IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 March 2024, accepted 31 March 2024, date of publication 3 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3384566

RESEARCH ARTICLE

Layer-Wise Personalized Federated Learning for Mobile Traffic Prediction

SEUNGYEOL LEE[1], **JIHOON SUNG**[2], **AND MYUNG-KI SHIN**[2]
[1]Department of Information and Communication Engineering, University of Science and Technology, Daejeon 34113, Republic of Korea
[2]Standard and Open Source Research Division, Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea

Corresponding author: Myung-Ki Shin (mkshin@etri.re.kr)

This work was supported by the Institute of Information Communications Technology Planning Evaluation (IITP) Grant funded by Korean Government (MSIT), Development of Distributed/Cooperative AI-Based 5G+ Network Data Analytics Functions and Control Technology, under Grant 2021-0-00739.

ABSTRACT With the evolution of mobile networks delivering high-performance network services to a myriad of devices, accurate mobile traffic prediction has become increasingly important. In recent years, federated learning (FL) has emerged as a communication-efficient approach, enabling collaborative model training without the centralized data aggregation. Despite its promising potential, FL-based mobile traffic prediction has following two major challenges: 1) data heterogeneity across regions: The diverse communication and mobility patterns inherent to different regions can lead to uneven traffic distribution. Training on such heterogeneous data can result in the global model failing to capture the unique patterns of specific regions, compromising consistent prediction performance across all regions; 2) communication efficiency concerns: The frequent exchange of large model weights during training leads to substantial signaling overhead in the FL. This added communication can pose a significant burden on the limited network bandwidth, potentially causing performance degradation in mobile networks. In this paper, we propose a novel personalized FL framework to address these challenges. Our framework enables a fine-grained federation through a layer-wise aggregation for the global model. This approach personalizes the global model to capture unique regional characteristics such as traffic spikes and other irregular patterns. In addition, we introduce an adaptive layer freezing mechanism to reduce communication costs during training. By selectively transmitting only the layers that require further training, our framework effectively enhances communication efficiency without sacrificing prediction performance. Extensive experiments on a real-world mobile traffic dataset demonstrate that our approach not only provides superior prediction accuracy compared to baselines but also achieves significant communication cost saving.

INDEX TERMS Mobile traffic prediction, traffic management, personalized federated learning.

I. INTRODUCTION

Mobile traffic prediction is a critical task in mobile networks, particularly in the context of the upcoming 6G networks. With the proliferation of high-performance devices and the need to provide high-quality services, the demand for sophisticated network management and control is expected to increase significantly [1]. 6G networks are expected to leverage AI-based solutions to support advanced use cases such as autonomous driving, and augmented/virtual reality [2]. These use cases require ultra-reliable and low-latency connectivity, which can be achieved through accurate traffic prediction.

In the field of mobile traffic prediction, deep learning (DL) algorithms [3] have shown superior performance compared to conventional statistical-based models such as ARIMA and HA. DL models, especially those based on recurrent neural networks (RNNs) [4] and convolutional neural networks (CNNs) [5], have been widely used for mobile traffic prediction tasks owing to their ability to capture complex temporal and spatial patterns.

A notable recent research trend in the domain of mobile traffic prediction is the shift from data-centralized DL

The associate editor coordinating the review of this manuscript and approving it for publication was Maurice J. Khabbaz.

models to data-distributed federated learning (FL) approaches [6], [7]. In contrast to centralized learning which sends vast amounts of raw data to a central server, FL ensures that only model weights obtained by local training are sent to the central server. This approach reduces the frequent transmission of training data and signaling overhead, thereby preventing potential adverse effects on payload transmissions. Furthermore, due to the advancements in edge computing technology to support high QoS requirements of specific use cases in mobile networks, powerful data centers are deployed across multiple regions. This distributed paradigm has paved the way for FL [8], where data can be trained locally in each region, and effectively learned models can be shared. Given these multifaceted advantages, FL is expected to play a pivotal role in next-generation communications [9].

Despite its promising potential, mobile traffic prediction under the FL framework has two major research challenges.

1) **Data Heterogeneity across Regions:** Traffic distribution varies across different regions due to the diverse communication and mobility patterns exhibited by users in each region. Training on such diverse data under the FL framework is very challenging because generalizing these unique regional patterns into a global model can lead to degradation of performance in specific regions, making it difficult to achieve consistently high predictive performance across all regions.
2) **Communication Efficiency Concerns:** FL framework is generally known to offer communication benefits over centralized learning. However, when dealing with small training data and a large number of model parameters, the expected communication cost savings are reduced. Given that the relatively low-volume training data such as communication logs, it is essential to design FL framework that improves communication efficiency.

In recent years, there has been a great effort to tackle these challenges by researchers. For instance, a knowledge distillation-based approach [10] has been leveraged to personalize local models. The global model is then refined by fine-tuning itself using the local models from all regions. Meta-learning has been explored to address the challenges of heterogeneous mobile data, requiring only a few steps of fine-tuning on local dataset [11]. Dual-attention-based FL (FedDA) [12] is another notable attempt in this domain, which addresses data heterogeneity through data augmentation strategies. While these approaches have yielded promising prediction performance, certain limitations persist. Most notably, the reliance on a single globally-shared model often fails to effectively capture the unique and rare traffic patterns to specific regions. Furthermore, a significant shortcoming in these methods is the lack of consideration for communication efficiency, which is crucial in mobile networks.

In this paper, we address the aforementioned challenges and limitations by introducing a layer-wise personalized FL framework. This novel approach enables fine-grained FL, whereby each layer of the global model is trained separately for personalization. To effectively separate the layers based on distinct traffic characteristics, we conducted a preliminary analysis that precisely explores the differences in traffic across regions. We observe that the decomposition scheme, which separates traffic data into trend and seasonality components, provides a detailed comparison of disparate traffic patterns across regions.

Driven by this observation, our framework adopts a decomposition scheme that semantically separates the model layers into trend and seasonality layers. To aggregate these separated layers in a layer-wise manner, we introduce a decomposition-based clustering strategy. This strategy groups clients based on their decomposed traffic characteristics, aligning clients with similar traffic patterns for more efficient layer-wise aggregation. Subsequently, the server combines aggregated layers to obtain a personalized global model for each specific region. This approach captures the unique and rare temporal characteristics of individual regions by multiple global models, each derived from a distinct combination of learned layers. Consequently, our framework ensures high prediction performance across all regions.

Furthermore, to enhance the communication efficiency of our framework, we incorporate an adaptive layer freezing mechanism. This technique allows selective updating of model layers based on the difference of weight changes after each communication round. Those layers with relatively stable weights are frozen, curtailing the transmission costs and enhancing communication efficiency. The main contributions of this paper are as follows:

- We propose a novel personalized FL framework that separately trains each layer of the global model. This fine-grained FL enables a granular understanding of mobile traffic patterns, achieving a personalized global model tailored to capture unique regional characteristics.
- We introduce a decomposition-based clustering strategy within our FL framework, designed for layer-wise aggregation. This strategy groups clients based on refined temporal characteristics obtained through decomposition.
- We introduce an adaptive layer freezing mechanism that minimizes the communication cost for training. By selectively transmitting model layers based on their convergence, the framework reduces communication overhead without sacrificing prediction accuracy.
- Our extensive experiments with real-world mobile traffic datasets demonstrate the superior performance of our framework over existing methods, highlighting its effectiveness in both improving prediction accuracy and reducing communication costs.

Our findings demonstrate that the decomposition-based layer-wise approach effectively achieves model personalization. This enhanced personalization capability successfully captures region-specific traffic patterns, leading to improved prediction accuracy compared to existing works. Furthermore, our framework addresses the issue of communication efficiency during training, a critical challenge in FL for mobile traffic prediction not thoroughly explored by existing works.

The rest of this paper is organized as follows. The related literature in mobile traffic prediction is introduced in Section II. Next, the preliminary analysis and problem description are presented in Section III. The proposed framework and learning procedure are described in Section IV. The prediction performance comparison with real-world datasets is presented in Section V. The conclusions and plans for future work are presented in Section VII.

II. RELATED WORKS

In this section, we provide a review of existing research in the field of mobile traffic prediction, focusing on recent trends that incorporate FL approaches.

A. MOBILE TRAFFIC PREDICTION

In recent years, the field of mobile traffic prediction has gained significant attention due to the increasing demand for accurate traffic modeling and prediction in various telecommunication tasks. This problem is essentially a time series prediction task, and existing methods can be broadly categorized into two groups: statistical-based models and DL-based models.

Statistical-based models predict mobile traffic using statistical and probabilistic tools. There are well-known prediction methods such as the Historical Average (HA) [13] and Autoregressive Integrated Moving Average (ARIMA) [14]. HA predicts future values based on the average or the last observation of historical data. While these methods are straightforward to implement, they often fail to capture the underlying patterns in mobile traffic, resulting in relatively poor prediction performance. ARIMA has been explored to characterize the self-similarity and bursty nature of mobile traffic. Variants of ARIMA [15] have also been proposed to handle different aspects of mobile traffic prediction, such as decomposing traffic into regularity and randomness components. Another variant, SARIMA, highlighted for its ability to capture seasonal patterns, was effectively used in [16] for improved mobile traffic prediction.

In the meantime, with the advancements in machine learning and AI techniques, DL-based models [17] have emerged as strong competitors to statistical-based models in mobile traffic prediction. Particularly, DL-based models have demonstrated their potential to capture complex patterns in temporal and spatial domains. Various DL frameworks, such as RNN and LSTM, have been applied to capture the spatial and temporal dependencies among different cells in mobile networks. In [4], spatiotemporal correlations was proposed to explore similarities and differences between cells using RNNs within a multi-task learning framework [18]. Additionally, [5] proposed a densely connected CNN to model the nonlinear dynamics of mobile traffic, complemented by a novel fusion scheme to learn the influence of spatial and temporal dependencies. Beyond using mobile traffic data, [19] proposed a spatial-temporal cross-domain neural network (STC-N) model, leveraging data characteristics from similar services and regions.

While the aforementioned works mainly focus on centralized approaches for mobile traffic prediction, our proposed framework in this paper takes a different approach. We aim to solve the mobile traffic prediction problem using a distributed learning architecture with FL, leveraging the benefits of communication efficiency.

B. FEDERATED LEARNING

The recent trend in mobile traffic prediction research notably shifts from data-centralized DL models toward data-distributed FL approaches [6], [7]. The adoption of FL maintains data privacy across multiple devices or nodes by eliminating the need to share raw data. Moreover, FL also minimizes the communication overhead by performing computations on the edge devices themselves, instead of transmitting massive amounts of raw data to a central server. This not only leads to more efficient use of network resources but also enables rapid analytics and decision-making in distributed mobile network environments.

Therefore, many recent studies over the past few years have leveraged the potential of FL within the domain of mobile traffic prediction. For example, widely known as FedAvg [7], this foundational FL approach aggregates parameters from client models across various regions and averages them to obtain a global model. Similarly, FedAtt [20] enhances this approach by introducing an attention mechanism to the aggregation process, which assigns weights to client contributions based on their relevance. These studies reveal that when the client data is independent and identically distributed (IID), models achieve good performance, but there is a significant performance degradation with non-IID data. This degradation also manifests in the context of mobile traffic, which inherently exhibits non-IID characteristics due to its uneven distribution across different regions. This challenge has led to the adoption of personalized FL, a strategy that tailors models to better fit the specific data distributions of each client, thereby enhancing model performance and accuracy in mobile traffic prediction.

One such strategy is the Dual-attention-based FL (FedDA) [12], which demonstrates enhanced prediction performance by leveraging data augmentation to address data heterogeneity. Similarly, the Graph Attention Spatial-Temporal Network (GASTN) [10] employs knowledge distillation to mitigate the heterogeneity issue in mobile traffic prediction. Another notable approach is a federated meta-learning model [11], which adapts to heterogeneous data through minimal local dataset fine-tuning. While these proposed methods utilize various personalized techniques to handle the heterogeneity of mobile traffic in the FL framework, they still have areas that can be enhanced. A single globally-shared model frequently falls short in capturing rare traffic patterns inherent to certain regions. Moreover, these methods often do not extensively consider the aspect of communication efficiency. In contexts where network resources are limited and numerous clients (i.e., base stations, mobile devices) participate, minimizing communication cost becomes crucial.

In contrast to the methods mentioned above, our approach stands out with several unique features that set it apart. Firstly, while traditional methods typically train a single global model for all clients, our approach trains multiple global models tailored to individual clients. This means each client maintains its own personalized model, a crucial differentiation for addressing heterogeneity in the traffic data across regions. Secondly, by employing layer-wise aggregation with decomposed trend and seasonality layers, our method enables a fine-grained FL among clients that exhibit similar temporal patterns. Lastly, we design our method to emphasize on communication efficiency, minimizing communication costs during the learning process.

III. PRELIMINARY ANALYSIS AND PROBLEM DESCRIPTION

In this section, we first introduce the mobile traffic dataset sourced from Telecom Italia,[1] covering the regions of Milan and Trento, and detail its structure and key attributes. Next, we present a decomposition approach to the dataset, separating the raw traffic data into trend and seasonality components, and conduct a spatial correlation analysis that unveils intriguing disparities in correlation when considering the original data versus its decomposed constituents. Lastly, we define the primary problem we aim to address in this paper.

A. MOBILE TRAFFIC DATA

The mobile traffic dataset used in this paper, which is publicly available as part of the Big Data Challenge [21], is sourced from Telecom Italia. This dataset, covering both Milan and Trento in Italia city, was collected from data spanning from 11/01/2013 to 01/01/2014, recorded in 10-minute intervals. The areas of Milan and Trento are composed of grid overlays of 10,000 and 6,575 squares, respectively. Each square, approximately 235 × 235 meters, is referred to as a cell. Within every cell, three different types of mobile traffic are recorded: short message service (SMS), Call service, and Internet service. For any given service type s chosen from { SMS, Call, Internet }, we can represent the city's mobile traffic as a spatiotemporal sequence labeled as $\mathbf{X_s} = \{\mathbf{X_{s,t}} | t = 0, 1, 2, \ldots, T\}$, where T denotes the total count of time intervals. $\mathbf{X_{s,t}}$ denotes the traffic matrix for the t-th interval across an area segmented into $M \times N$ cells and it can be described as

$$\mathbf{X_{s,t}} = \begin{bmatrix} x_{s,t}^{(1,1)} & x_{s,t}^{(1,2)} & \cdots & x_{s,t}^{(1,n)} \\ x_{s,t}^{(2,1)} & x_{s,t}^{(2,2)} & \cdots & x_{s,t}^{(2,n)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{s,t}^{(m,1)} & x_{s,t}^{(m,2)} & \cdots & x_{s,t}^{(m,n)} \end{bmatrix}, \tag{1}$$

where $x_{s,t}^{(m,n)}$ denotes the mobile traffic of service type s in (m, n) cell at time t.

Fig.1a, Fig.1b, and Fig.1c illustrate the temporal dynamics of mobile traffic, where the x-axis denotes the time index (scaled to hours), and the y-axis represents the volume of a specific traffic type. SMS and Call traffic exhibit strong daily and weekly periodic characteristics. In contrast, Internet traffic exhibits relatively complex temporal dynamics, characterized by a mixture of weak periodicity and irregular peaks and troughs. Fig.1d displays the spatial distribution of Internet traffic in Milan within a specific time interval. To effectively visualize the entirety of the city, cells from across the city are uniformly sampled to represent the region and are mapped onto a 20 × 20 grid (i.e., x and y axes), with the z-axis indicating the volume of Internet traffic. This visualization reveals that traffic is distributed unevenly across the city, highlighting the need to consider the heterogeneity of traffic across different regions.

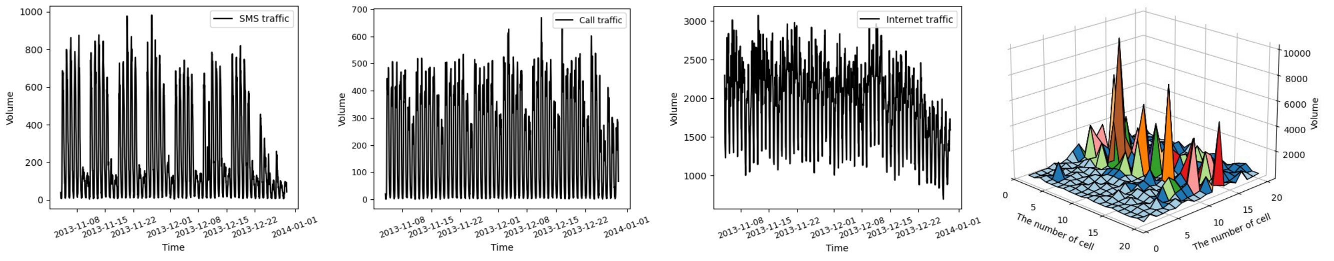
B. KEY OBSERVATION

To explore the dataset in detail, we decompose the raw traffic data into trend and seasonality components. As illustrated in Fig.2a, the decomposition scheme offers a clear understanding of the temporal characteristics in traffic dynamics. The decomposition process also yields a residual component, representing the portion of the original traffic data that remains after the extraction of trend and seasonality values. Although the residual is not utilized in our study, it is depicted in Fig.2a to ensure completeness in demonstrating the decomposition concept. In Fig.2c, the trend values represent the continuous changes in the data over time, capturing consistent patterns essential for understanding the overarching trajectory of the traffic. On the other hand, in Fig.2d, the seasonality values represent the repetitive and periodic fluctuations in the data, often observed within fixed intervals. Decomposing the traffic into trend and seasonality provides a more granular insight into its unique patterns across different regions. We define the trend τ and seasonality σ of each region as follows:

$$\tau_{s,t}^{(m,n)} = \frac{1}{w} \sum_{e=-\frac{w-1}{2}}^{\frac{w-1}{2}} x_{s,w+e}^{(m,n)}, \tag{2}$$

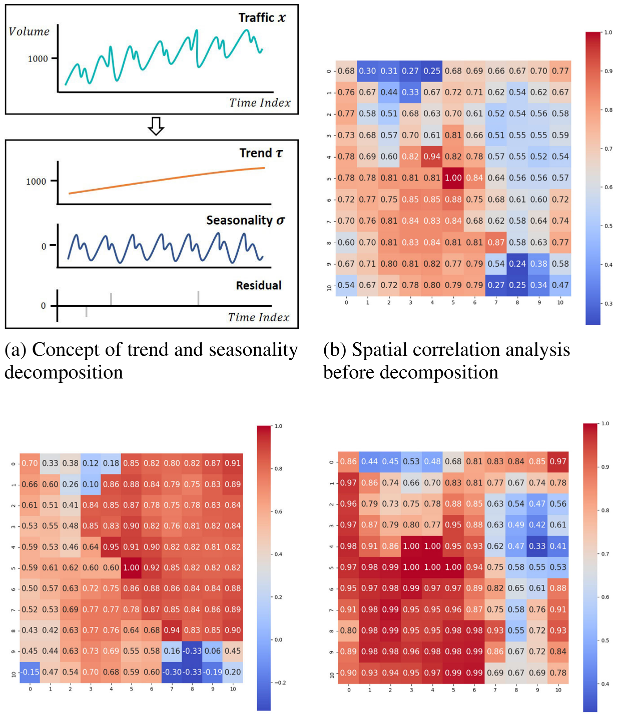
$$\sigma_{s,t}^{(m,n)} = x_{s,t}^{(m,n)} - \tau_{s,t}^{(m,n)}, \tag{3}$$

[1] This dataset is recognized for its accurate representation of regionally distributed mobile traffic. As one of the most comprehensive publicly available datasets, it continues to be widely utilized in various research domains.



(a) SMS traffic dataset (b) Call traffic dataset (c) Internet traffic dataset (d) Spatial distribution of Internet traffic

FIGURE 1. **The temporal and spatial dynamics of mobile traffic data.**



(a) Concept of trend and seasonality decomposition

(b) Spatial correlation analysis before decomposition

(c) Spatial correlation analysis of trend

(d) Spatial correlation analysis of seasonality

FIGURE 2. **Spatial correlation analysis using SMS traffic data: raw data to decomposed components.**

where t is the time index, w denotes and odd window size and e denotes the offset from the central data point.

Before decomposition, we analyze the spatial correlation of the original traffic, selecting SMS traffic as a case study for our analysis, using the Pearson correlation coefficient [17], denoted as ρ.[2] The ρ value between a target cell (m, n) and its neighboring cell (m', n') is calculated as:

$$\rho_{(m,n),(m',n')} = \frac{Cov\left(x_{s,t}^{(m,n)}, x_{s,t}^{(m',n')}\right)}{\delta_{x_s^{(m,n)}} \cdot \delta_{x_s^{(m',n')}}}, \tag{4}$$

where Cov represents the covariance operator which measures the linear relationship of the traffic data of the two cells, and δ is the standard deviation, indicating the extent to which the traffic data in cells (m, n) and (m', n') deviate from their respective mean value. A target cell within an 11×11 cell grid is selected as an example to show the spatial correlation. The obtained ρ values, representing correlations between the target cell $(6, 6)$ and its neighboring cells, are demonstrated in Fig.2b. After decomposing the original traffic data into its trend and seasonality components, the spatial correlation analyses for the same target cell are demonstrated in Fig.2c and Fig.2d, respectively.

An interesting observation derived from the spatial analysis is that the correlation of the original traffic data is not sufficient to fully capture the trends and seasonality. This suggests that even if the correlation coefficient ρ of the original traffic data is relatively high or low, the ρ for the trend or seasonality may not be. For example, the top-right portion of Fig.2b shows a moderate correlation; however, its corresponding trend correlation in Fig.2c is high, while its seasonality correlation in Fig.2d is low. Conversely, the bottom-left portion of Fig.2d shows a high correlation with the target cell, but its trend correlation in Fig.2c is relatively low and its seasonality correlation in Fig.2d is high. Based on the above observations, it becomes evident that the decomposed scheme enables a detailed comparison of disparate traffic patterns across regions. Inspired by this realization, we considered integrating this decomposition approach into the FL framework. In the context of FL, understanding the subtle similarities and differences across regions is critical for model performance. The details of the proposed approach are addressed in subsequent section.

C. PROBLEM DESCRIPTION

In this paper, our goal is to collaboratively train personalized models across multiple regions. We consider a mobile network with I clients (i.e., base stations) geographically distributed with each client i covering one of these regions. As described in Section III-A, each client i covering region (m, n) has local data $x_{s,t}^{i}$ corresponding to the traffic matrix. For this data, let $y_{s,t}^{i}$ denote the ground truth value. Here, client i is associated with a specific region represented by coordinate (m, n). Note that the notation for service type s is omitted in the following, implying that both $x_{s,t}^{i}$ and $y_{s,t}^{i}$

[2] The Pearson correlation coefficient is widely used to measure the linear relationship between two sets of data, in this case, SMS traffic volumes.

are simplified to x_t^i and y_t^i respectively. This is to improve readability, as the formulation is the same for each service type.

Let $\mathcal{X}_t^i = \{x_{t-o+1}^i, x_{t-o+2}^i, \ldots, x_t^i\}$ denote the sequential mobile traffic for client i at the time index t, where o is the length of observation. Correspondingly, the ground truth values over the prediction horizon are represented by $\mathcal{Y}_t^i = \{y_{t+1}^i, y_{t+2}^i, \ldots, y_{t+p}^i\}$, where p is the length of prediction. Each client i has a local data set denoted by $D^i = \{\mathcal{X}_k^i, \mathcal{Y}_k^i\}_{k=1}^K$, where $i \in \{1, 2, \ldots, I\}$ and K is the number of data sets obtained using a sliding window scheme. Then, mobile traffic prediction of client i at time t is as follows:

$$\hat{\mathcal{Y}}_t^i = \mathcal{F}(\mathcal{X}_t^i; \theta^i), \tag{5}$$

where $\mathcal{F}(\cdot)$ denotes the prediction model and θ^i denotes its parameters.

To be concrete for our layer-wise personalized federated scenario, let $\mathcal{F}(\cdot; \theta^i)$ denote a personalized model for client i, parameterized by θ^i. The objective function is:

$$\underset{\mathbf{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{I} \mathcal{L}^i(D^i; \theta^i), \tag{6}$$

where $\mathbf{\Theta}$ is the set of clients' model parameters denoted by $\mathbf{\Theta} = \{\theta_i\}_{i=1}^I$. The loss function is formulated as

$$\mathcal{L}^i(D^i; \theta^i) = \frac{1}{K} \sum_{k=1}^{K} (\mathcal{F}(\mathcal{X}_k^i; \theta^i) - \mathcal{Y}_k^i)^2. \tag{7}$$

For layer-wise aggregation, client i model parameters θ^i are split into θ_t^i for the trend layer and θ_s^i for the seasonality layer, denoted by $\theta^i = \{\theta_t^i, \theta_s^i\}$. During training, model parameters $\theta^i = \{\theta_t^i, \theta_s^i\}$ from all participating clients are aggregated at the server via

$$\bar{\theta}_t = \frac{1}{I} \sum_{i=1}^{I} \theta_{t,(r)}^i, \quad \bar{\theta}_s = \frac{1}{I} \sum_{i=1}^{I} \theta_{s,(r)}^i, \tag{8}$$

where $\bar{\theta}_t$ and $\bar{\theta}_s$ represent the aggregated parameters of the trend and seasonality layers, respectively, at the r-th global communication round. Then, client i updates its local model $\mathcal{F}(D^i; \theta^i)$ using the aggregated parameters from the server, represented as $\mathcal{F}(D^i; \bar{\theta}_t, \bar{\theta}_s)$. Therefore, the loss function (7), can be rewritten as follows:

$$\mathcal{L}^i(D^i; \bar{\theta}_t, \bar{\theta}_s) = \frac{1}{K} \sum_{k=1}^{K} (\mathcal{F}(\mathcal{X}_k^i; \bar{\theta}_t, \bar{\theta}_s) - \mathcal{Y}_k^i)^2. \tag{9}$$

Consequently, the detailed objective function is defined as:

$$\underset{\mathbf{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{I} \left(\frac{1}{K} \sum_{k=1}^{K} (\mathcal{F}(\mathcal{X}_k^i; \bar{\theta}_t, \bar{\theta}_s) - \mathcal{Y}_k^i)^2 \right), \tag{10}$$

where $\mathbf{\Theta} = \{\theta^1, \theta^2, \ldots, \theta^I\}$ represents the collection of parameter sets for personalized models of each client i.

IV. PROPOSED FRAMEWORK

In this section, we present a layer-wise personalized FL framework. We first introduce the client clustering strategy designed for layer-wise aggregation approach. Next, we outline the overall workflow of our proposed framework and describe on the specifics of each step. Lastly, we address our adaptive layer freezing to enhance communication efficiency during the training process.

A. DECOMPOSITION BASED CLIENT CLUSTERING

For our personalized FL approach, it is needed to perform clustering based on the decomposition of mobile traffic per region. Firstly, the trend and seasonality values of each region are determined using (2) and (3). Subsequently, each client prepares proxy datasets[3] intended for clustering on the server-side:

1) Trend Proxy Dataset: For each client i, the trend proxy dataset, denoted as $\mathbf{P_T}^i$, comprises the directly computed trend values over all time steps. These trend values $\tau_{s,t}^i$ are calculated by (2) for $t = 0$ to T. Specifically, for every time step, the trend proxy dataset can be expressed as:

$$\mathbf{P_T}^i = \tau_{s,t}^i, \quad \forall t \in [0, T] \tag{11}$$

2) Seasonality Proxy Dataset: For each client i, the seasonality proxy dataset, denoted as $\mathbf{P_S}^i$, is constructed using the Fast Fourier Transform (FFT) to transform the seasonality values $\sigma_{s,t}^i$ from the time domain to the frequency domain for every t. The FFT is given by:

$$FFT[\sigma_{s,t}^i](f) = \sum_{t=0}^{N-1} \sigma_{s,t}^i e^{-j(2\pi f t/N)}, \tag{12}$$

where N is the total number of samples in $\sigma_{s,t}^i$, f represents the frequency domain components, and j is the imaginary unit. From the FFT spectrum, we identify the top 3 predominant frequency components, hypothesized to correspond to daily, weekly, and monthly seasonality. The selected frequencies, denoted as f_1, f_2, f_3, correspond to the three largest magnitudes in the FFT spectrum. Thus, the seasonality proxy dataset for client i can be expressed as:

$$\mathbf{P_S}^i = \left\{FFT[\sigma_{s,t}^i](f_k)\right\}_{k=1}^3, \forall t \tag{13}$$

Note that these proxy datasets encapsulate the necessary information about the client's data, allowing the server to perform clustering without direct access to the full local datasets. This not only facilitates efficient clustering at the server-side but also aligns with the principles of FL.

Upon receiving the aforementioned proxy datasets, the server independently conducts k-means clustering for both trend and seasonality. The number of clusters for both trend

[3] In the domain of machine learning, proxy datasets are typically summarized versions of the original datasets, designed to approximate distributions of the full data without including every detail. In our context, they encapsulate the essential features of the original mobile traffic data, allowing the server to perform clustering efficiently without directly accessing the complete data from each client.

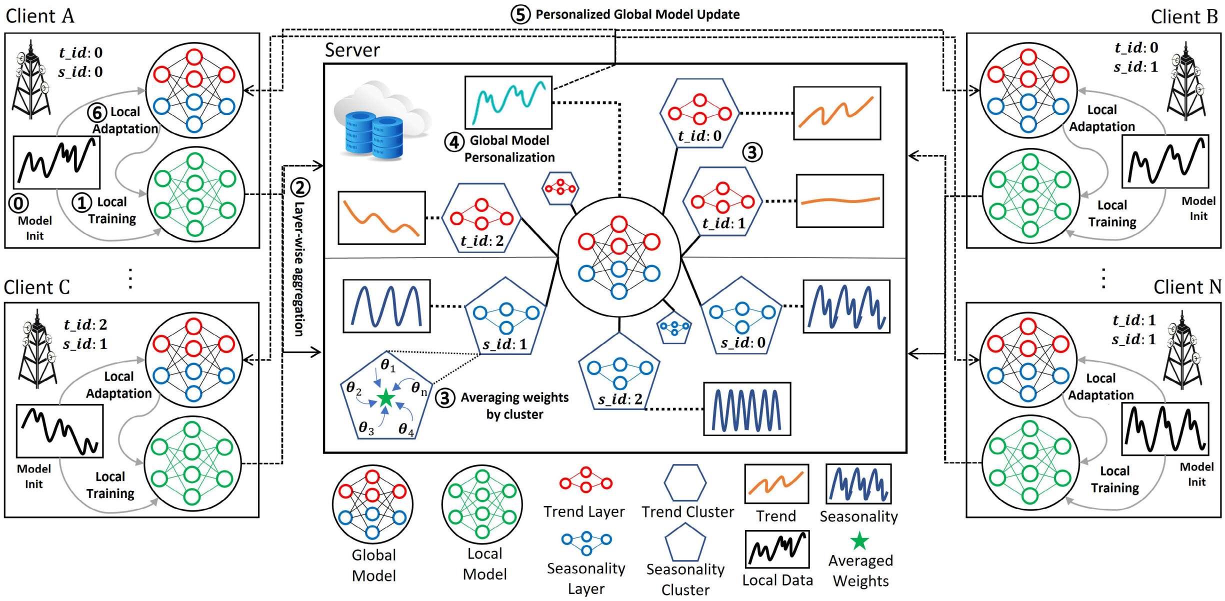


FIGURE 3. The framework of the proposed method for mobile traffic prediction.

```
Algorithm 1 Client Clustering Strategy With Decomposition Components
Input: $\mathbf{X}_{\mathbf{s},\mathbf{t}}$ - Traffic volume matrix for service type $s$
       at time point $t$ from all clients;
Output: $t\_id$ - trend cluster id for each client; $s\_id$ -
        seasonality cluster id for each client;
// Client-side
for each client $i$ do
    if given $x^i_{s,t}$ then
        Compute trend values $\tau^i_{s,t}$ by (2)
        Compute seasonality values $\sigma^i_{s,t}$ by (3)
        Compute $FFT[\sigma^i_{s,t}]$ by (12)
        Construct proxy datasets $\mathbf{P_T}^i$ and $\mathbf{P_S}^i$ by (11),
          (13), respectively
    Send proxy datasets $\mathbf{P_T}^i$ and $\mathbf{P_S}^i$ to the server
// Server-side
Perform k-means clustering on all received trend
  proxy datasets and assign $t\_id$ for each client
Perform k-means clustering on all received
  seasonality proxy datasets and assign $s\_id$ for each
  client
```

and seasonality is empirically determined based on silhouette scores[4] As an outcome of this clustering process, the server determines the t_id and s_id for each client. Our detailed clustering strategy is summarized in Algorithm 1.

B. LAYER-WISE PERSONALIZED FEDERATED LEARNING APPROACH

In traditional FL, a generic global model is trained uniformly across different data distributions. However, mobile traffic data from various regions inherently exhibit diverse temporal characteristics and patterns due to differences in local behaviors, infrastructure, population density, and other factors. This diversity can result in compromised performance when using a generalized global model.

To tackle this challenge, we introduce the layer-wise personalized FL, an approach that separately aggregates each layer decomposed into trend and seasonality. This novel approach enables fine-grained FL what each layer of the global model is learned separately. The proposed approach is illustrated in detail in Fig.3 and the whole procedure is described in Algorithm 2. Note that layer-wise freezing mechanism in Algorithm 2 will be addressed in next subsection. Specifically, our framework consists of seven steps:

1) **Model initialization:** The server initializes a model with a pre-defined architecture and random weights. Each client i then downloads this initialized model from the server.
2) **Local training:** Each client i trains locally using its dataset D^i, leading to an updated local model. The local update for client i during communication round r is given by:
$$\theta^i_{(r+1)} = \theta^i_{(r)} - \eta_1 \nabla \mathcal{L}(D^i; \theta^i_{(r)}) \tag{14}$$
where $\theta^i_{(r)}$ is the model parameter for client i at communication round r, η_1 is the learning rate, and $\nabla\mathcal{L}$ denotes the gradient of the loss function.
3) **Layer-wise aggregation:** After the local training, each client i provides its updated weights θ^i. These weights

[4] Silhouette scores measure the quality of clusters in a dataset. They indicate how similar an object is to its own cluster compared to other clusters. A higher score indicates better-defined clusters. They are commonly used to determine the optimal number of clusters.

Algorithm 2 Proposed Layer-Wise Personalized FL

```
Input: $D^i = \{\mathcal{X}_k^i, \mathcal{Y}_k^i\}_{k=1}^K$ - Mobile traffic data; $f$ -
       Fraction of the client; $R$ - Number of
       communication rounds; $\eta_1, \eta_2$ - Learning rates
       for local training and local adaption,
       respectively; $t\_id, s\_id$ - Cluster identifiers for
       trend and seasonality respectively by
       Algorithm [algo1]1; $\epsilon$ - Threshold for layer
       freezing.
Output: $\mathbf{\Theta} = \{\Theta^{local,1}, \Theta^{local,2}, \dots, \Theta^{local,I}\}$ - Set of
        clients' personalized model weights;
Initialize a global model on the server: $\theta^{init}$
for each client $i$ do
    Download the initialized model from the server:
      $\theta^i \leftarrow \theta^{init}$
for communication round $r = 1$ to $R$ do
    $p \leftarrow max(I \cdot f, 1)$
    $\mathcal{S}_i \leftarrow$ a random set of $p$ clients
    // Client-side
    for each client $i \in \mathcal{S}_i$ do
        $\theta^i_{(r+1)} \leftarrow \theta^i_{(r)} - \eta_1 \nabla \mathcal{L}(D^i; \theta^i_{(r)})$
        Selectively send $\theta^i_{(r+1)}$ to the server based on
          its freezing status
    // Server-side
    for each $\theta^i$ from client $i \in \mathcal{S}_i$ do
        if $t\_id(i) = a$ then
            $C_a^T \leftarrow C_a^T \cup \{\theta_t^i\}$
        if $s\_id(i) = b$ then
            $C_b^S \leftarrow C_b^S \cup \{\theta_s^i\}$
    $\overline{\theta}_{t,a} \leftarrow \frac{1}{|C_a^T|} \sum_{i \in C_a^T} \theta_t^i, \overline{\theta}_{s,b} \leftarrow \frac{1}{|C_b^S|} \sum_{i \in C_b^S} \theta_s^i$
      Determine layers to freeze based on $\epsilon$
    for each client $i \in \mathcal{S}_i$ do
        Obtain $\tilde{\Theta}^i$ with $\overline{\theta}_{t,a}$ and $\overline{\theta}_{s,b}$
        Send $\tilde{\Theta}^i$ to client $i$ and notify about the
          freezing status
    // Client-side again
    for each client $i \in \mathcal{S}_i$ do
        Receive $\tilde{\Theta}^i$ from the server
        $\Theta^{local,i} \leftarrow \tilde{\Theta}^i - \eta_2 \nabla \mathcal{L}(D^i; \tilde{\Theta}^i)$
```

contain information about both trend and seasonality, which can be represented as θ_t^i and θ_s^i. The server then aggregates these trend and seasonality weights separately for all clients. After aggregation, the weights are categorized and assigned to their respective clusters by t_id and s_id.

4) **Averaging intra-cluster weights:** The server averages the weights for both the trend and seasonality layers of clients that belong to the same cluster using the following equations. For the trend layer in a given cluster C_a^T:

$$\overline{\theta}_{t,a} = \frac{1}{|C_a^T|} \sum_{i \in C_a^T} \theta_t^i, \tag{15}$$

where a is the cluster identifier for the trend layer, $\overline{\theta}_{t,a}$ is the averaged weight for the trend layer, θ_t^i is the weight from client i's trend layer. For the seasonality layer in a given cluster C_b^S:

$$\overline{\theta}_{s,b} = \frac{1}{|C_b^S|} \sum_{i \in C_b^S} \theta_s^i, \tag{16}$$

where b is the cluster identifier for the seasonality layer, $\overline{\theta}_{s,b}$ is the averaged weight for the seasonality layer, θ_s^i is the weight from client i's seasonality layer.

5) **Global model personalization:** During this step, the server personalizes the global model for each participating client. For client i with specific t_id and s_id, the server fetches weights from the respective trend and seasonality clusters. Formally, for client i with $t\_id = a$ and $s\_id = b$, the personalized global model's weights are:

$$\tilde{\Theta}^i = \{\bar{\theta}_{t,a}, \bar{\theta}_{s,b}\}, \tag{17}$$

where $\bar{\theta}_{t,a}$ are the averaged weights from the trend cluster a, $\bar{\theta}_{s,b}$ are the averaged weights from the seasonality cluster b, and $\tilde{\Theta}^i$ represents the personalized weights for client i.

6) **Personalized global model update:** In this step, the server disseminates the personalized global model weights to the respective participating clients. For each client i, the server sends the corresponding weights $\tilde{\Theta}^i$ in (17). The client then updates its local model using:

$$\Theta^{updated,i} = \tilde{\Theta}^i, \tag{18}$$

where $\Theta^{updated,i}$ represents the local model weights for client i after receiving the personalized weights.

7) **Local adaption:** Upon receiving its personalized global model weights from the server, client i proceeds to fine-tune its model on its local dataset. Using these personalized weights as a starting point, the client adjusts its model to better fit its unique data distribution. This can be represented as:

$$\Theta^{local,i} = \Theta^{updated,i} - \eta_2 \nabla \mathcal{L}(D^i; \Theta^{updated,i}) \tag{19}$$

where η_2 is the local learning rate and $\nabla \mathcal{L}$ represents the loss function evaluated on the local dataset D^i. Once the local adaptation is complete, the updated weights from client i are ready to be sent back to the server for the next communication round.

To provide an intuitive understanding of our approach, we refer to a simple example as depicted in Fig.3. Consider client N which has a local dataset visualized in black. The traffic data of client N displays a nearly consistent trend with minimal fluctuations, yet it prominently exhibits a marked seasonality pattern. Through our approach, client N not only refines its model based on its own unique data but also benefits from the learned patterns of other clients with similar seasonality characteristics, specifically clients C and B.

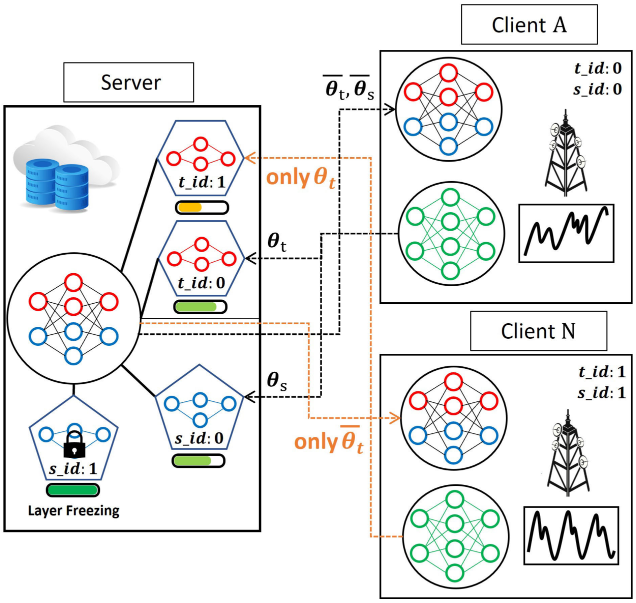


FIGURE 4. The concept of adaptive layer-wise aggregation.

C. ADAPTIVE LAYER-WISE FREEZING

To consider the communication cost in the training phase, we introduce an adaptive layer-wise freezing technique. As illustrated in Fig.4, client A typically transmits both θ_t and θ_s to the server and subsequently updates its local model with personalized global model's weights $\tilde{\Theta}^i = \{\bar{\theta}_{t,a}, \bar{\theta}_{s,b}\}$. In contrast, client N only sends θ_t owing to the frozen state of the cluster associated with the seasonality layer. Consequently, client N receives only $\bar{\theta}_{t,a}$ and updates its local model as $\tilde{\Theta}^i = \{\bar{\theta}_{t,a}, \bar{\theta}_{s,b,(r-1)}\}$. Thus, as illustrated in Fig.5, avoiding the transmission of frozen layers results in more substantial communication cost savings than the full model aggregation approach. Such a design is motivated by the fact that frequent transmission of model weights between the server and the clients can be a significant source of communication overhead.

To determine which layers of the model should be frozen and which should be updated, we perform an analysis on the magnitude of weight changes after each communication round. Those layers where the weights remain relatively stable are considered for freezing, implying that their weights will not be sent in the subsequent communication rounds. For each model layer associated with the trend cluster C_a^T, the difference between the current averaged weights $\bar{\theta}_{t,a,(r)}$ and the weights from the previous round $\bar{\theta}_{t,a,(r-1)}$ is computed by (20). Similarly, for the seasonality cluster C_b^S, the difference is computed between the current averaged weights $\bar{\theta}_{s,b,(r)}$ and the weights from the preceding round $\bar{\theta}_{s,b,(r-1)}$ by (21). If the difference is under a predetermined threshold ϵ, the corresponding layer is frozen:

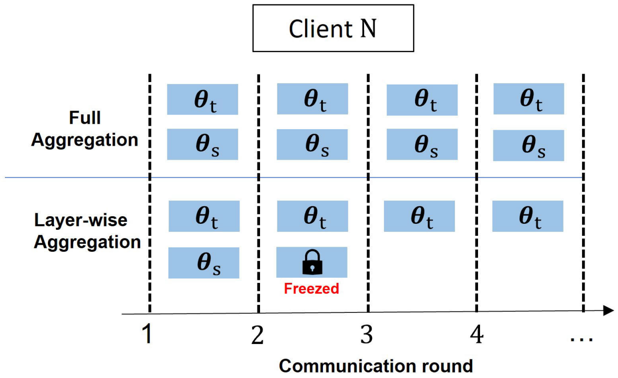


FIGURE 5. The comparison of our layer-wise aggregation and full model aggregation.

$$\Delta\theta^t = ||\bar{\theta}_{t,a,(r)} - \bar{\theta}_{t,a,(r-1)}||_2 \tag{20}$$

$$\Delta\theta^s = ||\bar{\theta}_{s,b,(r)} - \bar{\theta}_{s,b,(r-1)}||_2 \tag{21}$$

$$\text{if } \Delta\theta^t < \epsilon \text{ or } \Delta\theta^s < \epsilon, \text{ then freeze the corresponding layer.} \tag{22}$$

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conducted extensive experiments to validate the effectiveness of our proposed method for mobile traffic prediction. We detail a prediction model used for FL, along with the experimental settings. Subsequently, we explain the evaluation and baselines for comparison. We then present a comparative analysis of the prediction performance between the proposed method and baselines on various kinds of mobile traffic. Notably, predictions and ground truth from arbitrarily chosen cells within the city are provided. Moreover, we emphasize how our method outperforms in both prediction accuracy and communication efficiency.

A. FEDERATED MODEL AND EXPERIMENT SETTINGS

We utilized *Dlinear* [22] as our federated model, which has recently demonstrated remarkable performance in time series forecasting. *Dlinear* is a decomposition scheme model that separates the model into trend and seasonality layers. We trained the *Dlinear* model to predict mobile traffic for each region and exchanged the model weights during the FL process.

For this experiment, ensuring no loss of generality, we randomly selected 100 cells from each dataset and conducted experiments on three types of mobile traffic from these cells. A span of seven weeks was used for training and the last week for testing. Both the training and test datasets were constructed using a sliding window scheme with a sequence length $o = 72$ and a prediction length $p = 1$. For the *Dlinear* model, we set the hyperparameters as follows: the number of channels c was set to 1 and the kernel size was set to 25 as usual. The FL-related parameters are as follows: The fraction of the client f, determining the number of participating clients in the training, was set to 0.1. The model was trained for 100 consecutive communication rounds R with a batch size of 32, utilizing the Adaptive Moment Estimation (Adam) optimizer with learning rates $\alpha = 0.01$. During local adaptation, we also employed Adam with a learning rate of $\alpha' = 0.0001$. The whole parameters used in experiments are summarized in Table 1 for quick understanding.

TABLE 1. Experiment setup parameters.

Parameter	Value
Traffic types	3
Number of selected cells	100
Training duration	7 weeks
Testing duration	1 week
Window sequence length (o)	72
Prediction length (p)	1
Channels (c)	1
Kernel size	25
Client fraction (f)	0.1
Communication rounds (R)	100
Batch size	32
Optimizer	Adam
Learning rate (α)	0.01
Local adaption learning rate (α')	0.0001

B. BASELINES AND EVALUATION METRICS

To demonstrate the effectiveness of our proposed method, we primarily compare it with other widely used FL methods. For a comprehensive evaluation, we also include comparisons with popular statistical and DL models. The baseline methods used in our study are described as follows:

- *HA* [13]: Utilizes past traffic data to predict future patterns through simple averaging, serving as a baseline for its simplicity and direct approach.
- *ARIMA* [14]: A widely recognized basic model that captures various aspects of time series data such as mobile traffic.
- *SARIMA* [16]: An extension of ARIMA that specifically incorporates seasonality, enhancing prediction accuracy for traffic patterns with clear cyclical behaviors.
- *MLP* [23]: A fundamental neural network architecture for regression, offering a solid comparison point for traffic prediction models. In our experiments, the MLP model is designed with three fully connected layers, configured with 64, 128, and 64 units, respectively.
- *GRU* [24]: Optimized for sequential data, GRU is capable of capturing temporal dependencies of traffic with a more efficient training process than traditional RNNs. In our experiments, the GRU model comprises three hidden layers each with 64 units.
- *LSTM* [25]: Designed to capture short-term and long-term dependencies in time series data such as traffic prediction. In our experiments, the LSTM model comprises five hidden layers each with 64 units.
- *FedAvg* [7]: It is a foundational approach in FL that averages model parameters across multiple clients. It is known for its simplicity in many federated scenarios.
- *FedAtt* [20]: Introducing an attention mechanism, this method refines the aggregation process in FL. By doing so, it gives more weight to more relevant client models, potentially enhancing the overall model accuracy.
- *FedDA* [12]: It utilizes a unique data augmentation approach to handling data heterogeneity of mobile traffic in FL. By grouping clients into clusters based on augmented datasets and using a dual attention mechanism, it achieves a more accurate model aggregation compared to simply averaging weights.

We evaluate our method and baselines with two commonly used metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE is a statistical metric that calculates the average of the squares of the errors or deviations between the predicted values and the ground truth values. MAE is a statistical metric used to quantify the average magnitude of the errors between the predicted values and the corresponding ground truth values. These two metrics are defined as follows:

$$MSE = \frac{1}{I}\sum_{i=1}^{I}\sum_{t=1}^{T}(\mathcal{Y}_t^i - \hat{\mathcal{Y}}_t^i)^2, \tag{23}$$

$$MAE = \frac{1}{I}\sum_{i=1}^{I}\sum_{t=1}^{T}\left|\mathcal{Y}_t^i - \hat{\mathcal{Y}}_t^i\right|, \tag{24}$$

where $\hat{\mathcal{Y}}_t^i$ and $\mathcal{Y}_t^i$ are the predicted value and ground truth of mobile traffic for each client i at time step t, *i.e.,* each client represents a region in the target city, and I is the total number of clients.

TABLE 2. Performance comparisons on SMS traffic.

Strategy	Methods	Milano		Trento	
		MSE	MAE	MSE	MAE
Statistical-based (Centralized)	HA	0.8580	0.7111	5.1515	1.5182
	ARIMA	0.7819	0.6358	4.8181	1.3355
	SARIMA	0.7412	0.6255	4.7550	1.3212
Deep learning-based (Centralized)	MLP	0.5983	0.4465	3.6137	0.9844
	GRU	0.5697	0.4312	3.5634	0.9428
	LSTM	0.5581	0.4264	3.5134	0.9333
Deep learning-based (Federated)	Proposed	**0.3067**	**0.3403**	**1.8586**	**0.7512**

C. COMPARISONS OF PREDICTION PERFORMANCE

First, we present a detailed performance comparison of our proposed method against traditional statistical methods and centralized DL models. Tables 2,3, and 4 show the results of these comparisons for SMS, Call, and Internet traffic, respectively. For SMS traffic in Milano, our proposed method achieves a 58.62% gain in MSE and a 45.60% gain in MAE over the best-performing statistical-based method (*SARIMA*).

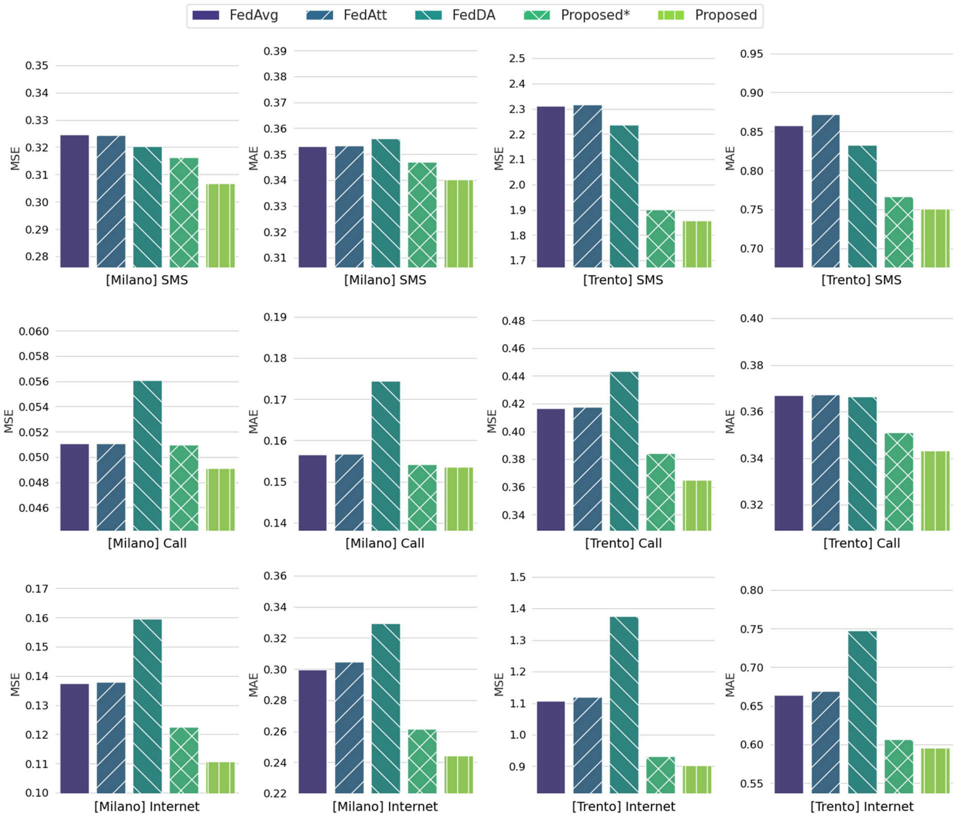


FIGURE 6. Comparison of prediction performance for three mobile traffic types in Milano and Trento.

TABLE 3. Performance comparisons on call traffic.

Strategy	Methods	Milano		Trento	
		MSE	MAE	MSE	MAE
Statistical-based (Centralized)	HA ARIMA SARIMA	0.3421 0.2948 0.2505	0.5217 0.4447 0.3353	1.9823 1.8644 1.7416	1.0056 0.9541 0.7185
Deep learning-based (Centralized)	MLP GRU LSTM	0.1499 0.1365 0.1345	0.2553 0.2435 0.2411	1.2145 1.1394 1.1344	0.5237 0.5053 0.5000
Deep learning-based (Federated)	Proposed	**0.0491**	**0.1535**	**0.3651**	**0.3433**

TABLE 4. Performance comparisons on internet traffic.

Strategy	Methods	Milano		Trento	
		MSE	MAE	MSE	MAE
Statistical-based (Centralized)	HA ARIMA SARIMA	0.4315 0.3142 0.2873	0.5466 0.5284 0.4450	6.8441 6.1207 5.9583	1.9275 1.7542 1.4470
Deep learning-based (Centralized)	MLP GRU LSTM	0.1785 0.1715 0.1648	0.3014 0.2987 0.2900	4.9574 4.6971 4.6127	1.2101 1.1995 1.1057
Deep learning-based (Federated)	Proposed	**0.1108**	**0.2443**	**0.9040**	**0.5962**

For Call traffic, it shows an 80.40% gain in MSE and a 54.22% gain in MAE. For Internet traffic, it shows a 61.43% gain in MSE and a 45.10% gain in MAE. When compared to the best-performing DL-based method (*LSTM*), our method achieves a 45.05% gain in MSE and a 20.19% gain in MAE for SMS traffic. For Call traffic, the gains are 63.49% in MSE and 36.33% in MAE. Lastly, for Internet traffic, the gains are 32.77% gain in MSE and 15.76% gain in MAE.

Statistical-based methods are effective in predicting regular traffic patterns but often fail to capture the full range of traffic across different regions. While DL-based methods excel in generalizing across diverse traffic patterns, they typically struggle with unique and rare temporal patterns (e.g., traffic spikes). However, our method leverages FL to train a global model, ensuring generalization that covers traffic patterns across all regions. Additionally, our approach personalizes the global model through layer-wise aggregation, where regions with similar temporal characteristics are grouped together for training. The combination of generalization and personalization capabilities is the key reason why our method achieves superior prediction performance. It adeptly captures both the broad regular traffic applicable across all regions and the unique, region-specific patterns crucial for precision in each local prediction.

After comparing our method's performance with traditional statistical-based and DL-based methods, the next phase of our analysis focuses on its efficacy relative to other FL baselines. The results, as depicted in Fig.6, extend our evaluation to three distinct types of mobile traffic. Among all the sub-figures, the first row represents results from Milano and Trento for the SMS dataset in terms of MSE and MAE. Similarly, the second row represents results for the Call dataset, and the third row represents results for the Internet dataset.

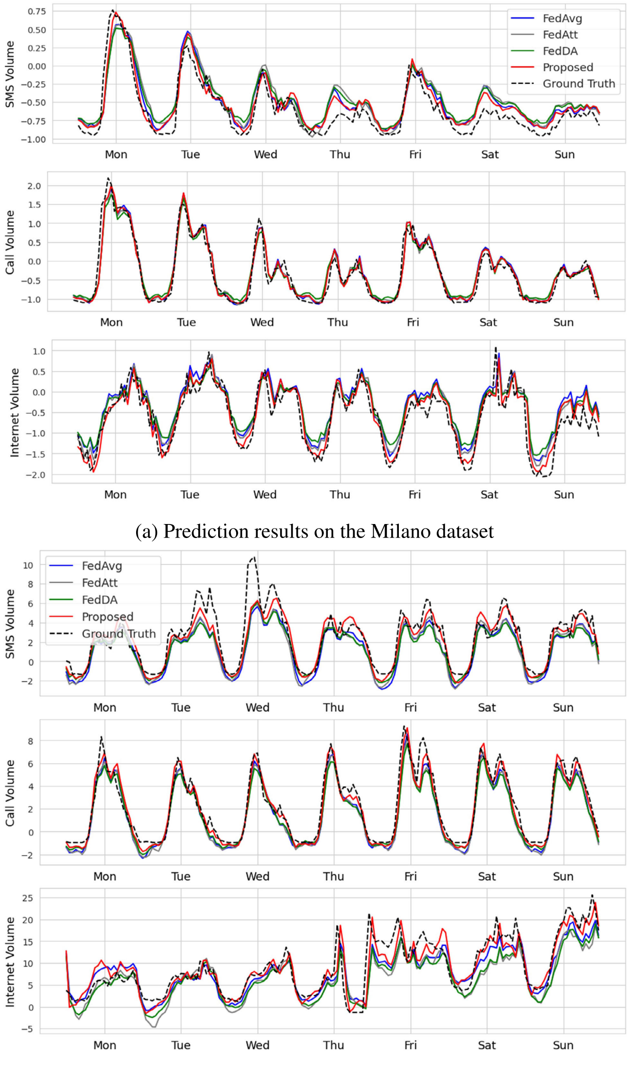
As depicted in Fig.6, our proposed method consistently outperforms all baselines across both datasets. Specifically, we observe significant performance gains when compared to the best-performing baselines for each dataset. For the SMS dataset in Milano, our method achieves gains of 4.27% and 4.46% in MSE and MAE metrics over *FedDA*, respectively. For the Call dataset in Milano, our method achieves gains of 3.57% in MSE and 1.78% in MAE over *FedAvg*. For the Internet dataset in Milano, our method achieves gains of 16.68% in MSE and 16.77% in MAE against *FedAvg*. Similarly, for the SMS dataset in Trento, our method achieves gains of 16.95% and 9.75% in MSE and MAE metrics over *FedDA*, respectively. For the Call dataset in Trento, our method achieves gains of 11.63% in MSE and 6.44% in MAE over *FedAvg*. Lastly, for the Internet dataset in Trento, our method achieves gains of 14.79% in MSE and 9.11% in MAE against *FedAvg*.

The performance enhancements observed are primarily due to two factors: 1) Our method enables a fine-grained federation among regions that share similar temporal characteristics. 2) Our method trains the global models tailored for local traffic data in each specific region. This personalized approach yields more accurate predictions compared to a coarse-grained global model designed for all regions. Note that there is a variation of our proposed method, represented as proposed*, specifically introduced to evaluate the impact of local adaptation. This variant omits the local adaptation process and focuses solely on personalization at the server side. By comparing the performance of this variation, we can ascertain the effectiveness of our local adaptions step.

To further evaluate the predictive capabilities of our method and the baseline models, prediction results and ground truths are presented in Fig.7. The x-axises denote the time index of the test dataset and the y-axises are the mobile traffic volume. The cells are randomly selected for each service type in the Milano and Trento datasets, respectively. Each sub-figures show our method can accurately predict the values for all the kinds of mobile traffic. Especially, we can see that our proposed method demonstrates not only enhanced accuracy in predicting peaks and troughs compared to baselines but also maintains a minimal error when irregular patterns occurs. This can be attributed to our approach's ability to better capture the local characteristics of a specific region.

D. PERSONALIZATION ANALYSIS

In Fig.9, we explore the trained layers across clusters to verify model personalization. The weights of the layers are visualized as a 2D heatmap. The x-axis represents the length of the input data, corresponding 72 time index in the past. The y-axis represents the length of the output prediction, corresponding to the prediction value for the subsequent time index. This heatmap provides a visual representation of how much each point in the input data affects the predicted value.



(a) Prediction results on the Milano dataset

(b) Prediction results on the Trento dataset

FIGURE 7. Prediction results and ground truth compared to the baseline for randomly selected cells in each of the two citys.

Fig.9a shows the trend layer's attention to the input sequence for each cluster. For cluster 0, the attention given to the data around time point 60 significantly influences predictions. This reveals the model's effective learning from areas where recent trends remain consistent. On the other hand, cluster 1 places its attention more on the latest data, indicating the recent data strongly influences the prediction. This reveals the model's effective learning from areas where recent trends change significantly. In Fig.9 (b), we can pinpoint the areas where the seasonality layer directs its attention within the input sequence for each cluster. The visualization of the seasonal layer reveals the nature of periodic patterns learned by each cluster. Cluster 1 shows a strong periodic pattern with a cycle close to 24 units,

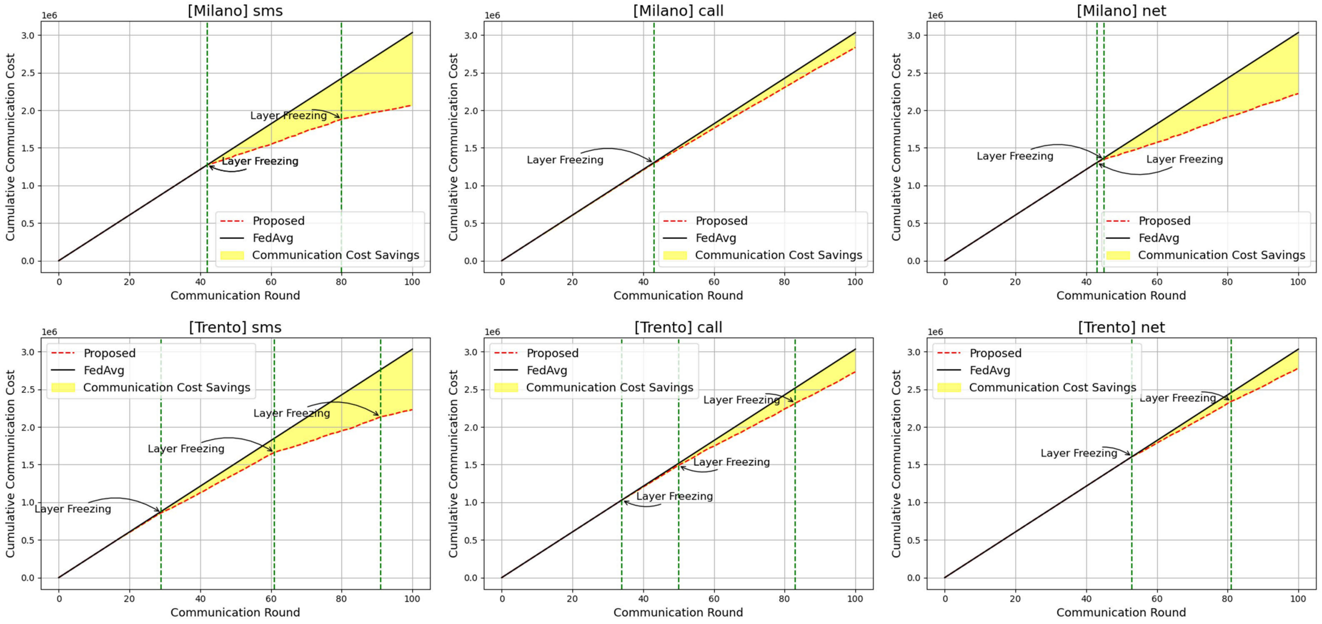
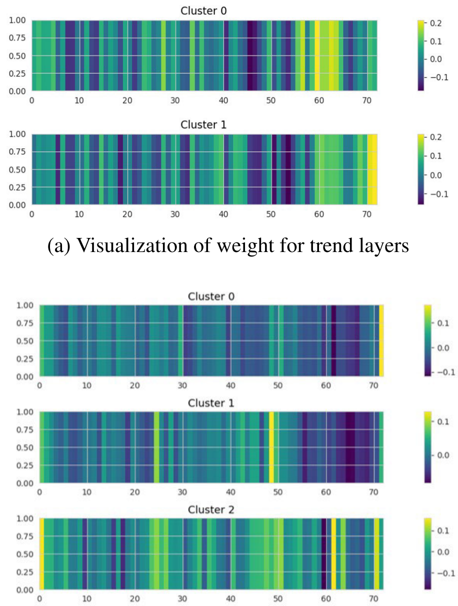


FIGURE 8. The communication cost induced during whole training.



(a) Visualization of weight for trend layers

(b) Visualization of weight for seasonality layers

FIGURE 9. Visualization of weight for trend and seasonality layers.

while cluster 0 shows a periodicity close to 20 units. Conversely, cluster 2 shows a weak periodic pattern and more irregular pattern.

The visualization results demonstrate that each cluster has effectively learned the temporal patterns for specific regions. By combining these learned layers on the server, a personalized global model is tailored to the characteristics of each region. We conclude that our personalized approach provides a more precise learning of regional mobile traffic patterns, yielding improved prediction accuracy.

E. COMMUNICATION COST ANALYSIS

To evaluate the communication efficiency of our proposed method, we measured the communication cost throughout the entire learning process, as illustrated in Fig.8. The x-axis denotes the communication rounds, representing global FL rounds with clients from diverse regions, while the y-axis indicates the communication cost required for training. The vertical green line represents the points at which layer freezing occurs (i.e., when communication costs start to save). It is evident that the earlier and more frequent layer freezing occurs, the greater the efficiency in terms of communication costs.

The varied timing of layer freezing across the data can be attributed to the different complexities associated with learning mobile traffic data patterns for various service types. Layers in clusters that learn simpler trend and seasonality patterns tend to converge and freeze faster than their counterparts. This early layer freezing indicates the presence of consistent patterns in the traffic of specific regions, suggesting that some service traffic exhibits stable trends and distinct periodic behaviors (e.g., traffic patterns with minimal fluctuations in trend or those showcasing pronounced seasonality).

VI. DISCUSSION

In this study, we presented a novel personalized FL framework, specifically tailored for mobile traffic prediction. When comparing our results with existing statistical methods and DL methods, including other FL approaches, several key distinctions become evident. Our approach has the capability for both generalization and personalization, offering accurate predictions of each region's unique temporal characteristics.

This contrasts with traditional statistical models, which typically lack the ability to capture complex non-linear relationships. On the other hand, DL methods excel in generalization but often fall short in personalization, leading to lower accuracy for atypical patterns in specific regions. Moreover, our method stands in contrast to existing FL approaches. Existing FL methods typically train a single global model across all regions, which may not efficiently capture unique regional traffic characteristics. Our personalized FL framework, however, separately trains each layer of the global model. This fine-grained approach allows for a more granular understanding of mobile traffic patterns, thereby achieving a prediction model that is precisely tailored to regional specifics.

However, it is important to note that the practical implementation of our method necessitates computational resources at the base station or edge server in each region. Looking forward, with the progression towards beyond 5G and 6G networks, anticipated advancements in edge computing are expected to significantly enhance the decentralization of computing resources [8], [26]. This evolution will likely make our approach even more feasible and effective by providing the necessary computational resources at the local level, further bolstering the capabilities of implementations [27].

Unfortunately, our study used 4G data due to the lack of publicly available datasets for more advanced networks like 5G. However, it is essential to recognize that with the advent of the beyond 5G and 6G era, mobile traffic is expected to display increasingly complex patterns [28]. This complexity will stem from a broader range of communication and mobility patterns generated by a variety of devices, including IoT devices, AR and VR systems, and autonomous vehicles [29]. Such diversity in device types and their respective applications is likely to lead to significant regional variations in traffic patterns. Traditional single-global model approaches are likely to face limitations in capturing these regional characteristics and making accurate traffic predictions. In contrast, our proposed personalized approach is designed to adaptively learn models that are specifically tailored to the unique characteristics of each region. We anticipate that this approach will be increasingly effective in handling the evolving complexity of mobile traffic data over time.

VII. CONCLUSION

In this paper, we focus on FL-based mobile traffic prediction for future mobile networks. There are two primary challenges in this domain: First, data heterogeneity stems from the varied communication and mobility patterns across regions, posing a challenge for maintaining consistent predictive performance. Second, the challenge of communication efficiency arises as the frequent transmission of a large number of model parameters significantly increases communication costs. To address these challenges, we propose a layer-wise FL framework that enables fine-grained training of each layer in the global model. This fine-grained federation personalizes the model for each region, effectively addressing the data heterogeneity issue. Additionally, we integrate an adaptive layer freezing mechanism that selectively updates model layers based on their convergence, reducing communication costs during training. The experiment results highlight the superiority of our approach, attributing the enhanced performance to its robust personalization capabilities that adeptly capture rare and unique temporal patterns across different regions. Additionally, our method significantly reduces communication costs by strategically updating model parameters only when necessary, without compromising prediction accuracy.

In future work, we intend to explore additional personalization techniques to capture rare traffic patterns in specific region. Furthermore, a comprehensive convergence analysis of layer-wise FL is planning to be explored. Proving the convergence of our approach theoretically would be an interesting direction of future research. Additionally, an intriguing extension of this work for future study is the exploration of joint problems, such as addressing traffic congestion, where our method could be further applied.

REFERENCES

[1] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electron.*, vol. 3, no. 1, pp. 20–29, Jan. 2020.

[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] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.

[4] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, and S. Cui, "Spatio-temporal wireless traffic prediction with recurrent neural network," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 554–557, Aug. 2018.

[5] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1656–1659, Aug. 2018.

[6] J. Konečný, H. Brendan McMahan, F. X. Yu, P. Richtárik, A. Theertha Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.

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

[8] Q. Duan, J. Huang, S. Hu, R. Deng, Z. Lu, and S. Yu, "Combining federated learning and edge computing toward ubiquitous intelligence in 6G network: Challenges, recent advances, and future directions," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2892–2950, 4th Quart., 2023.

[9] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2019, pp. 1387–1395.

[10] K. He, X. Chen, Q. Wu, S. Yu, and Z. Zhou, "Graph attention spatial–temporal network with collaborative global-local learning for citywide mobile traffic prediction," *IEEE Trans. Mobile Comput.*, vol. 21, no. 4, pp. 1244–1256, Apr. 2022.

[11] L. Zhang, C. Zhang, and B. Shihada, "Efficient wireless traffic prediction at the edge: A federated meta-learning approach," *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1573–1577, Jul. 2022.

[12] C. Zhang, S. Dang, B. Shihada, and M.-S. Alouini, "Dual attention-based federated learning for wireless traffic prediction," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2021, pp. 1–10.

[13] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2018.

[14] B. K. Nelson, "Time series analysis using autoregressive integrated moving average (ARIMA) models," *Academic Emergency Med.*, vol. 5, no. 7, pp. 739–744, 1998.

[15] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, and Y. Li, "Big data driven mobile traffic understanding and forecasting: A time series approach," *IEEE Trans. Services Comput.*, vol. 9, no. 5, pp. 796–805, Sep./Oct. 2016.

[16] S. Medhn, B. Seifu, A. Salem, and D. Hailemariam, "Mobile data traffic forecasting in UMTS networks based on SARIMA model: The case of addis ababa, Ethiopia," in *Proc. IEEE AFRICON*, Sep. 2017, pp. 285–290.

[17] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.

[18] S. P. Sone, J. J. Lehtomäki, and Z. Khan, "Wireless traffic usage forecasting using real enterprise network data: Analysis and methods," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 777–797, 2020.

[19] Q. Zeng, Q. Sun, G. Chen, H. Duan, C. Li, and G. Song, "Traffic prediction of wireless cellular networks based on deep transfer learning and cross-domain data," *IEEE Access*, vol. 8, pp. 172387–172397, 2020.

[20] S. Ji, S. Pan, G. Long, X. Li, J. Jiang, and Z. Huang, "Learning private neural language modeling with attentive aggregation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[21] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Sci. Data*, vol. 2, no. 1, 2015, Art. no. 150055.

[22] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 9, pp. 11121–11128.

[23] A. Y. Nikravesh, S. A. Ajila, C.-H. Lung, and W. Ding, "Mobile network traffic prediction using MLP, MLPWD, and SVM," in *Proc. IEEE Int. Congr. Big Data (BigData Congress)*, Jun. 2016, pp. 402–409.

[24] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[25] H. D. Trinh, L. Giupponi, and P. Dini, "Mobile traffic prediction from raw data using LSTM networks," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1827–1832.

[26] E. Kartsakli, J. Perez-Romero, O. Sallent, N. Bartzoudis, V. Frascella, S. K. Mohalik, T. Metsch, A. Antonopoulos, Ö. F. Tuna, Y. Deng, X. Tao, M. A. Serrano, and E. Quiñones, "AI-powered edge computing evolution for beyond 5G communication networks," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Jun. 2023, pp. 478–483.

[27] S. Iftikhar, S. S. Gill, C. Song, M. Xu, M. S. Aslanpour, A. N. Toosi, J. Du, H. Wu, S. Ghosh, D. Chowdhury, M. Golec, M. Kumar, A. M. Abdelmoniem, F. Cuadrado, B. Varghese, O. Rana, S. Dustdar, and S. Uhlig, "AI-based fog and edge computing: A systematic review, taxonomy and future directions," *Internet Things*, vol. 21, Apr. 2023, Art. no. 100674.

[28] D. Sabella, D. Micheli, and G. Nardini, "The power of data: How traffic demand and data analytics are driving network evolution toward 6G systems," *J. Sensor Actuator Netw.*, vol. 12, no. 4, p. 49, Jun. 2023.

[29] J. R. Bhat and S. A. Alqahtani, "6G ecosystem: Current status and future perspective," *IEEE Access*, vol. 9, pp. 43134–43167, 2021.

SEUNGYEOL LEE received the B.S. degree in applied IT and engineering from Pusan National University, in 2018, and the M.S. degree in information and communication engineering from the University of Science and Technology (UST), Daejeon, South Korea, in 2020, where he is currently pursuing the Ph.D. degree in information and communication engineering. His current research interests include network intelligence, 5G/6G, and distributed learning.

JIHOON SUNG received the B.E. degree in electrical engineering from Chungnam National University, Daejeon, South Korea, in 2008, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2010 and 2016, respectively. He is currently a Senior Researcher with the Standards Research Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon. His current research interests include mobile network intelligence, content delivery networks, and wireless network control technologies.

MYUNG-KI SHIN received the Ph.D. degree in computer engineering from Chungnam National University, in 2003, with a focus on IPv6 multicast and mobility. From 2004 to 2005, he was a Guest Researcher with NIST, USA. He is currently a Principal Researcher with the Electronics and Telecommunications Research Institute (ETRI), South Korea. He is a Professor with the ETRI School, University of Science and Technology (UST). He is also the Technical Leader of 5G/6G network and system standardization projects sponsored by Korean Government in ETRI. He has been working on Internet and mobile/wireless protocols, since 1994. He is the author of several IETF RFCs, such as RFC 3338, RFC 4038, RFC 4489, and RFC 5181. He is actively involved in 3GPP SA and SA2. His research interests include 5G/6G, future internet, mobility, network virtualization, and SDN/NFV technologies.