Mobile Network Traffic Prediction Based on User Behavior and Machine Learning

Mohammad Sadegh Rostami[1], Ali Shahzadi*[2]

***Abstract*— One of the challenges of any network is user traffic analysis and prediction. Mobile networks have achieved notable growth in recent years. Traffic information on these networks play a crucial role in service quality control, user access control, and optimization. There are various methods for traffic prediction, including harmonic analysis and mathematical transformations, time series methods, and machine learning. Due to the increasing volume of user data on mobile networks, machine learning methods have gained popularity in recent decades. In this paper, we introduce a probabilistic behavioral model and propose a new method that focuses on human behavior by categorizing users through clustering and utilizing their similarities. Assuming a history of past user data is available, users with similar behavior are grouped into categories using clustering methods, and each category is assigned a label. The average of each cluster represents the traffic in that category. To predict the traffic of new users, our proposed method utilizes classification functions to determine the most appropriate category. Subsequently, weighted averages are used to calculate the overall network traffic. We compare our proposed method with time series and Fourier transform methods through three different scenarios. The results indicate that our method exhibits significant superiority over the other methods.**

***Index Terms*— Prediction, Traffic, Mobile Network, User Behavior, Machine Learning, Probabilities.**

I. INTRODUCTION

Network traffic is one of the parameters that helps with better design and more accurate analysis of networks. Different users have different traffic patterns, which are essential for the system administrator to know. The growth of mobile network traffic over the years exhibits a graph that closely resembles an exponential function. [1], and it is always increasing. In this context, we define traffic as the amount of data transferred from one point to one or more points within a specific time period. Understanding traffic information has various applications, including the management of available resources such as bandwidth, delay, capacity, and the number of channels. Additionally, it aids in controlling new user acceptance, user entry and exit from the network, and ensuring quality of service.

The activities associated with network traffic can be categorized into various areas, including traffic analysis, traffic prediction, and the optimization of specific network parameters based on traffic information. Some other applications include: finding user data usage patterns [2], which applied an unsupervised machine learning approach to cluster mobile user types by considering various factors and demonstrating the correlation between the types of applications installed by users and their data consumption patterns; classifying user types [3] which proposed a framework for mobile big data to handling the immense amount of data traffic by providing massive data traffic collection, storage, processing, analysis, and management functions; investigating user behavior [4] which utilized passive sensing data from mobile phones to investigate the extent to which within-person variability in behavioral patterns can predict self-reported personality traits, collected data from 646 college students over a period of 14 days, and identified several significant correlations between the features of within-person variability and the self-reported personality traits; detecting network anomalies [5] which used a new hybrid model based on various anomaly detection techniques such as GARCH, K-means, and Neural Network to determine the anomalous CDR data; traffic engineering [6,7] which proposed a heuristic traffic engineering approach in Software-Defined Networking (SDN) based on multipath forwarding and switching of flows between paths, and CFR-RL (Critical Flow Rerouting-Reinforcement Learning) scheme that learns a policy to select critical flows for each given traffic matrix automatically in SDN, respectively; traffic smoothing [8] When a significant amount of mobile device traffic becomes concentrated in a wireless access network at a specific time, user throughputs experience a significant decrease, leading to a decay in communication quality, so they proposed a new mechanism that instructed users to delay their traffic, aiming to shift a portion of the peak-time traffic to off-peak time in order to temporarily smooth the traffic. Long-term network development [9], in which this study analyzes and forecasts the number of 5G (5-generation) users using a logistic model, consumer preference for 5G services using a mixed logit model, and 5G data traffic increase using sensitivity analysis; allocating resources to users by short-term predictions [10,11], which extended an algorithm by introducing a new approach that incorporates personalized pricing based on the load pressure imposed by an allocation decision on a radio access technology and examined the impact of the arrival distribution of primary user traffic on secondary users specifically in terms of secondary user network performance, respectively; spectrum or power optimization [12] which proposed and analyzed dynamic spectrum access

1- Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran.

2- Associate Professor, Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran.

Corresponding author: shahzadi@semnan.ac.ir

scheme by also taking balking and reneging behavior into account, etc. Hence, acquiring comprehensive knowledge about network traffic is of utmost importance. Furthermore, different methods are employed based on the type of data. In the following section, some of these methods are mentioned.

Based on the research conducted by the authors of this paper, the prevailing approach to traffic analysis has traditionally been top-down or holistic. In other words, the focus has been on studying the overall data transfer traffic within a large-scale network or its specific segments [13]. So far, only a limited number of researchers have addressed the issue of network traffic from the perspective of end users. This approach involves considering the traffic generated by individual users and aggregating it to analyze or predict the overall network traffic. While some efforts have been made to identify the behavior patterns of various devices (machine users) and human users and extract relevant patterns [14,15], adopting a bottom-up perspective (traffic analysis from the end user's point of view) opens up diverse research prospects. One notable advantage of the bottom-up perspective is its ability to adapt to evolving user behavior over time. As users move from one location to another, their historical traffic data can be transferred, thus ensuring the analysis remains up-to-date [16].

Numerous nodes in mobile networks gather diverse user data, including time stamps, location, mobile data consumption, and types of data transfers. System administrators are able to obtain a more profound understanding of individual users because of the development of data-driven methodologies and the creation and preservation of massive amounts of user data. These data may be analysed to gain a deeper comprehension of user behavior.

The unpredictability of user behavior poses a challenge to network traffic prediction. If a user exhibits deterministic behavior, prediction becomes relatively straightforward. However, obtaining more detailed information about a user and understanding the factors that influence their behavior can lead to more accurate predictions of their future actions based on past behavior and current conditions. In this regard, the utilization of probabilistic graph models, in combination with machine learning methods, emerges as a viable approach to accomplishing this objective.

The main purpose of this paper is to propose a methodology for predicting network traffic from the perspective of end users, incorporating user behavior, probabilities, and machine learning techniques. By leveraging these tools, we aim to enhance the accuracy of network traffic prediction.

II. A REVIEW OF TRAFFIC PREDICTION METHODS

The mobile network provides a wealth of diverse and extensive information compared to other networks, primarily due to its intended purpose of facilitating voice calls, messaging, and internet data applications. In addition to internet data, call detailed record (CDR) data is also available, which includes essential information such as user location during calls, call duration, connection time, and the number of text messages exchanged. These records are of significant importance to mobile service providers. Therefore, the collection, storage, and analysis of this data hold great importance. Given the various types of traffic data available, a range of methods are employed for their analysis and prediction. In this section, we briefly discuss five of the most commonly used methods, taking into account the diversity of data analysis and prediction approaches. Table I provides a concise overview of some applications associated with each of these methods.

A. Machine learning

The utilization of artificial intelligence (AI) has experienced an unprecedented expansion, driven by advancements in powerful processors, the collection of diverse datasets, and the development of robust mathematical methods. These three pillars—powerful processors, abundant data, and mathematical principles—serve as the fundamental foundations of AI. Within the realm of AI, machine learning (ML) emerges as a prominent subset. In this approach, machines are trained through the utilization of data and mathematical models. Machine learning methods encompass various aspects, including supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning. Each of these aspects serves a distinct purpose within the field. Machine learning methods find wide-ranging applications, such as classification, clustering, regression, dimensionality reduction, and association rules, among others. These applications are implemented through diverse methods, including decision trees, k-nearest neighbors, random forests, support vector machines, neural networks, deep learning, and more.

In their work [17], the authors have described the applications of machine learning in network traffic analysis. Additionally, the authors [18] have provided insights into the potential, limitations, and future prospects of employing machine learning techniques in the context of 5th-generation (5G) mobile networks and beyond. Moreover, a comparative analysis of various machine learning methods for wireless network traffic prediction was conducted by the authors in [19]. Furthermore, [13] serves as a valuable review paper that focuses on the utilization of machine learning for mobile network traffic prediction. In [20], the authors developed a deep learning model that integrates stationary processes into a well-designed hierarchical structure and models non-stationary time series using multi-scale stable features for 5G mobile network long-term traffic prediction.

B. Mathematical Transforms

It is commonly known that mathematical transformations are effective tools for data analysis because they allow signals to be converted from one domain—like time—to another, like frequency or scale. For signal processing, there are many different transformations available; the most well-known are the Fourier and Wavelet transformations. Different types exist for signals, such as discrete and continuous signals, as well as periodic and non-periodic signals. The examination of traffic data can also benefit from these categories. Network traffic signals can sometimes show self-similarity, which means that certain data will recur with only minor changes in the future.

In order to demonstrate how traffic can be predicted using the Fourier transform, the authors [21] identified various periodicities by examining the behavior of mobile network traffic in a large region. By performing the discrete Fourier transform (DFT) on the massive traffic data, they discovered that the magnitude of the signal in the frequency domain exhibits three peaks. By analyzing the frequency domain, it was

determined that these three peaks correspond to half-day, daily, and weekly intervals. Consequently, by analyzing the peaks of the traffic graph in the frequency domain, periodic traffic behaviors can be extracted.

The Fourier transform is a valuable tool for analyzing signals over a given time interval. However, when dealing with signals that contain time-varying periodic components, where the periodicity changes over time, the traditional Fourier transform falls short in identifying the points of transition. To address this challenge, researchers have introduced the short-time Fourier transform. The short-time Fourier transform overcomes this limitation by dividing the signal into shorter segments using a fixed-length window. By applying the Fourier transform to these smaller time intervals and fewer samples, it becomes possible to capture the changes in periodicity over time. While this method proves effective, it does come with its own set of challenges. For instance, the choice of window shape can impact the signal in the frequency domain, and selecting an appropriate window length presents another consideration.

To address the aforementioned challenges, the wavelet transform offers a viable solution. The wavelet transform is an alternative transformation commonly employed in network traffic analysis, either as a standalone method or in conjunction with other techniques [22]. In addition to overcoming the limitations of the Fourier transform, the wavelet transform boasts several advantages. Unlike the Fourier transform, the wavelet transform facilitates the transfer of signals not only from the time domain to the frequency domain but also to the scale domain, where scale can be considered as an equivalent to frequency. Furthermore, the wavelet transform incorporates time-shifting capabilities. One of the key strengths of the wavelet transform lies in its ability to provide sufficient resolution in both the time and frequency domains. It achieves this by allocating longer time spans to lower frequencies and shorter time spans to higher frequencies, thereby enabling a more precise analysis across different scales.

C. Time Series

Time series components can be classified into four main categories: (1) Trend, which represents the overall tendency of the time series to increase, decrease, or remain constant over the entire duration; (2) Cycle, which refers to repetitive and similar changes occurring in the midterm; (3) Seasonality, which encompasses recurring patterns that take place within periods shorter than the cycle; and (4) Irregular changes, which arise from unknown factors, exhibiting random and unpredictable behavior. When expressing time series models, there are two primary approaches. Assuming y represents the time series and T, C, S, and I denote the trend, cyclic, seasonal, and irregular components, respectively, the models are defined as follows:

- Additive model; that combines all or a subset of the components by summing them together. It can be expressed as $y = T + C + S + I$

- Multiplicative model; that represents the interaction between the components by multiplying them. It can be represented as $y = T \times C \times S \times I$

These two models, additive and multiplicative, provide the foundation for various methods used by researchers to analyze and predict time series data. These methods include MA (Moving Average), AR (Autoregressive), ARMA (Autoregressive Moving Average), SARMA (Seasonal Autoregressive Moving Average), ARIMA (Autoregressive Integrated Moving Average), among others. The choice of method depends on the specific problem at hand and the characteristics of the data. In the research conducted by the authors in [23] , they focused on the analysis and prediction of mobile traffic using time series methods. This study delves into the application of time series techniques to gain insights into mobile traffic patterns and make predictions based on historical data. Furthermore, in [22] , an example is presented where three different methods, neural network, time series analysis, and wavelet transformation, are compared. This comparative analysis sheds light on the strengths and limitations of each method in the context of the specific problem being addressed. The authors in [24] proposed an ARIMA-based mobility prediction model for forecasting the future mobility speed of a node in mobile ad hoc networks (MANETs). In [25], the authors examined and compared the performance of the LSTM and ARIMA methods on real data.

D. Statistics, Probabilities and Curve Fitting

Statistical indicators are also used in traffic analysis and investigation. They provide valuable insights into the characteristics and patterns of the data. Various statistical measures are utilized, including central indices such as the mean, median, and mode, which provide information about the central tendency of the data. Dispersion indices such as the range, mean deviation, variance, and standard deviation capture the spread or variability of the data. Distribution indices such as skewness and kurtosis measure the shape and symmetry of the data distribution. These statistical indicators are particularly useful for long-term data analysis, providing a comprehensive understanding of the data alongside other analytical methods. Additionally, entropy, a measure of uncertainty or randomness, and conditional probabilities are employed in traffic analysis to uncover hidden patterns and relationships within the data.

Statistical distributions are widely utilized for traffic prediction, often represented through probability density functions or cumulative distribution functions. These distributions can take on various forms, including discrete or continuous, as well as univariate or multivariate. To leverage these distributions effectively, it is crucial to identify the most appropriate distribution that aligns with the characteristics of the data under investigation. Once a suitable distribution is determined, curve-fitting methods are employed to achieve the best fit between the distribution and the data. Curve fitting techniques enable researchers and analysts to estimate the parameters of the chosen distribution, ensuring that the distribution closely aligns with the observed data. This fitting process helps to capture the underlying patterns and behaviors within the data, thereby enabling accurate traffic prediction. Using statistical distributions, the authors [26] have determined the general pattern of traffic based on user behavior. In [27], by using curve fitting and some probability distributions, the long-term changes in users' data usage patterns were extracted, and [28] determined CDR patterns in numerous users. In [29], the authors introduced a different approach to designing predictors for mobile network environments. Their research leveraged recent advancements in time series prediction to propose a hybrid method that combined statistical modeling and machine learning through a joint training process. To meet the specific

demands of network traffic, they introduced a customized model called TES-RNN (Thresholded Exponential Smoothing and Recurrent Neural Network).

E. Probabilistic Graphical Models

A combination of modeling, graphs, and probabilities constitutes probabilistic graphical models (PGMs). These models incorporate data and extract algorithms while incorporating specialized information from experts. Many problems often involve uncertainties, such as limited knowledge in vast and complex domains, the presence of observation noise, unaccounted-for related factors, and inherent randomness in the system.

Probability theory serves as a powerful tool to effectively describe uncertain scenarios, with conditional probabilities being one of its major advantages. Probabilistic graphical models typically manifest in two forms: Markov networks and Bayesian networks. Markov networks are undirected and utilized to represent correlations, whereas Bayesian networks are directed and used to depict cause-and-effect relationships. The integration of probabilistic graphical models with probabilistic programming languages has facilitated the application of machine learning techniques alongside these models, resulting in more accurate and dynamic modeling capabilities.

Probabilistic graphs find applications in various areas related to network traffic analysis. For instance, utilizing these graphs to estimate the probability of user presence at specific times and locations, as well as their usage of particular services, can aid in detecting network anomalies. In [12], researchers employed Markov chains to optimize spectrum allocation in a network with heterogeneous traffic. By leveraging probabilistic and traffic models, they were able to enhance the quality of service (QoS). Similarly, in [30] , the authors introduced a method for predicting the traffic of mobile network nodes by utilizing user movement graphs over time and employing tools such as time series analysis. These approaches demonstrated the effectiveness of probabilistic graphs in enhancing network analysis and prediction capabilities.

At the end of this section, Table I is provided to present a summary and highlight some applications of the reviewed methods.

TABLE I
Some Applications of the Mentioned Methods in Network Traffic

No	Method	Some applications in mobile network traffic	Ref.
1	Machine Learning	Obtaining traffic patterns, traffic prediction, network anomaly detection, changes in user data usage patterns, obtaining the distribution of network traffic resource usage, user segmentation, etc.	[13,17–20,31,32]
2	Transform	Obtaining traffic patterns, traffic prediction.	[16,22]
3	Time Series	Obtaining traffic patterns, network anomaly detection.	[22–24]
4	Sta. and Curve fitting	Finding the pattern of users' data usages, distribution of network resource consumption, statistical analysis of different logs, detection of network anomalies, analysis of changes in traffic behavior patterns, etc.	[26–29]
5	Prob. graphs	Prediction of users' behavior, detection of network anomalies	[12,30,33]

III. PROPOSED METHOD

Based on our research, previous studies focusing on network traffic from the perspective of user behavior have predominantly examined data traffic with various protocols or CDR (Call Detail Record) information, often combined with spatiotemporal models [34,35]. However, we have observed a limited consideration or absence of a probabilistic perspective specifically based on user types, particularly human users. This viewpoint involves calculating traffic based on a dynamic and time-varying probabilistic model derived from the behavior of end-users.

The proposed model takes into account the dynamic nature of user behavior, enabling the prediction of overall network traffic based on users' past actions. In this model, each user exhibits individual behavior, and the combination of users contributes to determining the traffic of any given node. This approach follows a bottom-up or part-to-whole methodology. One of the key motivations for adopting this approach is the increased distribution and density of networks compared to the past, along with a higher number of users in specific areas and a greater diversity of network applications.

There are two main approaches to identifying the end-user or end node: either complete information regarding the user's type is available, or the user remains unknown. In the case of a mobile network, each user is assigned a known ID, allowing for identification by the network administrator. Thus, for the purpose of this discussion, we assume that the user's identity is known to the system administrator. Categorizing users from different perspectives can contribute to the development of a probabilistic model, taking into account the inherent uncertainties in user behavior. Fig. 1 presents our proposed categorization tree, which aims to organize users into distinct categories.

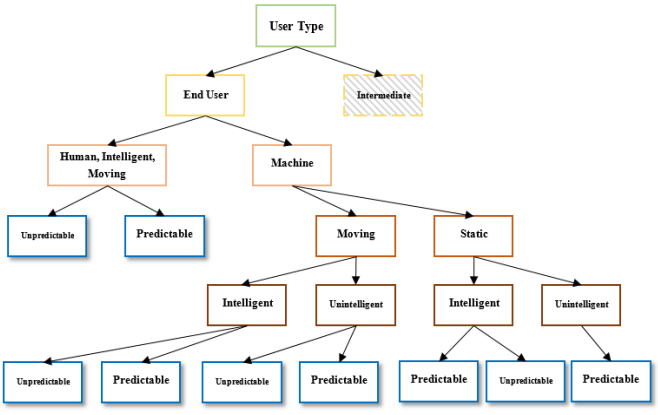

Fig. 1. Categorization of users in a network

These categories encompass two generic divisions: intermediate users and end-users. The end-user category branches into two subcategories: machine and human users. Machine users can be grouped into static or mobile users, each with the potential to exhibit intelligent or unintelligent behavior. The final categorization pertains to predictable (or certain) and unpredictable (or uncertain) behavior. Within the human user category, intelligence and mobility are combined into a single block with two states: predictable or unpredictable. In this paper, we have focused only on human users for two reasons. Firstly, several studies have already been conducted on machine traffic in the past. For instance, [14] explores research in the field of IoT traffic investigation. Secondly, according to Cisco's forecast for 2023, machine users are projected to account for approximately one-third of the total, while the remaining traffic is generated by human users in mobile networks [36].

User behavior can be a function of various factors. Based on this categorization and considering the human user, we take into account three factors here. Equation (1) expresses the general formula of this relationship.

$$B = f(U,T,L) \tag{1}$$

According to (1), the behavior of B is a function of the type of user U at time T and at location L. We will use (2) to predict the traffic.

$$Tr_v = P_{user\_id} \times V_{data} \times \Delta T \tag{2}$$

In this equation, Tr_v represents the volume of traffic requested by the user, P_{user_id} denotes the probability of network usage by a user with the characteristic *user_id*, V_{data} represents the required data volume, and ΔT represents the time required to transfer data over the network.

The probability of each user's data usage (P_{user_id}) is obtained from equation (3). Specifically, it represents the probability that a user with the *user_id* requests network usage (*data_request*), which is dependent on both the user's location and the specific time.

$$P_{user\_id} = P(data\_request \mid user\_id\ ,\ T\ ,\ L) \tag{3}$$

To provide a more detailed description of this equation, let's consider a scenario where a human user visits multiple locations at different times over consecutive days. By analyzing the user's historical records, we can calculate the probability of the user being present at a specific time in a given location. For a fixed user who remains in a constant location, the probability of network data usage depends solely on time. To facilitate this analysis, we propose Table II, which presents these probabilities for each user identified by their *user_id*. In this table, the 24-hour day is partitioned into intervals denoted as T_i. The time span between these intervals can be either constant or variable, with no overlap. The number of intervals, denoted as N (a natural number), is chosen based on the specific application requirements. The selection of N and whether the interval distances are fixed or variable significantly impact the accuracy of the prediction. It is important to note that excessively reducing or increasing the interval distances can influence the prediction results. Furthermore, to account for different possible locations where the user may be present, denoted as L_j, we employ categorization. The total number of locations, represented by the natural number M, varies depending on the user type and characteristics.

In Table II, $n_{T_i.L_1}$ is a natural number and represents the number of times the user with *user_id* ID was present at time T_i and at location L_j. Having this table for each user enables us to determine the probability of their presence at any location and time.

TABLE II
Probability Matrix of User Presence at Location L And Time T

	T_1	T_2	...	T_N	Total
L_1	$n_{T_1.L_1}$	$n_{T_2.L_1}$	…	$n_{T_N.L_1}$	$\sum_{i \in N} n_{T_i.L_1}$
…	…	…	…	…	…
L_M	$n_{T_1.L_M}$	$n_{T_2.L_M}$	…	$n_{T_N.L_M}$	$\sum_{i \in N} n_{T_i.L_M}$
Total	$\sum_{j \in M} n_{T_1.L_j}$	$\sum_{j \in M} n_{T_2.L_j}$	…	$\sum_{j \in M} n_{T_N.L_j}$	…

In (2), the kind of data, time, and user location are some of the elements that affect how much data the user needs (Vdata). We take into account the location, the matching time, and the particular data that the user is anticipated to request when calculating the amount of data. This computation is done using equation (4), and the resultant value can be stated in bit/s or in terms of descriptive adjectives (such as low, medium, high, etc.). A numerical representation of the value can be created if it is given a detailed description. As an illustration, let's look at the following case (4): a human user who uses social networking software to make a video chat while on their lunch break at work while connected to the network. In this case, we know the precise time (noon), the place (workplace), and the kind of data (live video call). With this data, we can use the equation to determine the traffic volume.

$$V_{data} = g(T,L,data\_type) \tag{4}$$

The last parameter in (2) is the duration of the user's network connection (ΔT), which is also influenced by time and location. Similar to what was discussed regarding the volume of data, the specific time and location when a user is connected to the network and the type of data being transferred can be expressed using Equation (5). The value of h, similar to g, is determined based on the user type.

$$\Delta T = h(T,L,data\_type) \tag{5}$$

The user information collected enables the creation of a probabilistic model that captures the user's behavior over time. As more data is accumulated, a more accurate probabilistic model can be obtained. For instance, past information allows us to determine the probability of a human user utilizing a social network at work around noon. Therefore, when a user initiates a new request, the required volume can be calculated by referring to this information. Consequently, the corresponding table is updated.

Our proposed method is based on analyzing the similarity of behavior between new users and previous users. We assume that we have access to the historical data of a number of users. Using clustering methods, we group (or categorize) these users

based on their similar traffic behavior patterns. We consider each cluster as a separate group and assign a label to it. The notion of similarity here refers to the categorization of users based on their traffic behavior. It is obvious that clustering methods employ different approaches to group data, with each method categorizing similar data in its own unique manner. Subsequently, we calculate the average traffic volume for each group. For instance, if we consider hourly traffic, we would have 24 data points for each 24h-day. The traffic volume assigned to each group for each hour is determined by the average traffic of all users within that group. For example, if we have 8 clusters, it means we have 8 distinct groups, each exhibiting similar behavior amongst its members. We compute the average traffic for each hour of the 24-hour day within these groupings. This data is then utilized to predict the traffic patterns of new users. In other words, whenever a new user is encountered, we utilize distance functions, such as the Euclidean distance, to determine which group the user belongs to. Consequently, we assign the average value of that group to the new user.

To further elaborate on this method, let's consider a scenario where the aforementioned information is available for a certain number of users (N_u users, as historical data). As mentioned, these users are categorized into different groups using clustering methods, and each group is assigned a label. To make it simple, let's assume that in (2), P_{user_id} is equal to 1 and the time interval (ΔT) remains constant, and we have already calculated the average traffic for each group within this time interval. Now, to predict the traffic in a network node consisting of several new users (N_p), we first use distance functions to determine the related group of each user. We determine the traffic of a new user by considering the average traffic of the group that has the least distance to the user. For example, let's consider a scenario where three clusters named g1, g2, and g3 have been generated through clustering. The average traffic for each group has been calculated. Consider that the average traffic for the time period of 8 AM is 0.2, 0.8, and 0.5 for g1, g2, and g3, respectively. Now, suppose we have 100 new users, with 35, 25, and 40 users assigned to groups g1, g2, and g3, respectively. To calculate the traffic of the new users at 8 AM, we follow the subsequent steps: First, we compute the product of 0.2 multiplied by 35, 0.8 multiplied by 25, and 0.5 multiplied by 40. Next, we sum these products, and finally, we divide the obtained value by the total number of users (here is 100). In this example, the calculated traffic for 8 AM is 0.47. This approach enables us to estimate traffic patterns more accurately. Further details and equations related to this method will be provided in the subsequent section.

IV. SIMULATIONS AND RESULTS

To simulate the proposed method, we utilized traffic data synthesis. By analyzing the traffic patterns of over 9,600 telecommunications cells, encompassing more than 150,000 users in a city in China, the authors [21] extracted hourly, daily, and weekly traffic models related to different user data usages. For our simulations, we employed the synthesized hourly and daily traffic patterns from their research. Additionally, the following assumptions were made:

- In (3), we set the value of P_{user_id} to 1, indicating that there is a probability of a user connection at any time of the day and at any location.

- In (4), we assigned the following values for different data types: video, search, voice call, and text message with 4, 1, 0.1, and 0.01 Mbps, respectively.

- In (5), we considered ΔT as a constant value.

- For simplicity, we did not consider the effect of location (L) in any of these equations.

To evaluate the efficiency of the proposed method and compare it with other methods, we employed the Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE) criteria as described in Equations (6), (7), and (8), respectively. In these equations, N_s represents the number of prediction samples, t_r is the real traffic, and $\widetilde{tr}$ denotes the predicted traffic. MAPE is expressed as a percentage.

$$RMSE: \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s}(tr(i)-\widetilde{tr}(i))^2} \quad (6)$$

$$MAPE: 100 \times \frac{1}{N_s}\sum_{i=1}^{N_s}\left|\frac{tr(i)-\widetilde{tr}(i)}{tr(i)}\right| \quad (7)$$

$$MAE: \frac{1}{N_s}\sum_{i=1}^{N_s}|tr(i)-\widetilde{tr}(i)| \quad (8)$$

According to the above assumptions and utilizing random functions, we have simulated the traffic patterns of 1000 users in a manner that ensures the overall traffic pattern aligns with the hourly and daily traffic patterns in [21]. Fig. 2-top shows the synthesis of normalized hourly traffic for 1000 users, and Fig. 2-bottom is an example of normalized hourly traffic for one user. The value of ΔT is set to 5 minutes.

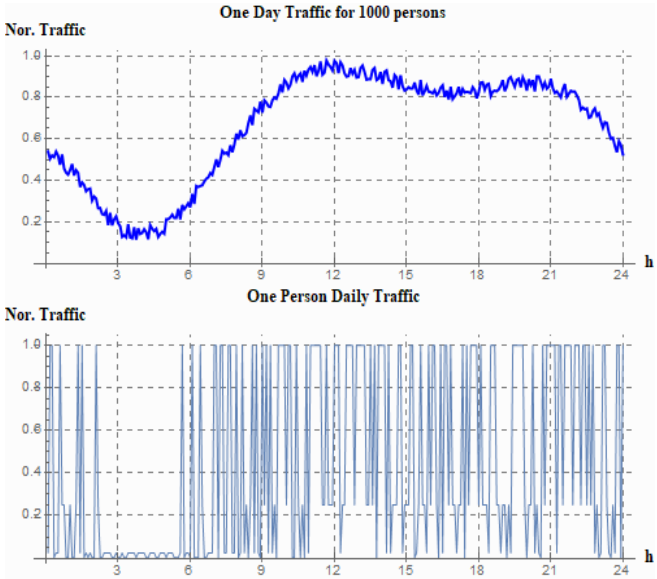

Fig. 2. Top: Daily traffic for 1000 users (synthesized and normalized), Bottom: One user daily normalized data usage.

In the first simulation, we set the value of ΔT to 5 minutes and considered 1000 users to predict the hourly traffic pattern. This means that for every 24 hours, we have 288 samples per user. Also, the traffic of 800 users (N_u) was utilized for clustering and training a machine learning model. The normalized synthesized traffic data of 200 users (N_p) was included for prediction. In other words, using various methods, we aimed to predict the traffic patterns for these new users.

In this scenario, we considered a total of 576 samples (equivalent to 2 days). Also, 514 samples were used as historical data, while 62 samples were reserved for prediction (N_s). The traffic behavior of the 800 users had already been categorized and labeled using three clustering methods: k-means, k-medoids, and spectral, with values of k set to 8 and 16. For every 5-minute interval, the average traffic of each cluster, denoted as $m_{c_i}(t)$, was calculated. Here, c_i represents a natural number ranging from 1 to k, and t corresponds to the time intervals starting from 1 and incrementing by ΔT. The value of k was determined through experimental extraction [37]. The method for selecting k is implemented by running the proposed model for different values of k (using iterative loops) and choosing the value that yields the lowest prediction error. Additionally, in machine learning, there exist various methods for clustering, and depending on the nature of the problem, one or several of these methods can be utilized. For selecting the appropriate clustering method, we have experimented with several methods and selected those that exhibited lower prediction errors.

To predict the 62 traffic samples for the 200 new users, we employed a weighted average approach. Firstly, each user was assigned to a specific cluster (or category), and then the number of users belonging to each cluster among the 200 new users, denoted as n_{c_i}, was determined. Finally, the weighted average for each t was calculated using Equation (9).

$$\tilde{tr}(t) = \frac{1}{N_p}\sum_{i=1}^{k} m_{c_i}(t) \times n_{c_i} \tag{9}$$

In the following, we compare the proposed method with Fourier and time series methods using three criteria: RMSE, MAPE, and MAE. The time series methods employed are ARIMA (2,2,2) and SARIMA (2,1,2) $(0,1,0)_1$. For simulations and plot generation, we utilized Mathematica software.

Table III presents the comparison results among the different methods. According to the table, the spectral method with k=8 outperformed the other methods in all three error criteria. Next, the same method with k=16 exhibited promising performance, followed by the Fourier method. Fig. 3-top illustrates the comparison between the simulation results of the proposed method, the time series methods, and the Fourier transform method against the ground truth data. Additionally, Fig. 3-bottom provides a more detailed view of this comparison.

TABLE III
Comparison of the Proposed Method with Existing Methods for Hourly Prediction (ΔT = 5 Minutes)

Pred. Err. Method	Machine Learning						Time Series		Fourier
	k=8			k=16					
	k-Means	k-Medoids	Spectral	k-Means	k-Medoids	Spectral	ARIMA	SARIMA	
RMSE	0.0652	0.0355	0.0059	0.0748	0.0612	0.0103	0.0449	0.0448	0.0171
MAPE	8.852	4.841	0.784	10.61	8.881	1.502	5.427	5.726	1.911
MAE	0.0515	0.0282	0.0046	0.0604	0.0512	0.0088	0.0385	0.0385	0.0137

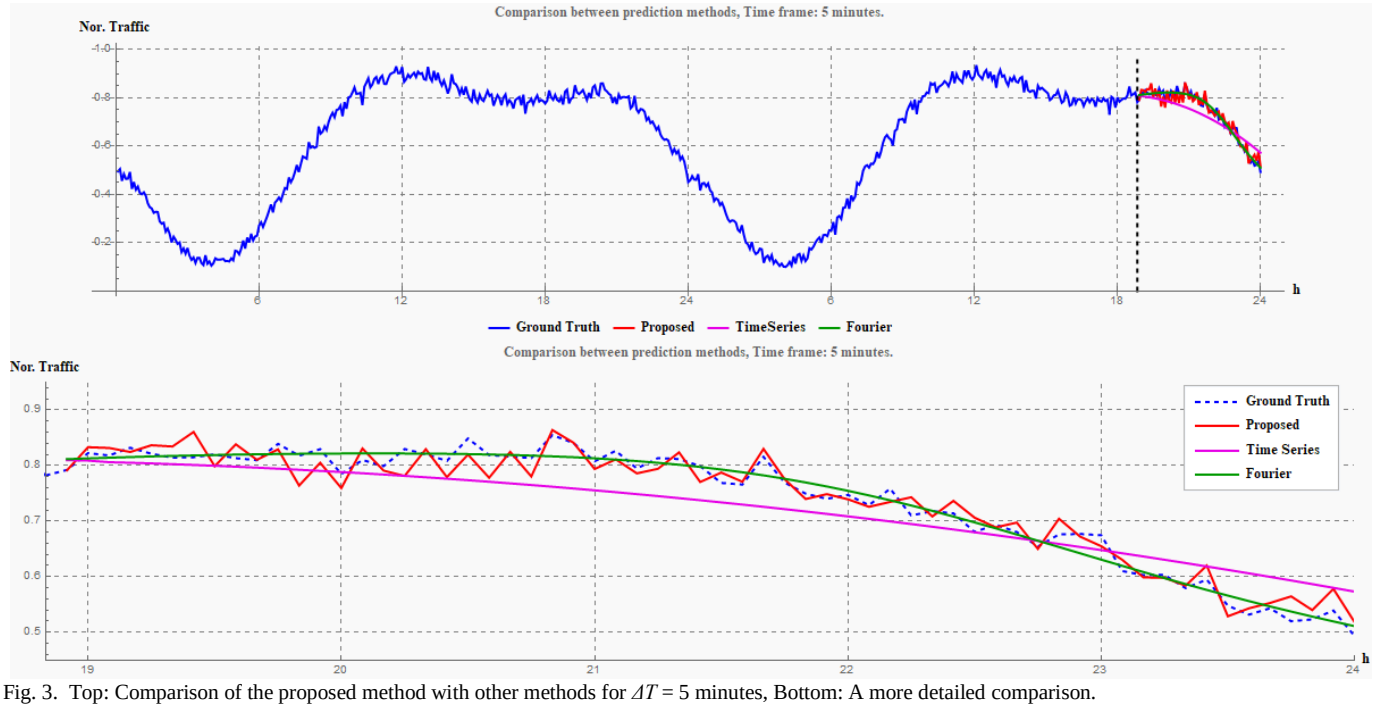

Fig. 3. Top: Comparison of the proposed method with other methods for ΔT = 5 minutes, Bottom: A more detailed comparison.

In the second simulation, we aimed to predict the daily traffic patterns. For this purpose, we set the value of ΔT to one hour (60 minutes) and considered a total of 6 days. This means that each user had a total of 144 samples, with 16 samples allocated for prediction and 128 samples for historical data.

In this simulation, we employed the ARMA (5,1) and SARIMA (1,0,1) $(0,1,2)_{24}$ time series methods, while keeping the rest of the description consistent with the previous

simulation. Fig. 4-top presents a comparison between the simulation results of the proposed method, the time series methods, the Fourier transform method, and the ground truth data. Additionally, Fig. 4-bottom provides a more detailed view of this comparison. Table IV displays the comparison results among the different methods. According to the table, the spectral method with k=8 demonstrated superior performance in all three error criteria compared to the other methods. Next, the Fourier and SARIMA methods ranked second and third, respectively.

TABLE IV
Comparison of the Proposed Method with Existing Methods for Hourly Prediction (ΔT = 60 Minutes)

Pred. Err. Method	Machine Learning						Time Series		Fourier
	k=8			k=16					
	k-Means	k-Medoids	Spectral	k-Means	k-Medoids	Spectral	ARMA	SARIMA	
RMSE	0.0245	0.0232	0.0115	0.0228	0.0737	0.0155	0.1732	0.0181	0.0141
MAPE	3.099	3.040	1.412	2.701	9.578	2.196	16.434	2.064	1.447
MAE	0.0194	0.0193	0.0091	0.0167	0.0605	0.0136	0.1277	0.0155	0.0110

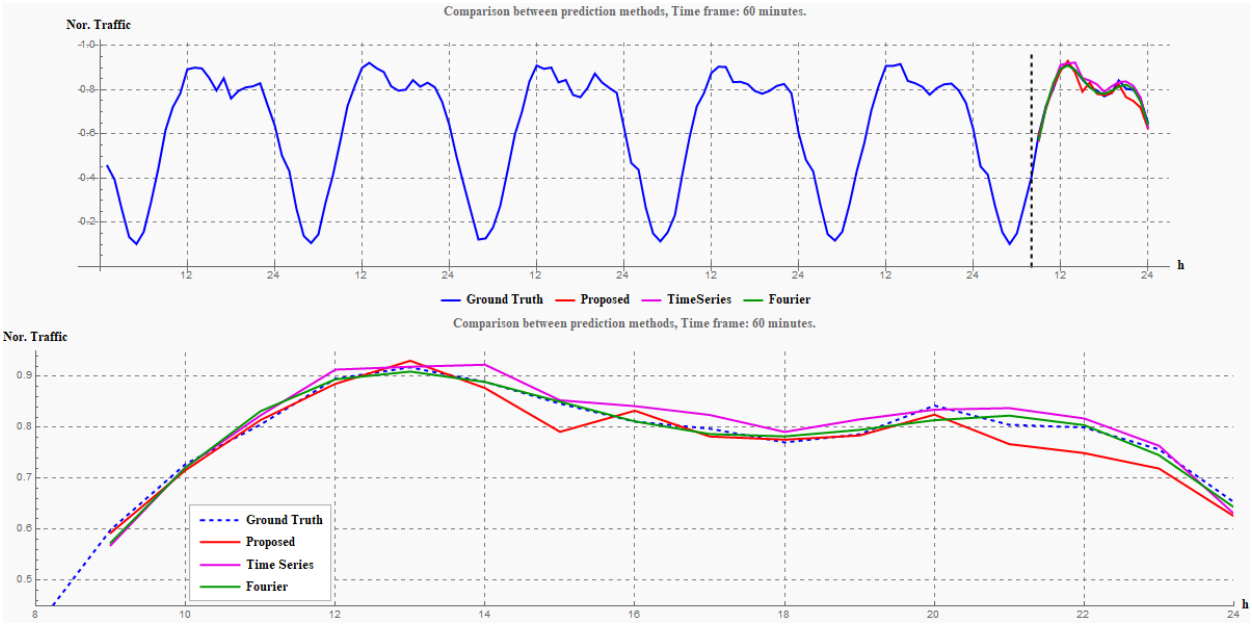

Fig. 4. Top: Comparison of the proposed method with other methods for ΔT = 60 minutes, Bottom: A more detailed comparison.

With ΔT set to 30 minutes, the final simulation focuses on the daily time period. We took into account a period of 6 days in this case, yielding 288 samples per user. Of these, 32 samples were set aside for prediction, while the remaining 256 samples were utilized for historical data.

For this simulation, we utilized the ARMA (4,1) and SARIMA (1,1,1) $(1,1,2)_{48}$ time series methods. Fig. 5-top illustrates a comparison between the simulation outcomes of the proposed method, other methods, and the original data. Additionally, Fig. 5-bottom provides a detailed view of this comparison. Table V presents the comparison results among the different methods. According to the table, the spectral method with k=8 exhibited superior performance in all three error criteria compared to the other methods. The Fourier methods and spectral with k=16 ranked second and third, respectively. By comparing Table IV with Table V, we observe that the results of the 30-minute time frame are slightly better than the 60-minute time frame. It is important to note that we maintained the same random seed for both simulations.

TABLE V
Comparison of the Proposed Method with Existing Methods for Hourly Prediction (ΔT = 30 Minutes)

Pred. Err. Method	Machine Learning						Time Series		Fourier
	k=8			k=16					
	k-Means	k-Medoids	Spectral	k-Means	k-Medoids	Spectral	ARMA	SARIMA	
RMSE	0.0231	0.0262	0.0088	0.0231	0.0488	0.0134	0.2204	0.0262	0.0186
MAPE	3.107	3.213	1.1458	3.003	6.516	1.794	21.046	2.439	1.886
MAE	0.0193	0.0206	0.0073	0.0187	0.0414	0.0113	0.165	0.0184	0.0149

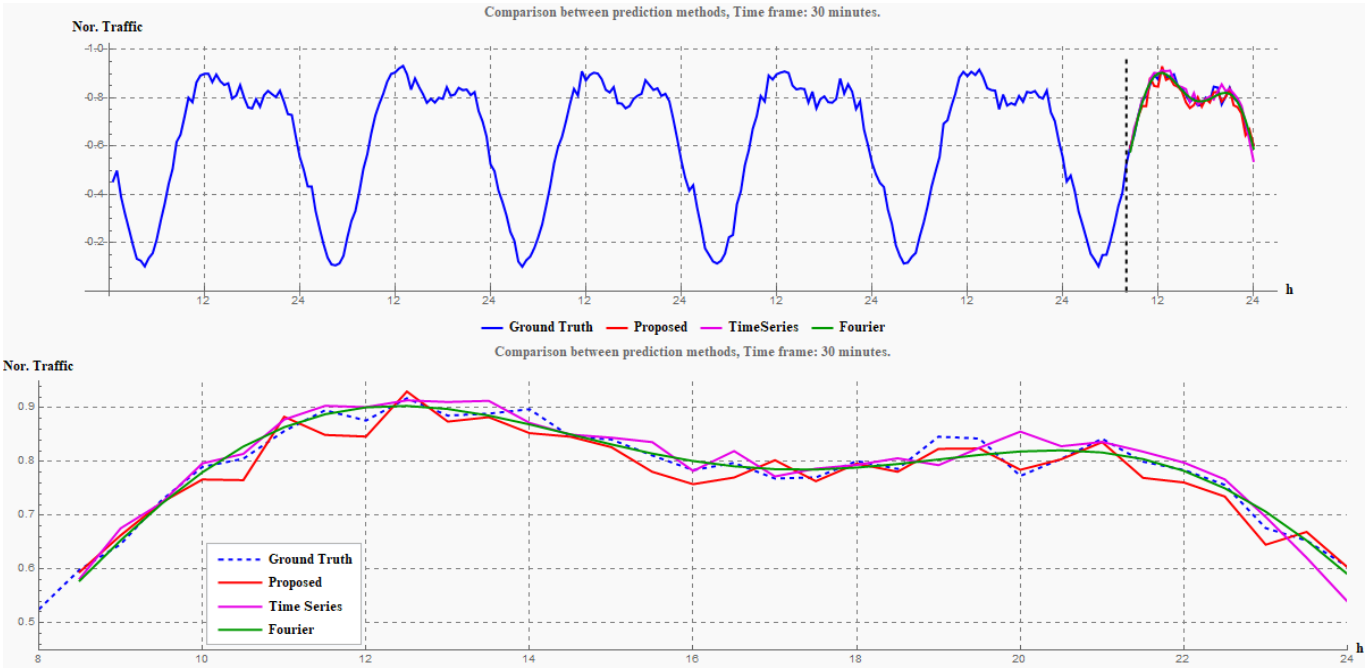

Fig. 5. Top: Comparison of the proposed method with other methods for ΔT = 30 minutes, Bottom: A more detailed comparison.

V. CONCLUSION

In this study, we offer a machine learning approach that uses user behavior probabilities to estimate mobile network traffic. Using clustering techniques, this strategy groups users with similar behavior probabilities and makes predictions based on the average traffic in each cluster.

To demonstrate the effectiveness of the proposed method, we utilized synthetic traffic data. For this purpose, we generated synthetic traffic data for 1000 users using the data [21] at three different time frames: 5 minutes, 30 minutes, and 60 minutes. We labeled the 5-minute time frame as hourly and generated 576 samples (equivalent to 2 days), where 62 samples were used for prediction and the rest for historical data. The 30-minute and 60-minute time frames were labeled as daily, and we generated 288 and 144 samples, respectively, which corresponds to 6 days. In the daily time frame, we considered 32 and 16 samples, respectively, for prediction. In the proposed machine learning-based method, we considered 800 users for training and 200 users for testing. Based on our simulations, the proposed method outperformed the Fourier and time series methods in all three metrics (RMSE, MAPE, and MAE). Also, Figs. 3 to 5 and Tables III to V have been provided to compare our proposed method with other existing methods.

To continue our work, we are interested in making further developments. In this paper, we have employed some simplifications. For instance, in (2), we considered the value of P_{user_id} to be 1, and in equations (3) to (5), we ignored the influence of location (L). Additionally, synthetic traffic data was utilized. In future studies, we will examine our proposed method using real data without the mentioned simplifications. Furthermore, while this paper only compares our proposed method with two other approaches (the time series method and Fourier), we will incorporate additional methods in the future. In this paper, we considered a user count of 1000, but in the future, we will explore different values (both fewer and greater than 1000). Moreover, we will modify the prediction time frame and utilize other time frames as well. We hope that our proposed method proves to be effective for these advancements.

VI. REFERENCES

[1] Navarro-Ortiz J, Romero-Diaz P, Sendra S, Ameigeiras P, Ramos-Munoz JJ, Lopez-Soler JM (2020) A Survey on 5G Usage Scenarios and Traffic Models. *IEEE Commun Surv Tutorials* 22, pp. 905–929.

[2] Walelgne EA, Asrese AS, Manner J, Bajpai V, Ott J (2021) Clustering and predicting the data usage patterns of geographically diverse mobile users. *Comput Networks* 187, pp. 107737.

[3] Qiao Y, Xing Z, Fadlullah ZM, Yang J, Kato N (2018) Characterizing Flow, Application, and User Behavior in Mobile Networks: A Framework for Mobile Big Data. *IEEE Wirel Commun* 25, pp. 40–49.

[4] Wang W, Harari GM, Wang R, Müller SR, Mirjafari S, Masaba K, Campbell AT (2018) Sensing Behavioral Change over Time. *Proc ACM Interactive, Mobile, Wearable Ubiquitous Technol* 2, pp. 1–21.

[5] Mokhtari A, Sadighi L, Bahrak B, Eshghie M (2020) Hybrid Model for Anomaly Detection on Call Detail Records by Time Series Forecasting. *arXiv* pp. 1–12.

[6] Truong Dinh K, Kukliński S, Osiński T, Wytrębowicz J (2020) Heuristic traffic engineering for SDN. *J Inf Telecommun* 4, pp. 251–266.

[7] Zhang J, Ye M, Guo Z, Yen CY, Chao HJ (2020) CFR-RL: Traffic Engineering with Reinforcement Learning in SDN. *IEEE J Sel Areas Commun* 38, pp. 2249–2259.

[8] Shinkuma R, Tanaka Y, Yamada Y, Takahashi E, Onishi T (2018) User instruction mechanism for temporal traffic smoothing in mobile networks. *Comput Networks* 137, pp. 17–26.

[9] Shin H, Jung J, Koo Y (2020) Forecasting the video data traffic of 5 G services in south korea. *Technol Forecast Soc Change* 153, pp. 119948.

[10] Passas V, Miliotis V, Makris N, Korakis T (2020) Pricing Based Distributed Traffic Allocation for 5G Heterogeneous Networks. *IEEE Trans Veh Technol* 69, pp. 12111–12123.

[11] Peng J (2020) Impact of the Arrival Distribution of Primary User Traffic in Dynamic Spectrum Access. *Procedia Comput Sci* 170, pp. 325–332.

[12] Shruti, Kulshrestha R (2020) Channel allocation and ultra-reliable communication in CRNs with heterogeneous traffic and retrials: A

dependability theory-based analysis. *Comput Commun* 158, pp. 51–63.
[13] Jiang W (2022) Cellular traffic prediction with machine learning: A survey. Expert Syst Appl 201, pp. 117163.
[14] Khedkar SP, Canessane RA, Najafi ML (2021) Prediction of Traffic Generated by IoT Devices Using Statistical Learning Time Series Algorithms. Wirel Commun Mob Comput 2021, pp. 1–12.
[15] Nie L, Ning Z, Obaidat MS, Sadoun B, Wang H, Li S, Guo L, Wang G (2021) A Reinforcement Learning-Based Network Traffic Prediction Mechanism in Intelligent Internet of Things. IEEE Trans Ind Informatics 17, pp. 2169–2180.
[16] Long P, Li J, Liu N, Pan Z, You X (2022) Antenna On/Off Strategy for Massive MIMO Based on User Behavior Prediction. In 2022 3rd International Conference on Computing, Networks and Internet of Things (CNIOT) IEEE, pp. pp. pp. 113–119.
[17] Alqudah N, Yaseen Q (2020) Machine Learning for Traffic Analysis: A Review. In Procedia Computer Science Elsevier B.V., pp. pp. pp. 911–916.
[18] Morocho-Cayamcela ME, Lee H, Lim W (2019) Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions. IEEE Access 7, pp. 137184–137206.
[19] Alekseeva D, Stepanov N, Veprev A, Sharapova A, Lohan ES, Ometov A (2021) Comparison of Machine Learning Techniques Applied to Traffic Prediction of Real Wireless Network. IEEE Access 9, pp. 159495–159514.
[20] Yang Y, Geng S, Zhang B, Zhang J, Wang Z, Zhang Y, Doermann D (2023) Long term 5G network traffic forecasting via modeling non-stationarity with deep learning. Commun Eng 2, pp. 33.
[21] Xu F, Li Y, Wang H, Zhang P, Jin D (2017) Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment. IEEE/ACM Trans Netw 25, pp. 1147–1161.
[22] Iqbal MF, Zahid M, Habib D, John LK (2019) Efficient Prediction of Network Traffic for Real-Time Applications. J Comput Networks Commun 2019, pp. 1–11.
[23] Xu F, Lin Y, Huang J, Wu D, Shi H, Song J, Li Y (2016) Big Data Driven Mobile Traffic Understanding and Forecasting: A Time Series Approach. IEEE Trans Serv Comput 9, pp. 796–805.
[24] Theerthagiri P (2022) Mobility prediction for random walk mobility model using ARIMA in mobile ad hoc networks. J Supercomput 78, pp. 16453–16484.
[25] Albeladi K, Zafar B, Mueen A (2023) Time Series Forecasting using LSTM and ARIMA. Int J Adv Comput Sci Appl 14, pp. 2023.
[26] Walelgne EA, Asrese AS, Manner J, Bajpai V, Ott J (2020) Understanding Data Usage Patterns of Geographically Diverse Mobile Users. IEEE Trans Netw Serv Manag pp. 1–1.
[27] Camacho J, McDonald C, Peterson R, Zhou X, Kotz D (2020) Longitudinal analysis of a campus Wi-Fi network. Comput Networks 170, pp. 107103.
[28] Jiang D, Wang Y, Lv Z, Qi S, Singh S (2020) Big Data Analysis Based Network Behavior Insight of Cellular Networks for Industry 4.0 Applications. IEEE Trans Ind Informatics 16, pp. 1310–1320.
[29] Lo Schiavo L, Fiore M, Gramaglia M, Banchs A, Costa-Perez X (2022) Forecasting for Network Management with Joint Statistical Modelling and Machine Learning. In 2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM) IEEE, pp. pp. pp. 60–69.
[30] Sun F, Wang P, Zhao J, Xu N, Zeng J, Tao J, Song K, Deng C, Lui JCS, Guan X (2021) Mobile Data Traffic Prediction by Exploiting Time-Evolving User Mobility Patterns. IEEE Trans Mob Comput 14, pp. 1–1.
[31] Ma H, Yang K, Pun M-O (2023) Cellular traffic prediction via deep state space models with attention mechanism. Comput Commun 197, pp. 276–283.
[32] Fu Y, Wang X (2022) Traffic Prediction-Enabled Energy-Efficient Dynamic Computing Resource Allocation in CRAN Based on Deep Learning. IEEE Open J Commun Soc 3, pp. 159–175.
[33] Zhao N, Wu A, Pei Y, Liang Y-C, Niyato D (2022) Spatial-Temporal Aggregation Graph Convolution Network for Efficient Mobile Cellular Traffic Prediction. IEEE Commun Lett 26, pp. 587–591.
[34] Liu M, Liu G, Sun L (2023) Spatial–temporal dependence and similarity aware traffic flow forecasting. Inf Sci (Ny) 625, pp. 81–96.
[35] Lee J-M, Kim J-D (2022) A Generative Model for Traffic Demand with Heterogeneous and Spatiotemporal Characteristics in Massive Wi-Fi Systems. Electronics 11, pp. 1848.
[36] Cisco (2020) Cisco Annual Internet Report - Cisco Annual Internet Report (2018 - 2023) White Paper. pp. 1–41.
[37] Pimpinella A, Giusto F Di, Redondi AEC, Venturini L, Pavon A (2022) Forecasting Busy-Hour Downlink Traffic in Cellular Networks. In ICC 2022 - IEEE International Conference on Communications IEEE, pp. 4336–4341.