) Base station 0. (b) Base station 1. (c) Base station 2. (d) Base station 3.

where $d_i = |D_i|$ is the total number of data samples of the i-th user. The $f(\boldsymbol{\omega}, \mathbf{l}_i^k, \mathbf{r}_i^k)$ is the error for input-output pairs given the model parameters ω. The overall objective is to find the optimal global parameters which minimizes the loss function on the whole training datasets $D = \{D_i, i = 1, 2, \ldots, M\}$, which is given by:

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}} J(\omega) \tag{3}$$

where the global loss function $J(\boldsymbol{\omega})$ can be written as:

$$J(\omega) = \frac{1}{d} \sum_{i=1}^{M} d_i F_i(\boldsymbol{\omega}) \tag{4}$$

where $d = \sum_{i=1}^{M} d_i$ is the total number of data samples of the M users. In the t-th round of learning, the near-RT RIC first selects $m(m \leq M)$ member users and distributes the global model ω^t to these members. The set of member users is denoted as S^t. Then the member users locally update the model for local epoch E_l using stochastic gradient descent (SGD):

$$\boldsymbol{\omega}_i^{t,j+1} = \boldsymbol{\omega}_i^{t,j} - \eta \nabla F_i(\boldsymbol{\omega}^{t,j}) \text{ for } j = 0, 1, \ldots, E_l - 1 \tag{5}$$

where η is the learning rate. After local training, the user upload the updated model to the near-RT RIC. When the near-RT RIC receives all updated model, it aggregate them by applying:

$$\boldsymbol{\omega}^{t+1} = \sum_{i=1}^{m} \frac{d_i}{d} \boldsymbol{\omega}_i^{t+1} \tag{6}$$

$\boldsymbol{\omega}_i^{t+1,E_l}$ given in (5) is recorded as $\boldsymbol{\omega}_i^{t+1}$ and uses in (6). The overall process of FedAVG deployed in the proposed framework can be referenced in Algorithm 1.

B. MULTI-HEAD FL ALGORITHM

Unlike the FedAVG algorithm, in our proposed Multi-head FL algorithm, the global model employs a Multi-head DNN as shown in Figure 4. The Multi-head DNN comprises a Backbone network along with multiple head networks. All the head networks have the same output dimension, which is consistent with the labels. The Input-to-Head Mapper module is used to determine which head network to connect based on the input features during the forward propagation.

Considering the overall communication overhead, the first step of the entire Multi-head FL involves the near-RT RIC preconfiguring all candidate users and determining the corresponding one or more head networks for each user. This step can be completed by collecting relatively low-privacy-level information from the users.

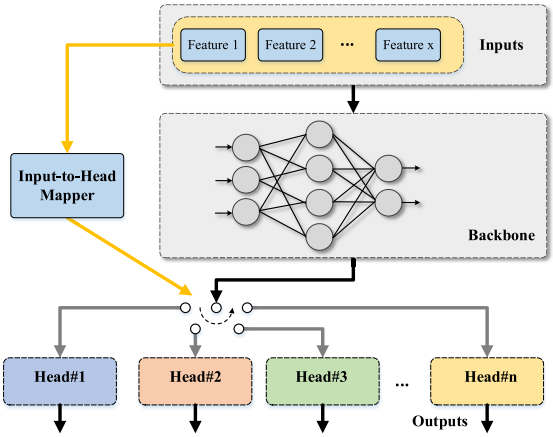


FIGURE 4. Schematic diagram of Multi-head DNN.

Algorithm 1 Pseudo-code for Federated Averaging (FedAVG) deployed in the pro-posed framework.

Input: Number of users, M; The training datasets, $D = \{D_i, i = 1, 2, \ldots, M\}$; Global Epoch, E; Local Epoch, E_l; Learning rate, η;
Output: The global model $\boldsymbol{\omega}^E$.

```
Initialize ω^0
for t = 0, 1, ..., E-1 do
    near-RT RIC determine m
    near-RT RIC determine S^t
    for each user in S^t do
        ω_i^{t,0} ← ω^t
        for j = 0, 1, ..., E_l-1 do
            ω_i^{t,j+1} ← ω_i^{t,j} − η∇F_i(ω_i^{t,j})
        ω_i^t ← ω_i^{t,E_l}
    near-RT RIC do
        ω^{t+1} ← Σ_{i∈S^t} (d_i/d) ω_i^t
near-RT RIC output the final global model ω^E
```

In the initial phase of federated learning, the near-RT RIC first determines the number of head network N and establishes the mapping relationships between inputs and head networks. Furthermore, considering that the backbone network typically contributes significantly to both uplink and downlink communication overhead, our Multi-head FL incorporates a designated epoch E_f for freezing the backbone network. When the global epoch reaches E_f, the backbone network will no longer be updated. This implies that after completing E_f global epochs, each user will only need to receive the frozen backbone network once more. After receiving the frozen backbone once, users can store it locally, eliminating the need for additional uploads and downloads, which will significantly reducing communication overhead.

We use ω to represent the backbone network, ${}_n\boldsymbol{\theta}$ to represent the n-th head network, C_i to represent the set of head network IDs corresponding to the i-th user. Same as FedAVG, at the start of the t-th round of learning, near-RT RIC first selects m member users. Then the near-RT RIC distributes the backbone $\boldsymbol{\omega}_i^t$ and the corresponding head networks $\{{}_n\boldsymbol{\theta}, n \in C_i\}$ to the member users. After receiving the complete model, the users divide their local training dataset into multiple subsets based on the mapping relationships between inputs and head networks, which can be written as:

$$D_i \rightarrow \{D_{i,k}, k = 1, \ldots, c_i\} \tag{7}$$

where $c_i = |C_i|$ is the number of head networks for i-th user. Therefore, the local loss function for each user and the global loss function become:

$$F_i(\boldsymbol{\omega}, \boldsymbol{\theta}) = \frac{1}{d_i} \sum_{n \in C_i} \sum_{k \in D_{i,n}} f(\boldsymbol{\omega}_i, {}_n\boldsymbol{\theta}_i, \mathbf{l}_i^k, \mathbf{r}_i^k) \tag{8}$$

$$J(\boldsymbol{\omega}, \boldsymbol{\theta}) = \frac{1}{d} \sum_{i=1}^{M} d_i F_i(\boldsymbol{\omega}, \boldsymbol{\theta}) \tag{9}$$

and the overall objective becomes:

$$(\boldsymbol{\omega}, \boldsymbol{\theta})^* = \arg\min_{(\boldsymbol{\omega}, \boldsymbol{\theta})} J(\boldsymbol{\omega}, \boldsymbol{\theta}) \tag{10}$$

The member users locally update the model for local epoch E_l using SGD:

$$\begin{aligned}&(\boldsymbol{\omega}_i^{t,j+1}, \boldsymbol{\theta}_i^{t,j+1}) \\ &= (\boldsymbol{\omega}_i^{t,j}, \boldsymbol{\theta}_i^{t,j}) - \eta \nabla F_i(\boldsymbol{\omega}^{t,j}, \boldsymbol{\theta}^{t,j}) \text{ for } j = 1, 2, \ldots, E_l\end{aligned} \tag{11}$$

After local training, the user upload the updated the backbone and heads network to the near-RT RIC. When the near-RT RIC receives all updated model, it aggregate backbone network by applying (6) and aggregate each head network by applying:

$${}_n\boldsymbol{\theta}^{t+1} = \sum_{i=1}^{z_n} \frac{d_{i,n}}{\sum_{j=1}^{z_n} d_{j,n}} {}_n\boldsymbol{\theta}_i^{t+1} \tag{12}$$

where z_n denotes the number of users that the corresponding head networks contain the n-th head network. When the global epoch $E > E_f$, the parameters of the backbone are frozen, which means the backbone still participates in forward propagation and does not participate in backward propagation. Thus equation (11) becomes:

$$\boldsymbol{\theta}_i^{t,j+1} = \boldsymbol{\theta}_i^{t,j} - \eta \nabla F_i(\boldsymbol{\omega}^{E_f}, \boldsymbol{\theta}^{t,j}) \text{for } j = 1, 2, \ldots, E_l \tag{13}$$

The near-RT RIC set a flag for each user, with the initial value set to False. When a user's flag is True, it means that the user already has the $\boldsymbol{\omega}^{E_f}$ locally, and the near-RT RIC does not need to send the backbone to that user. When the near-RT RIC sends the $\boldsymbol{\omega}^{E_f}$ to a user, it sets that user's flag to True. The overall process of Multi-head FL deployed in the proposed framework can be referenced in Algorithm 2.

V. EXPERIMENT AND DISCUSSIONS

In this section, we introduce the experimental setup in detail, the experimental results and the corresponding discussions.

Algorithm 2 Pseudo-code for the proposed Multi-head FL deployed in the proposed framework.

Input: Number of users, M; The training datasets, $D = \{D_i, i = 1, 2, \ldots, M\}$; Global Epoch, E; Local Epoch, E_l; Freezing Backbone Epoch, E_f; Number of head networks, N; Learning rate, η;

Output: The global model $(\boldsymbol{\omega}^E, \boldsymbol{\theta}^E)$.

Initialize the mapping relationships between inputs and head networks and $(\boldsymbol{\omega}^0, \boldsymbol{\theta}^0)$
for $t = 0, 1, \ldots, E_f$-1 **do**
  near-RT RIC determine m and S^t
  near-RT RIC send backbone and the corresponding heads to member users
  for each user in S^t **do**
    $(\boldsymbol{\omega}_i^{t,0}, \boldsymbol{\theta}_i^{t,0}) \leftarrow (\boldsymbol{\omega}^t, \boldsymbol{\theta}^t)$
    $D_i \rightarrow \{D_{i,k}, k = 1, \ldots, c_i\}$
    for $j = 0, 1, \ldots, E_l$-1 **do**
      $(\boldsymbol{\omega}_i^{t,j+1}, \boldsymbol{\theta}_i^{t,j+1}) \leftarrow (\boldsymbol{\omega}_i^{t,j}, \boldsymbol{\theta}_i^{t,j}) - \eta \nabla F_i(\boldsymbol{\omega}^{t,j}, \boldsymbol{\theta}^{t,j})$
    $(\boldsymbol{\omega}_i^{t+1}, \boldsymbol{\theta}_i^{t+1}) \leftarrow (\boldsymbol{\omega}_i^{t,E_l}, \boldsymbol{\theta}_i^{t,E_l})$
  near-RT RIC **do**
    $\boldsymbol{\omega}^{t+1} \leftarrow \sum_{i \in S^t} \frac{d_i}{d} \boldsymbol{\omega}_i^{t+1}$
    for $n = 1, 2, \ldots, N$ **do**
      ${}_n\boldsymbol{\theta}^{t+1} \leftarrow \sum_{i=1}^{z_n} \frac{d_{i,n}}{\sum_{j=1}^{z_n} d_{j,n}} {}_n\boldsymbol{\theta}_i^{t+1}$
for $t = E_f, \ldots, E$-1 **do**
  near-RT RIC **do** step 3
  near-RT RIC send the head network to member users
  for each user in S^t **do**
    $(\boldsymbol{\omega}_i, \boldsymbol{\theta}_i^{t,0}) \leftarrow (\boldsymbol{\omega}^{E_f}, \boldsymbol{\theta}^t)$
    $D_i \rightarrow \{D_{i,k}, k = 1, \ldots, c_i\}$
    for $j = 0, 1, \ldots, E_l$-1 **do**
      $\boldsymbol{\theta}_i^{t,j+1} \leftarrow \boldsymbol{\theta}_i^{t,j} - \eta \nabla F_i(\boldsymbol{\omega}^{E_f}, \boldsymbol{\theta}^{t,j})$ for $j = 1, 2, \ldots, E_l$
    $\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}^{t,E_l}$
  near-RT RIC **do** step 13-14
near-RT RIC output the final global model $(\boldsymbol{\omega}^E, \boldsymbol{\theta}^E)$

A. EXPERIMENTAL SETUP

In this paper, we investigated the impact of the ratio of head model weights to the overall model weights on the experimental results, which we refer to as Head weights Ratio (HWR). In this paper, multiple head networks in the same model have the same structure. We use N_h (weights + bias) to represent the number of parameters in one head network ${}_n\boldsymbol{\theta}$, and N_m (weights + bias) to represent number of parameters in one complete model $(\boldsymbol{\omega}, {}_n\boldsymbol{\theta})$ (one head network plus the backbone network). So the HWR in this paper can be written as:

$$HWR = \frac{N_h}{N_m} \tag{14}$$

We explore the model performance and communication overhead under two different values of HWR = 49.9% (referred to as "high HWR") and HWR = 5.96% (referred to as "low HWR"). To clearly demonstrate the algorithm's performance, we performed simple feature engineering by adding the distances between the measurement point and the four base stations, thus the input features consist of 6 dimensions. For the case HWR = 49.9%, the overall model (one head network plus the backbone network) comprises four fully connected (FC) layers ($6 \times 256 \times 1024 \times 256 \times 4$), with the last two layers being the head models. In this scenario, N_h = 1024*256 + 256 + 256*4 + 4 = 263428 and N_m = 6*256 + 256 + 256*1024 + 1024 + 1024*256 + 256 + 256*4 + 4 = 528388. Meanwhile, the experimental model used for FedAVG consists of only four fully connected layers, with the same size as the backbone plus one head model, as shown in Figure 5. For the case HWR = 5.96%, the overall model (one head network plus the backbone network) comprises five FC layers ($6 \times 256 \times 1024 \times 256 \times 128 \times 4$), with the last two layers being the head model (see Figure 6). In this scenario, N_h = 256*128 + 128 + 128*4 + 4 = 33412 and N_m = 6*256 + 256 + 256*1024 + 1024 + 1024*256 + 256 + 256*128 + 128 + 256*4 + 4 = 560772.

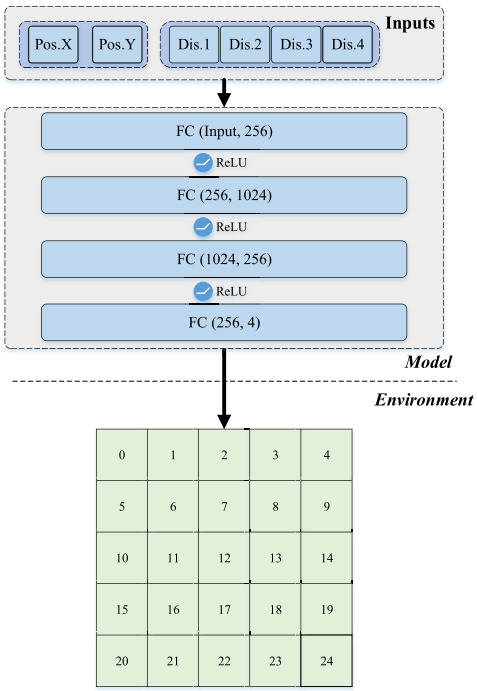
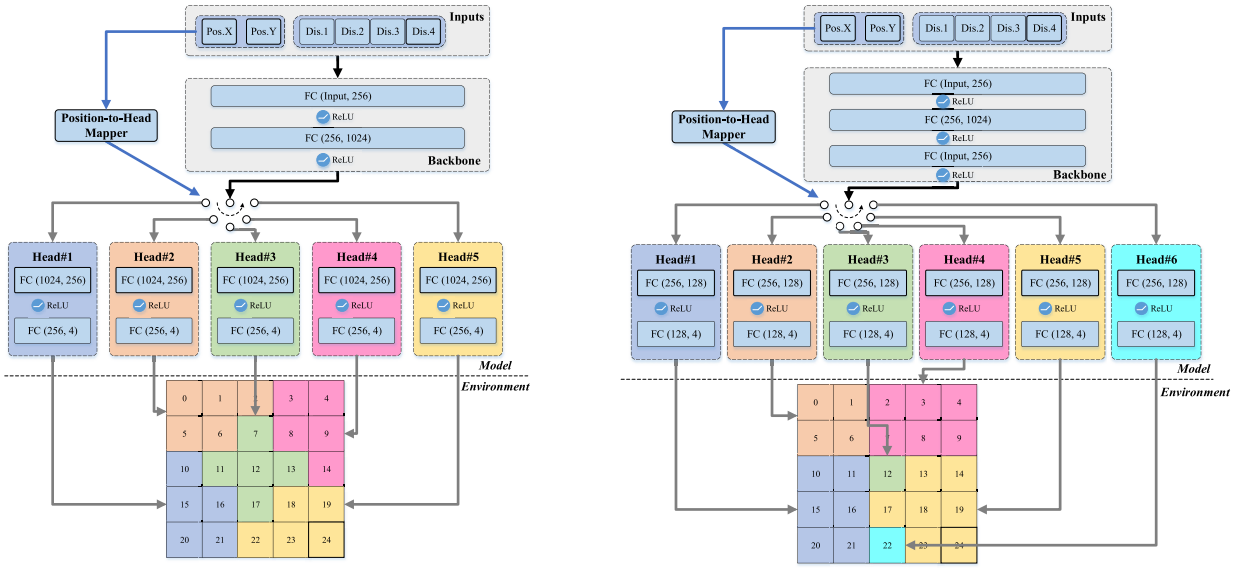


FIGURE 5. The experimental model structure of FedAVG.

Besides, we investigated the model performance and communication overhead under three different values of N, namely $N = 25$, $N = 5$, $N = 6$. For the case $N = 25$, each of the 25 sub-area in Figure 2(b) corresponds to one head network, which means there are 25 head networks in this case. For example, the moving area of user 1 is 0, 1, 2, 5, 6, 7, 10, 11, 12, so the near-RT RIC needs to send 9 head networks to user 1 in step 4 of algorithm 2. For the case of $N = 5$, we evenly divided the entire area into 5 small areas, with each small area corresponding to one head network,



(a) N=5 (b) N=6

FIGURE 6. The experimental model structure of Multi-head FL and the corresponding sub-areas of the head for (a) $N = 5$. (b) $N = 6$.

see Figure 6(a). In the case $N = 6$, we selected two of the most complex sub-area in the environment (sub-area ID is 12 and 22), each corresponding to one head network, while the remaining sub-areas are evenly corresponded, as shown in Figure 6(b).

After assigning the head models to their corresponding sub-areas, each user was assigned the corresponding head model. A user's head model includes the head models corresponding to the sub-areas within their moving sub-areas. For example, for User 89, based on Table 1, the moving sub-areas are [12], [13], [14], [17], [18], [19], [22], [23], [24]. In the scenario shown in Figure 6(a), User 89 corresponds to head#3, 4, 5, and in scenario shown in Figure 6(b), User 89 corresponds to head#3, 5, 6. This means that when selecting User 89 to be a member in the scenario shown in Figure 6(a), head#3, 4 and 5 need to be sent.

We randomly divide each user's dataset into training and testing datasets in a 99:1 ratio. The test loss of FedAVG and Multi-head FL are calculated on the testing datasets using formulas (4) and (9) respectively. In our training datasets and testing datasets, the input features is the location of the user, and the labels is a four-dimensional vector indicating the RSRP value between the user and 4 base stations at the location (shown in Fig 3), and the global model is applicable to all users in the entire region. The locations of users in the input vector are randomly distributed within their moving areas, and no sampling points at the same location. In the context of our datasets and federated learning, the primary evaluation criterion for model performance is the loss on the whole test dataset, while for communication costs, we assess it based on the total amount of parameters transmitted in federated learning.

In all following experiments, we set the global epoch $E = 5000$. For each epoch, we randomly select 5 members, and we calculated the global test loss and communication overhead for different values of N and E_f. All the parameters of interest and their values for the experiment are listed in Table 2.

TABLE 2. The experimental parameters of interest and corresponding values.

Parameters	Description	Values
M	Number of users	90
B	Number of base stations	4
m	Number of member users	5
HWR	Head weights ratio	[49.9%, 5.96%]
N	The number of head networks	[25, 6, 5]
E	Global epochs	5000
E_l	Local training epochs	20
E_f	Freezing backbone epoch	[1000, 2000, 3000]
η	Learning rate	3e-4

B. EXPERIMENTAL RESULTS

In this section, we introduce the experimental results of the case of high HWR and the case of low HWR.

1) THE RESULTS OF HIGH HWR SCENARIO

In this section, we present and discuss the experimental results for the case of high HWR where HWR = 49.9%. Figure 7 shows the global test loss versus global epochs for FedAVG and Multi-head FL when HWR = 49.9% under different E_f and N. Figure 8 shows the communication overhead

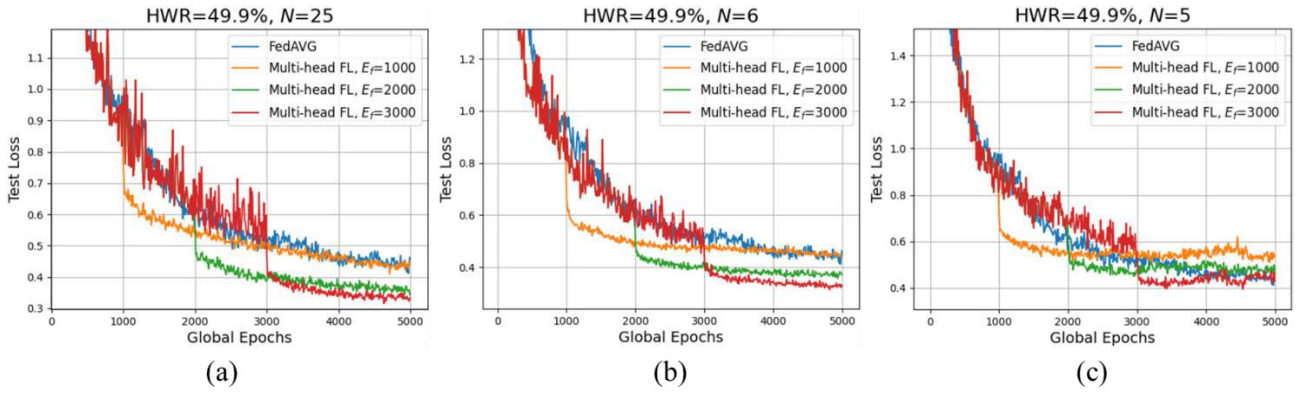


FIGURE 7. The global test loss versus global epochs for FedAVG and Multi-head FL when HWR = 49.9% and freezing backbone epoch E_f = 1000, 2000, 3000 respectively for (a) N = 25. (b) N = 6. (c) N = 5.

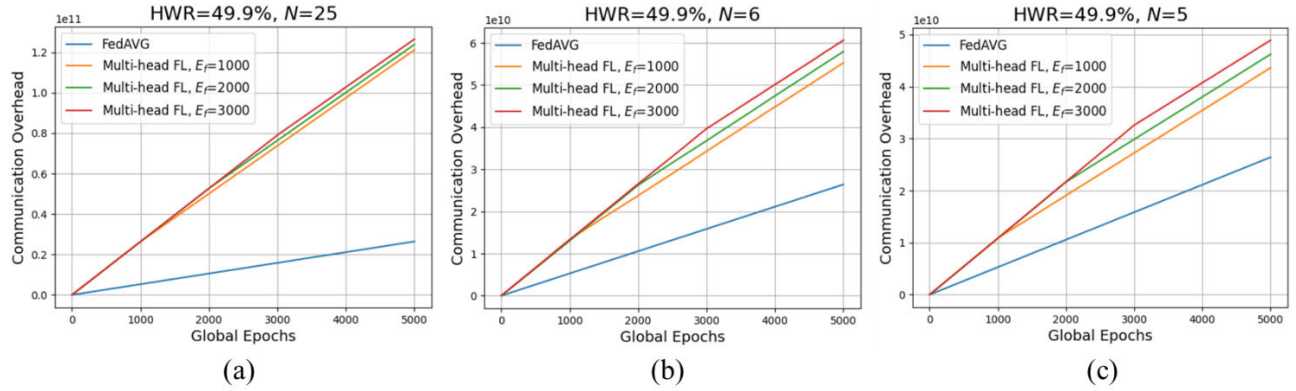


FIGURE 8. The communication overhead versus global epochs for FedAVG and Multi-head FL when HWR = 49.9% and freezing backbone epoch E_f = 1000, 2000, 3000 respectively for (a) N = 25. (b) N = 6. (c) N = 5.

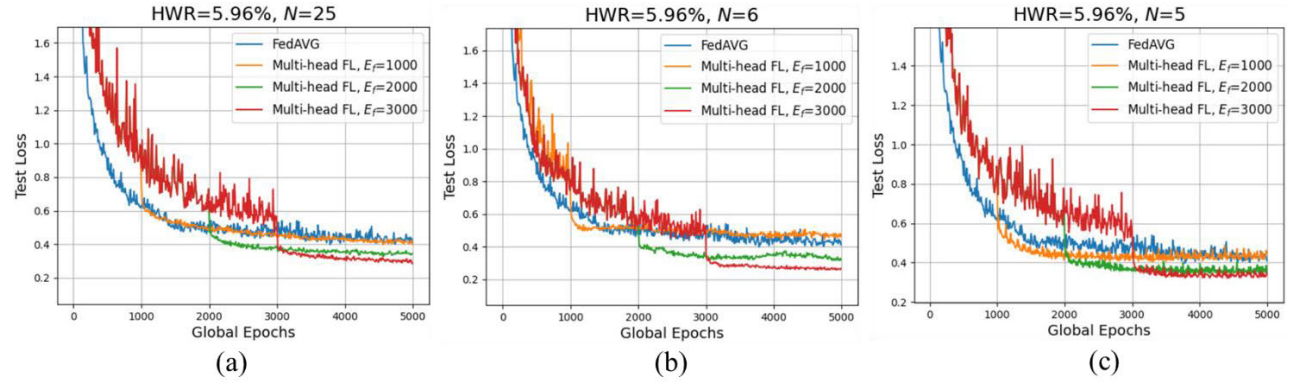


FIGURE 9. The global test loss versus global epochs for FedAVG and Multi-head FL when HWR = 5.96% and freezing backbone epoch E_f = 1000, 2000, 3000 respectively for (a) N = 25. (b) N = 6. (c) N = 5.

versus global epochs for FedAVG and Multi-head FL when HWR = 49.9% under different E_f and N.

From Figure 7, it can be observed that when freezing the backbone, there is a sharp drop in the global test loss. For the cases of $N = 25$ and $N = 6$, Multi-head FL can significantly improve global performance compared to FedAVG, and it can reduce the global test loss by up to approximately 26.7% for the case of $N = 6$ and $E_f = 3000$. This is because in FedAVG, data from different users or areas tends to be non-IID, resulting in a decrease in global model performance. Besides, after rounds of training, the global model may have overfitting on some users or areas. In our Multi-head FL, we use different

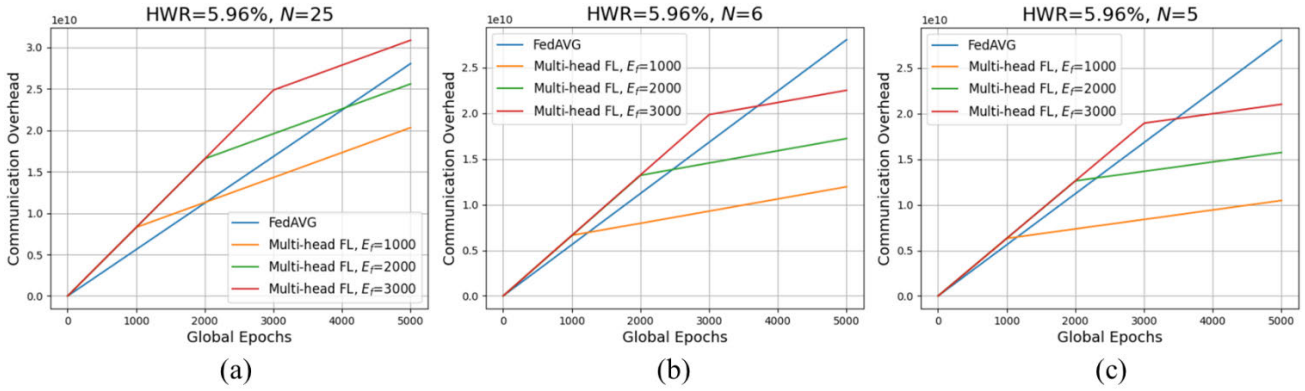


FIGURE 10. The communication overhead versus global epochs for FedAVG and Multi-head FL when HWR = 5.96% and freezing backbone epoch E_f = 1000, 2000, 3000 respectively for (a) $N = 25$. (b) $N = 6$. (c) $N = 5$.

head networks to represent different sub-areas. The smaller the granularity of a sub-area, the closer the dataset with the sub-area tends to be IID, and the head network can better present the characteristics within the sub-area. At the same time, it also to some extent avoids overfitting of the global model in some areas. Therefore, the Multi-head FL proposed in this paper can significantly improve the model performance compared to FedAVG. When $E_f = 1000$, the final global test loss values for two algorithms are quite similar. As E_f increases, the final global test loss decreases. This is because when E_f is relatively small, the backbone may not have converged to the global optimum before being frozen. Furthermore, it is evident that as N increases, the overall convergence speed becomes slower because with a larger N, there are more weights that need to be trained. From Figure 8, it can be observed that for high HWR scenario, Multi-head FL does not reduce communication overhead but rather increases it. Additionally, as N increases, the overall communication overhead increases. This is because in high HWR scenario, even though the back-bone network doesn't need to be transmitted, uploading and downloading multiple head models still require several times more communication overhead compared to FedAVG.

2) THE RESULTS OF LOW HWR SCENARIO

In this section, we present and discuss the experimental results for the case of low HWR where HWR = 5.96%. Figure 9 shows the global test loss versus global epochs for FedAVG and Multi-head FL when HWR = 5.96% under different E_f and N. Figure 10 shows the communication overhead versus global epochs for FedAVG and Multi-head FL when HWR = 5.96% under different E_f and N.

From Figure 9, it can be observed that in the low HWR scenario, when $E_f = 1000$, the global test loss for Multi-head FL is not significantly different from FedAVG. But when $E_f = 2000$ and 3000, Multi-head FL can effectively reduce the global test loss. In the case of $N = 6$ and $E_f = 3000$, Multi-head FL can reduce the global test loss by approximately 38.6%. Before freezing the backbone, the convergence speed of Multi-head FL is slower than FedAVG. However, after freezing the backbone, the global test loss of Multi-head FL sharply decreases below that of FedAVG. From Figure 10, it can be observed that in the low HWR scenario, the Multi-head FL can significantly reduce the overall communication overhead of federated learning. When $N = 5$ and $E_f = 1000$, Multi-head FL can reduce the overall communication overhead by approximately 62.7% compared to FedAVG while, as shown in Figure 9(c), the global test loss did not decrease. For the best-case scenario in terms of model performance, i.e., $N = 6$, $E_f = 3000$, Multi-head FL can still reduce the overall communication overhead by approximately 19.8%.

VI. CONCLUSION AND FUTURE WORKS

This paper presents an innovative Multi-head FL algorithm for 6G network, and deploys it within the proposed federated learning framework with the near-RT RIC as the server. Using the RSRP prediction scenario as the backdrop, we compared the Multi-head FL algorithm with FedAVG and analyzed the model performance and overall communication costs of federated learning under different values of various parameters, including the number of head models and the freezing backbone epoch. We designed two different scenarios, which we refer to as high HWR and low HWR, where the ratio of the weights of head network to the overall model weights differs. The experimental results demonstrate that for the high HWR scenario, Multi-head FL can improve the model performance compared to FedAVG, but increase the overall communication overhead. In contrast, for the low HWR scenario, Multi-head FL can not only significantly improve the model performance compared to FedAVG but also substantially reduce the overall communication overhead. Considering the constraints of bandwidth and computational capability of UE in real-life, the HWR can be adjusted reasonably in the practical deployment to achieve the best compromise between model performance and communication overhead.

In the future work, we will further improve the research work in this paper from the following aspects. Firstly, more diverse datasets and application scenarios will be considered

for the experiments and the performance evaluation. Secondly, our proposed Multi-head FL algorithm will be further studied with more complex models more than the DNN model. The road to 6G is still long. In the future, we will focus our research on the points mentioned above to enhance our algorithm from various perspectives.

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] Y. Xiao, G. Shi, and M. Krunz, "Towards ubiquitous AI in 6G with federated learning," 2020, *arXiv:2004.13563*.

[3] K. N. Nguyen, A. Ali, J. Mo, B. L. Ng, V. Va, and J. C. Zhang, "Beam management with orientation and RSRP using deep learning for beyond 5G systems," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2022, pp. 133–138.

[4] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[5] B. Balasubramanian, E. S. Daniels, M. Hiltunen, R. Jana, K. Joshi, R. Sivaraj, T. X. Tran, and C. Wang, "RIC: A RAN intelligent controller platform for AI-enabled cellular networks," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 7–17, Mar. 2021.

[6] S. Niknam, A. Roy, H. S. Dhillon, S. Singh, R. Banerji, J. H. Reed, N. Saxena, and S. Yoon, "Intelligent O-RAN for beyond 5G and 6G wireless networks," in *Proc. IEEE Globecom Workshops*, Dec. 2022, pp. 215–220.

[7] L. Tao, J. Lu, and X. Wang, "Feature engineering based intelligent wireless propagation model for RSRP prediction," *Proc. IOP Conf. Series, Mater. Sci. Eng.*, vol. 768, no. 4, pp. 1–23, Mar. 2020.

[8] Y. Zheng, J. Wang, X. Li, J. Li, and S. Liu, "Cell-level RSRP estimation with the image-to-image wireless propagation model based on measured data," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 6, pp. 1412–1423, Dec. 2023, doi: 10.1109/tccn.2023.3307945.

[9] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless federated learning (WFL) for 6G Networks4 Part I: Research challenges and future trends," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 3–7, Jan. 2022.

[10] H. Yu, Z. Hou, Y. Gu, P. Cheng, W. Ouyang, Y. Li, and B. Vucetic, "Distributed signal strength prediction using satellite map empowered by deep vision transformer," in *Proc. IEEE Globecom Workshops*, Dec. 2021, pp. 1–6.

[11] J. Stock, O. Hauke, and J. Weismann, "The applicability of federated learning to official statistics," 2023, *arXiv:2307.15503*.

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

[13] A. M. Kumar and D. P. Albena, "Dynamic pathloss model for place and time itinerant networks," *Wireless Pers. communications.*, vol. 100, no. 2, pp. 641–652, 2018.

[14] L. Pedraza, C. Hernandez, and O. Salcedo, "Spectrum forecast using propagation losses of the okumura-hata model," *Int. J. Commun. Antenna Propag. (IRECAP)*, vol. 6, no. 5, p. 328, Oct. 2016.

[15] H. F. Li, W. X. He, and X. G. He, "Review of wireless personal communications radio propagation models in high altitude mountainous areas at 2.6GHz," *Wireless Pers. Commun.*, vol. 101, no. 2, pp. 53–735, 2018.

[16] S. Wu, B. Ma, J. Zhang, S. Zheng, W. Shao, and W. Zheng, "Intelligent propagation model method for RSRP prediction based on machine learning," in *Proc. 7th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2021, pp. 2287–2291.

[17] U. Masood, H. Farooq, and A. Imran, "A machine learning based 3D propagation model for intelligent future cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[18] L. Dai, H. Zhang, and Y. Zhuang, "Propagation-model-free coverage evaluation via machine learning for future 5G networks," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–5.

[19] O. Haliloglu, E. U. Soykan, and A. Alabbasi, "Privacy preserving federated RSRP estimation for future mobile networks," in *Proc. IEEE Globecom Workshops*, Dec. 2021, pp. 1–6.

[20] E. Bakopoulou, M. Yang, J. Zhang, K. Psounis, and A. Markopoulou, "Location leakage in federated signal maps," 2021, *arXiv:2112.03452*.

[21] P. Pu Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," 2020, *arXiv:2001.01523*.

[22] N. Bhuyan and S. Moharir, "Multi-model federated learning," in *Proc. 14th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2022, pp. 779–783.

[23] Z. Yang, D. Gjorgjevikj, J. Long, Y. Zi, S. Zhang, and C. Li, "Sparse autoencoder-based multi-head deep neural networks for machinery fault diagnostics with detection of novelties," *Chin. J. Mech. Eng.*, vol. 34, no. 1, pp. 1–12, Dec. 2021.

[24] K. Fang, Q. Tao, Y. Wu, T. Li, X. Huang, and J. Yang, "On multi-head ensemble of smoothed classifiers for certified robustness," 2022, *arXiv:2211.10882*.

[25] Z. Yang, D. Gjorgjevikj, J. Long, Y. Zi, S. Zhang, and C. Li, "Sparse auto encoder-based multi-head deep neural networks for machinery fault diagnostics with novelty detection," *Chin. J. Mech. Eng.*, vol. 3, pp. 146–157, Jan. 2021.

MENGHAN YU received the M.S. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 2021. He is currently an Associate Engineer with the 6G Research Center, China Telecom Research Institute. His research interests include 6G network architecture, 6G native AI, and distributed learning.

XIONG XIONG received the B.S. and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), China, in 2013 and 2022, respectively. He is currently an Associate Engineer with the 6G Research Center, China Telecom Research Institute. His research interests include M2M networks, NTN, software defined radio, and reinforcement learning.

ZHEN LI received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2018. He currently works as a Research Engineer with China Telecom Research Institute. His research interests include intent-driven networks and 6G network architecture.

XU XIA (Member, IEEE) received the master's degree in information and communication from Chonbuk National University, South Korea, in 2005. He has been with China Telecom Research Institute, since 2012. Currently, he is the Standards Director of 6G, B5G, and industry convergence. His research interest includes core network innovation. He has made extensive contributions to GSMA, CCSA, and 5GAA standards and the Vice Chair of 3GPP SA1. He has also served on keynotes, program and organization committees of numerous leading wireless and networking conferences.