IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 January 2023, accepted 29 January 2023, date of publication 10 February 2023, date of current version 16 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242861

RESEARCH ARTICLE

Deep Learning Based Fusion Model for Multivariate LTE Traffic Forecasting and Optimized Radio Parameter Estimation

SYED TAUHIDUN NABI[1], (Graduate Student Member, IEEE), MD. RASHIDUL ISLAM[1], MD. GOLAM RABIUL ALAM[1], (Member, IEEE), MOHAMMAD MEHEDI HASSAN[2], (Senior Member, IEEE), SALMAN A. ALQAHTANI[3], (Member, IEEE), GIANLUCA ALOI[4], (Member, IEEE), AND GIANCARLO FORTINO[4], (Fellow, IEEE)

[1]Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh
[2]Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[3]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4]Department of Informatics, Modeling, Electronics, and Systems, University of Calabria, 87036 Rende, Italy

Corresponding author: Giancarlo Fortino (giancarlo.fortino@unical.it)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project under Project RSP2023R18. This work was also carried out in the framework of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 9 - Green-aware AI, under the NRRP MUR program funded by the NextGenerationEU.

ABSTRACT With the evaluation of cellular network internet data traffic, forecasting and understanding traffic patterns become the critical objectives for managing the network-designed Quality of Service (QoS) benchmark. For this purpose, cellular network planners often use different methodologies for predicting data traffic. However, traditional traffic forecasting approaches are erroneous. As well as most of the time, traditional traffic forecasts are high-level or a generously large regional cluster level. Also, eNodeB-level utilization with concerning traffic forecasting is not readily available. As a result, user experience degradation or unnecessary network expansion is triggered based on the traditional method. This research deals with extensive 6.2 million real network time series Long-Term Evolution (LTE) data traffic and other associate parameters, including eNodeB-wise Physical Resource Block (PRB) utilization, which focuses on building a traffic forecasting model with the help of multivariate feature inputs and deep learning algorithms. A state-of-the-art Deep Learning algorithm-based fusion model is proposed. The combination of different deep learning algorithms, namely Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM) and Gated Recurrent Unit (GRU), enables traffic forecasting at a granular eNodeB-level and also provides eNodeB-wise forecasted PRB utilization. In this research R^2 score value for the proposed fusion model is 0.8034, which outperforms traditional models. Apart from the PRB utilization, QoS threshold was devised as 70% from a real network experience to trigger soft parameter tuning decisions. Based on the forecasted PRB utilization, this research proposed a unique algorithm that estimates eNodeB-level soft capacity parameter optimization for a short-term step-up solution or long-term network expansion to ensure a guaranteed QoS benchmark.

INDEX TERMS LTE networks, machine learning in networking, traffic prediction, deep learning, mobile network capacity, physical resource block, resource management.

I. INTRODUCTION

Mobile Internet data traffic has been rapidly growing in the last decades. According to Ericsson Mobility Report (Nov 2021), by 2027, global mobile network data traffic will be close to 300 exabytes per month [1]. With the evaluation of technology, at present and next 2-3 years, LTE will carry most of the data traffic where 5G and beyond technology is still under development, mostly in developing countries. Apart from that, the LTE network is more mature for carrying a

The associate editor coordinating the review of this manuscript and approving it for publication was Hassan Tariq Chattha.

load of data traffic as it is already covered by 84% of the global population as of 2020 [2]. With the growth of mobile data traffic, several challenges are coming to the surface. Among them, eNodeB-wise utilization prediction is topmost because the network quality of services (i.e., Speed, Latency) is depends on this utilization parameter.

The number of users and their demand for mobile internet speed with quality is ever-increasing. In Ericsson Mobility Report 2021 [1], the monthly average internet usage per smartphone is approx. 11.4 GB, which will be almost four times by 2027. As well as, location-wise variety of user behavior gives some extra load to the LTE network or cell. For example, video traffic is currently getting more popular than any other chatty or browsing traffic, which is nearly 70% [1]. It has been predicted that video traffic will reach about 79% by 2027, which means more data transfer by mobile networks. For this reason, it can be easily predicted that, without proper forecasting of network utilization beforehand, Mobile Network Operators (MNO) will not handle the user demand on time which may cause QoS degradation.

A. MOTIVATION

Traffic forecasting is the most sophisticated part of network dimensioning [3]. Because to make cellular network business more profitable, investors are always looking for the proper Capital Expenditure (CAPEX) in the right cell/site/location and reducing Operational Expenditure (OPEX). Wrong traffic forecasting may mislead the network dimensioning, which causes additional CAPEX and OPEX as well as degradation of QoS.

Apart from that, Deep learning-based approaches have been studied recently to identify the pattern of sequential data and classify similar data types together [4]. Different Recurrent Neural Networks (RNN) algorithms are used to forecast multiple time series sequential data types. By knowing the enormous potential of deep learning algorithms for predictive measures, the authors were more interested in building a model to solve one of the most critical problems in cellular network dimensioning, which is traffic forecasting [5]. In this work, Modern GPUs are used to run complex deep learning algorithms with various features in optimistic run time.

The advance is knowing the more accurate traffic and user demand from the network's ability to promptly manage the resource allocation among the connected User, which will improve the quality of user experience [5]. This work will help to understand the Mobile network's traffic behavior. Also, recommend which eNodeB to trigger expansion with the help of deep learning algorithm-based traffic forecast and Utilization correlation chart.

B. PROBLEM FORMULATION

Understanding the traffic demands in a cellular network is a complex task due to the large and uneven densification of the mobile users attached to a particular network. Apart from that, this task becomes more challenging because of the huge number of different types of devices, and user patterns are different all the time. Different application data traffic consumption rate is not the same [6].

From an academic research perspective, one of the significant challenges is collecting a large set of data for training a model. eNodeB-wise detail datasets with several valuable features are not available by MNOs. In most cases, Call Detail Records (CDR) contain the traffic dataset in an aggregated format, where there is no segregation of technology, user count per technology, or per-protocol category [7]. So, the CDR dataset could not help this research work that much. Several efforts and initiatives had taken for data mining to collect the suitable dataset from Operation Support System (OSS) and radio and core network end.

Network traffic forecasting is one of the significant activities for network dimensioning because it affects eNodeB-level utilization. In other words, eNodeB-level traffic and utilization are directly proportional. As well as, eNodeB-level utilization affects overall network performance and user experience. Suppose utilization increased uncontrolled manner, which means physical resource block (PRB) shared by additional users. It will hamper user experience as the same resource is shared by more users initially not considered for design. We need to trigger network expansion before the capacity reaches the threshold point for those cases. If traffic is not increasing as per the prediction level, eNodeB will be underutilized, which means wasting resources. The greatest challenge is that both case network engineers traditionally can observe the scenario only after it happens in particular eNodeB in a real network. Then, engineers take the initiative for resolution with a step-up capacity solution or step-down in some cases. However, in the meantime (lead time), the customer suffers from QoS degradation. It would be much better for network planners/engineers to identify eNodeB-level traffic patterns and PRB utilization and estimate radio parameters before it happens in the network. So that network engineers can initiate actions on time and customers will be less suffered.

Consequently, we have devised a Non-deterministic Polynomial *(NP)* hard problem based above-sated situation to address the solution. *NP*-hard problems are commonly used in formalized research problems [8], [9]. This research question can be classified as an optimization problem as our objective is to find out the maximum user throughput (T_h)in a particular time for each eNodeB and which is inversely proportional with PRB Utilization ($PRBU_t$), and other network contains.

$$\text{Objective function, Max } T_h = \frac{1}{PRBU_t} + C_1 \qquad (1)$$

Here in the objective function (1), the value of constant C_1 will change according to the configured radio bandwidth of each eNodeB. In other words, a user throughput of a particular eNodeB can vary based on configured bandwidth even in the same PRB utilization.

Similarly, future PRB utilization can be computed based on predicted traffic volume on that node and other factors. Suppose we want to calculate a cluster of eNobeB (number of eNodeBs in same geographic are creates cluster) future performance or PRB Utilization ($PRBU$). In that case, this can be possible with predicted traffic volume (Vol), and other factors i.e., Average user equipment ($\overline{UE}$), Maximum user equipment $Max\,(UE)$, Downtime (D_T), and other unknown factors C_2. Thus, we can write the PRB Utilization equation as below for a cluster of eNodeBs:

$$median\,\{PRBU_T\} \times BW \triangleq \lim_{T\to+\infty} \frac{1}{T} \sum_{t+1}^{t+60} \sum_{e\in E} \times \left(Vol_{T,e} + UE_{T,e} - D_{T,e}\right) + C_2 \tag{2}$$

In the above equation (2), we have considered only $\overline{UE}$ (User Equipment), because $Max\,(UE)$ varies in a certain geographic area or cluster only because of special circumstances and social events. Prediction of max UE for a particular eNodeB could be another research question we will address in our future work. However, we have considered average $\overline{UE}$ in the computation process, representing the number of connected user equipment in a particular eNodeB for a specific time frame. eNodeB-wise count of $\overline{UE}$ for a particular hour depends on cellular network operators' market share and population of that eNodeBs coverage area. So, in most cases, the $\overline{UE}$ will not change drastically for the yearly business plan (BP). In standard network conditions, there is minimum eNodeB downtime (D_T), where D_T negatively impacts traffic volume and PRB Utilization. So, it's easily understandable that future traffic is the most vital thing for predicting utilization as well as user throughput. If we can rightly predict the traffic or user throughput, then it's possible to take action to maintain the quality of service. In equation (2), $\sum_{e\in E} Vol_{T,e}$ is the summation of all eNodeB (E) traffic in a cluster. By taking into consideration all of these actual network factors, we can simplify the PRB Utilization equation for one single eNodeB -

$$PRBU_{t\in T} \times BW = \lim_{T\to+\infty} \frac{1}{T} \left(Vol_{T,e} + \overline{UE}_{T,e} - D_{T,e}\right) + C_2 \tag{3}$$

As eNodeB-wise bandwidth (BW) and $\overline{UE}$ is almost constant for a particular network planning year, so it can assume that PRB utilization is directly proportional to traffic growth and bandwidth of a particular spectrum band. In equation (2) and (3), time $T = \{t+1, t+2, \cdots, t+60\}$, that means maximum 60 days hourly future PRB Utilization is denoted as $PRBU_{t+60}$, where Vol_{t+60} indicates predicted traffic volume (unit bits) in the same time frame.

Understanding the unleash potential of deep learning algorithms for time series data prediction, we have built prediction models for future traffic volume and PRB utilization with a unique deep learning algorithm fusion strategy. We have not limited our research work to identifying future traffic and PRB utilization. Based on an assessment of predicted PRB utilization from traffic, we develop an algorithm for radio network parameter estimation for triggering the action of maintaining network QoS benchmark.

C. RELATED WORK

This research majorly focused on network traffic forecasting and estimation of radio parameters based on forecasted traffic. Therefore, we have discussed the state-of-the-art network traffic forecasting methods and LTE radio parameter estimation method sequentially. There are some related works on cellular network traffic forecasting using different techniques. However, at the same time, only a few researchers dealt with LTE radio parameter optimization based on forecasted traffic. In this section, we will discuss how other related works are different from this research work and what are the unique contributions of this research paper.

Traffic prediction or forecasting is vital for anticipating cellular network status, identifying user usage patterns, and estimating quality-of-service or major resource allocation parameters [3]. Fang et al. [10] revealed city-scale level traffic forecasting based on a cell handover-aware graph neural network. Xu et al. [11] demonstrated the geographical distribution of forecasted traffic heatmap in a particular city by analyzing time series data. Kirmaz et al. from Nokia Bell Lab [12] also illustrate similar research by dividing the geographic area into pixels. All these three research [10], [11], [12], focuses on predicting traffic based on a geographical unit of measurement, and we know a geographical area could be covered by multiple LTE eNodeB or cell. However, our research focused on each eNodeB or cell level of traffic prediction. Another research by Trinh et al. [6] introduced mobile traffic forecasting with RNN in daily level time series. Sun et al. in their research work [13] represented network-level future mobile data estimation based on user mobility patterns. Any time scale of traffic forecasting may help network planning activity to some level of degree, but our developed model will predict traffic at an hourly level. Hourly level traffic prediction will enable more insights into time series data and easily convertible to daily level [6]. L. Lo et al. [14] developed a Thresholded Exponential Smoothing and Recurrent Neural Network (TES-RNN) model to manage network traffic and resources with Joint Statistical Modelling and Machine Learning as a hybrid approach. This research [14] only focused on predicting traffic anomalies at a particular time, rather than on hourly or daily regular traffic. Q. Yu et al. exercised Graph Attention Network (GAT) and Temporal Convolutional Network (TCN) to model to predict traffic overload considering large amounts of small-scale redundant data [15]. As long as the mobile operator's network planning is concerned, eNodeB or cell-level granularity is essential. So, unlike most related research work, we focused on cellular network traffic forecast at the granular cell level. Each cell is considered a different eNodeB. Apart from that, there is another benefit

of evaluating cell or eNodeB-level forecasting, network planners can easily convert the cell level forecast into city or province level by simply adding all eNodeB traffic in that geographic area. There is also different time horizon in traffic forecasting research works, but considering the real-life challenges of network planning, we focus on developing an hourly traffic forecast model. In Summary, granular data traffic prediction in terms of two major factors (hourly in terms of time and eNodeB in-network or geography) is the key difference between our research and other similar works.

The second part of our research is to identify future network utilization based on predicted traffic and propose an algorithm for handling expected traffic by estimating LTE radio parameters. There is some sporadic research on radio capacity analysis at different times. Jang et al. [16] develop a model to estimate the resource block usage rate to solve the fixed-length input problem in the traditional RNN model. However, this research [16] doesn't address how radio parameters can be used to estimate RB usage rate (RBUR). Hasan et al. [17] proposed an algorithm for Adaptive Mobility Load Balancing to maintain throughput in LTE Small-Cell networks, which was a reactive process; any proactive measure was not defined in [17]. Most importantly, none of these research works devises any radio parameter estimation model from predicted future traffic. Our research investigated this particular issue and proposed an algorithm that triggers radio parameter solutions based on forecasted traffic.

D. CONTRIBUTIONS

The major contribution in this research paper are summarized as follow:

1) We proposed a state-of-the-art fusion model for forecasting network traffic using different deep learning algorithms (LSTM, BiLSTM, and GRU), which increase overall model performance. Unlike most of the research, we have considered multivariate inputs for modeling of forecasting Mobile network data traffic. We have predicted eNodeB-level utilization (or cell load) based on forecasted traffic. The eNodeB-level utilization prediction from deep learning model-based traffic forecasting technique is one of the significant outcomes of this research work, which will help MNOs to decide about network expansion for maintaining benchmark QoS.
2) To forecast data traffic, we categorize all eNodeBs into different clusters. In order to obtain clustering accuracy of time series data, we have considered both Euclidean and Dynamic Time Wrapping (DTW) algorithm-based Self Organized Map (SOM). In addition, we have compared both algorithms for a particular cluster. However, due to computational accuracy, finally, we have recommended DTW-SOM for the rest of the analysis.
3) We also proposed a unique radio parameter estimation algorithm to ensure the Quality of Service (QoS) benchmark. In this way, the network planner can initiate plans to meet customer demand before it crosses the QoS threshold or causes massive degradation in customer experience.

The rest of the paper is demonstrated as follows. Dataset description is presented in section II. Apart from that, relevant definitions and formulas are also added in the Data-Set Description section. Then, the proposed System Model for this research is illustrated in section III, along with visual and text explanation. Section IV explains a step-by-step methodology for cellular network traffic prediction. Model Performance evaluation is discussed with standard performance criteria in Section V. After that, the Experimental Outcome derived from the proposed model is shown in section VI. Later on, the unique contribution of this research paper, Optimized LTE Radio Parameter Estimation based on predicted traffic, is articulated in section VII. Finally, in section VIII, we conclude the paper by summarizing the overall work and discussing future work.

II. DATASET DESCRIPTION

The LTE 4G dataset was identified and downloaded from the Operations Support System (OSS) of one of the MNO. Initially, hourly Data traffic from the Radio Network end was collected from around 890 eNodeB for 351 consecutive days, including 8424 samples. After collecting a total of approximately 6.2 million dataset, data masking was done to ensure data privacy. Apart from that, other associate features from eNodeB were collected, i.e., Utilization, Max_UE, Avg_UE, Cell_TP, User_TP, and traffic. This analysis only focuses on Downlink (DL) traffic, as it has the most significant contribution to cellular network utilization.

The Cellular Network dataset contains the following information and features used in this work:

- eNodeB: eNodeB is the Radio network element of the LTE network, which is also known as Evolved Node B.
- Traffic: Traffic means a combination of Uplink (UL) and Downlink (DL) internet Traffic from the Radio network end. The counter formula of traffic is as below:

$$\sum \text{downlink traffic volume for } PDCP + \sum \text{uplink traffic volume for } PDCP$$

  The unit of traffic is Gigabits here.
- Utilization: Utilization indicates the usage of Physical Resource Block (PRB) in LTE system. The higher number of utilization indicates more usage of LTE resources. Utilization can be formulated in counter level by the below formula:

$$\frac{AvgnumberofusedPRBs}{NumberofavailablePRBs}$$

- Max_UE: Maximum number of Users connected at an instance in a particular node considered as Max_UE.
- Avg_UE: Avg_UE is the average number of connected Users per hour in a particular node

- Cell_TP: Cell_TP means Cell Throughput, which is the sum of all users' throughput in a particular eNodeB or any node for a unit time frame. The counter level formula can be represented as below:

$$\frac{\sum \text{downlink traffic volume for PDCP}}{\sum \text{duration of downlink data transmission in a Node}}$$

- User_TP: A particular user receives an amount of data on average, known as User Throughput or User_TP. In other words, the average number of packets received by the User in a unit time frame. The counter level formula for User_TP as below –

$$\frac{\left(\sum \text{DL traffic} - \text{DL traffic volume sent in last TTI}\right)}{\text{Data transmit duration except last TTI}}$$

During data modeling of traffic forecasting for utilization prediction, all eNodeB has been classified in different classes according to their time-series behavior. The detailed classification procedure will share later part of this paper.

III. SYSTEM MODEL

In this system model section, the cellular network traffic prediction system model is presented. In Fig. 1, we have proposed a cellular network traffic prediction model using deep learning algorithms.

In the First step, eNodeB-wise traffic and other associate parameters have been collected through a rigorous data mining process. After collecting data, it is stored in the local database as LTE data is complex and combines several underlying features and information. As the key objective of this research, the model predicts traffic, so after data storage, we have done Exploratory Data Analysis (EDA) part. As we know, Exploratory Data Analysis (EDA) is the process of identifying major features and patterns in datasets. From EDA, we have identified any data missing in the datasets.

Missing data is filled in two steps. We use the mean of that particular day of the month for short or discrete data missing for a particular hour. In case of larger data missing (More than an hour), we use prediction from the previous trend and predict the missing period.

After EDA and missing data filling, the dataset is used in two folds per the system model. In one part, we try to make a cluster of eNodeB based on their time-series traffic pattern. The short-term objective of this part is to categorize similar consumption patterns eNodeBs. So that, later on, we can compare cluster-wise resource utilization concerning predicted traffic. As we have 890 eNodeB considered in this research, performance visualization will not be possible without clustering. We have followed the SOM-DTW-based clustering model for time series unsupervised data clustering. Details of the SOM-DTW model are discussed in the methodology part.

In the next part, we first extracted critical features of the dataset through feature engineering. Essential feature means which information has highly correlated with traffic data. As we considered multivariate inputs for the traffic prediction model, those inputs have different units. For this reason, data normalization is necessary to avoid systematic bias. We have used the min-max method in this work to transform all multivariate inputs from zero to one. Scaling input data helps reduce biasness as well as increase the accuracy of the traffic forecasting model. Equation (4) is used for data transformation:

$$z_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \left(New_{\max_x} - New_{\min_x}\right) + New_{\min_x} \quad (4)$$

Maximum data denoted as $x_{\max}$, and $x_{\min}$ is the minimum of the data. $New_{\min}$ and $New_{\max}$ is the zero and one respectively [18]. After Normalization and transformation, we divided data into two parts: test and train. In this research, we have split the training and test data ratio as 79:21 for 290 days hourly data of 890 eNodeB. The remaining 61 days of hourly data are kept as a validation dataset.

We have a training model with the help of the fusion strategy of deep learning algorithms LSTM, BiLSTM, and GRU, which are discussed in detail in the Methodology section.

After that, we completed the most crucial segment of this work, called traffic prediction. Based on predicted traffic with the help of Deep Regression, we forecast the utilization and compare overall cluster utilization.

IV. METHODOLOGY

In this section step by step deep learning structure-based cellular network traffic forecasting will be discussed, along with different algorithms for time series prediction.

A. AGGREGATION OF DATA SETS

This dataset contains all encrypted eNodeB-wise parameter information located in a densely populated city in South Asia. Let's assume the whole dataset as a $E_t = \{E_{c1t}, E_{c2t}, \ldots E_{cit}\}$, Where E_t is the sets of all eNodeB and E_{cit} is the all features of each individual eNodeB regardless of time (t). So, the aggregated eNodeB-wise traffic (Tr) in a time frame T is,

$$A(T) = \Sigma_{r(t)\in R(T)} E(t) \Sigma_{t\in T}\, a(t) \quad (5)$$

B. PATTERN IDENTIFY ON DATASET

Pattern identification on the dataset is one of the key elements before modeling any dataset. In this research, we try to understand the dataset first because it gives some idea about how traffic changes over time and the vital contributing hours for overall data traffic. From the above four figures, we can quickly identify the pattern of the hourly traffic dataset of 351 days. Below equation (6) used to identify data patterns per eNodeB.

$$E(t) = \sum_{i=1}^{351} (\text{Traffic in hours}) \,/\text{Number of Days} \quad (6)$$

From Fig. 2, it can be easily understood that traffic is increasing over the period. We have also noticed the hourly traffic difference between weekdays and weekends in Fig. 4.

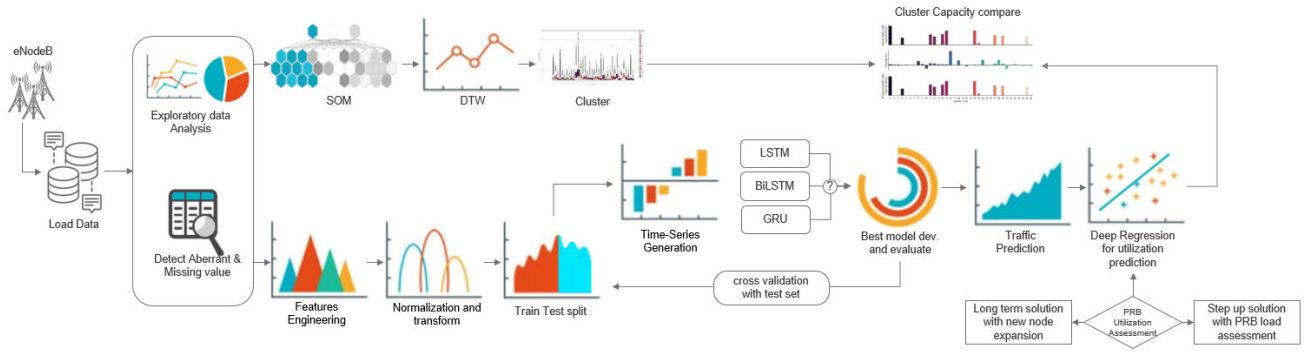


FIGURE 1. Proposed system model of cellular network traffic prediction and PRB utilization based optimized parameter estimation.

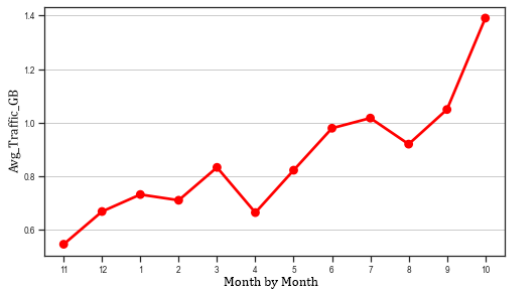


FIGURE 2. Month by month average traffic (GB) per eNodeB.

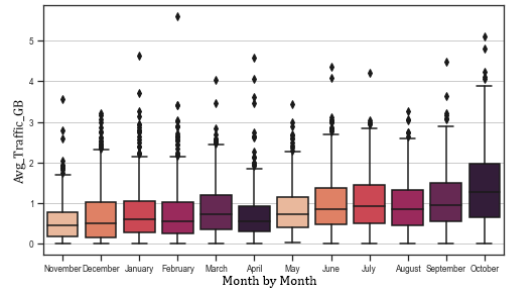


FIGURE 3. Boxplot of month-by-month average traffic (GB).

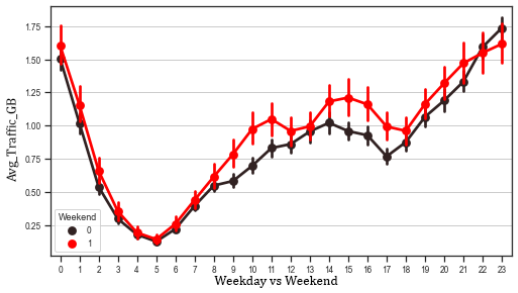


FIGURE 4. Weekday vs. Weekend hourly traffic pattern.

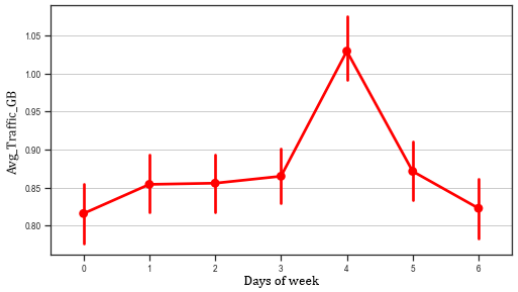


FIGURE 5. Daily average traffic (GB).

The Monthly average traffic box plot is showing the median (Q2) traffic is increasing in every month Fig. 3. In Fig. 5, it represents the one special day in a week when traffic is almost double the rest of the days.

C. FEATURE CORRELATION PLOT

In this research for predicting future eNodeB-wise traffic, we have collected five additional features, which are important parameters for predicting cellular network traffic and understanding the overall LTE network for dimensioning. For the Fig. 6 correlation (r) plot, we used the below equation (7) for each pair –

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{7}$$

The graphical plot of correlation Fig. 6, found that utilization is directly correlated with traffic, whereas User_TP TP (User Throughput) is negatively correlated with traffic and utilization. This implies that more traffic will cause higher utilization and lower User_TP or degradation of the quality of services (QoS). Our goal is to keep utilization at an optimum level from a network design perspective. If utilization becomes higher, it means intolerable traffic for that eNodeB; network expansion needs to trigger. By this, MNOs can keep standard utilization and QoS.

The later part of this research will devise more precious recommended utilization parameters for keeping QoS in the desired range.

D. SELF-ORGANIZING MAP (SOM) AND DYNAMIC TIME WRAPPING (DTW) BASED CLUSTERING

We found a lot of unstructured data from the traffic and associated parameter dataset. Leveling those data is an expensive, time-consuming, and challenging task. However,

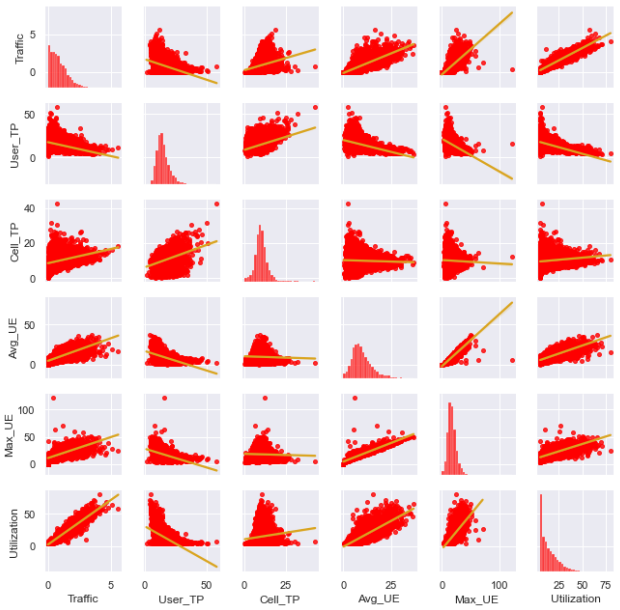


FIGURE 6. Correlation Plot from different features.

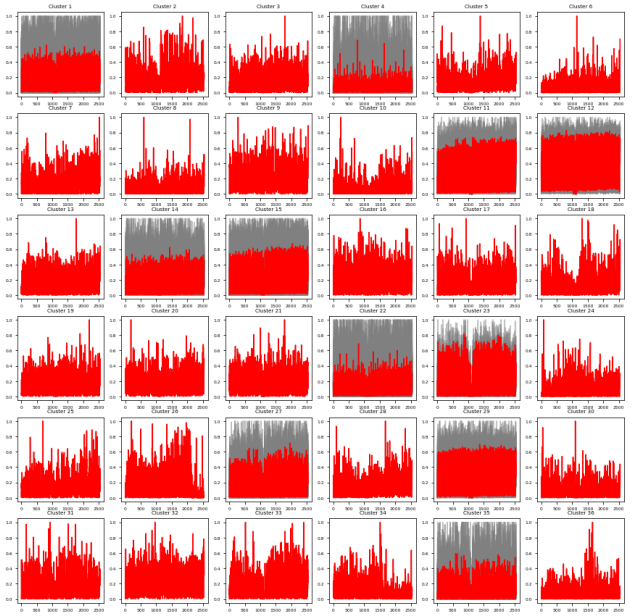


FIGURE 7. DTW matching based 36 SOM cluster according to eNodeB hourly time series data pattern of 351 days.

we had to overcome the unsupervised data leveling challenges to develop this intelligent traffic forecasting and eNodeB utilization model. It is well established that supervised algorithms perform better when data are labeled properly. More numbers of datasets will increase the accuracy of the model [19]. In this research work, we have applied Self-Organizing Map (SOM) based eNodeB clustering based on their hourly time series data.

SOM is one kind of unsupervised neural network with only two layers [20]. One input layer and another mapping layer also work as output layers. In SOM-based clustering, each neuron of the input and mapping layers is wholly connected. As per the working principle, each mapping neuron searches whose weight is most like input vectors while developing the SOM-based cluster iteration model. The best-matching neuron pair is known as Best Matching Unit (BMU) [21], [22].

This research work has tried to identify both Euclidean matching and DTW matching for creating SOM clusters. Finally, choose the matching algorithm for cluster creation. In the first phase of work for the Euclidean matching-based model, Mobile network time series data input is considered in d dimensional (i.e., there are d input units). So we can write the input patterns as $x = \{x_i : i = 1, 2, 3, 4, \ldots d\}$. if, i is the connecting weights between the input unit and the neurons j in the computing layer can be written $w_j = \{w_{ij} : j = 1, 2, 3, 4, \ldots N; i = 1, 2, 3, 4, \ldots d\}$. N Considered as a total number of neurons. According to Euclidean distance ED between the input vector x and the weight vector w_j for each neuron j –

$$ED_j(x) = \sqrt{\sum_{i=1}^{d} (x_i - w_{ij}) \wedge 2} \tag{8}$$

Using equation (8) of Euclidean distance matching, all 890 eNodeB divided into 36 clusters. Equation (8) is also known as a Pythagorean theorem in Cartesian coordinates. So as per the working principle, Euclidean distance is the length of a line segment between two points. While combining similar traffic carrying eNodeB's in a cluster, SOM-Euclidean matching will only calculate the based on two points (starting and endpoints). Which may not reflect the best-represented cluster trend-line for all eNodeBs [21], [23]. To overcome this problem of Euclidean Matching-based SOM clustering, we have proposed another method that is called Dynamic Time Wrapping (DTW) driven SOM clustering. Unlike Euclidean distance, DTW-based clustering is not only limited to staring and endpoint-based calculation. Here, in DTW, the Best Matching Unit (BMU) neuron is searched for a minimal DTW sample of eNodeB-wise traffic data Fig. 7. We have also used the distance decay kernel function in the DTW-based distance method [24].

$$W_{dt^{New}} = W_{dt^{previous}} + \Theta \bullet K_{rs} \bullet \left(x - W_{dt^{previous}}\right) \tag{9}$$

DTW matching updates the model based on learning rates, while Euclidean matching only calculates the distance from two points. As a result, DTW-SOM enables more accuracy in cluster representative trend-lines as similar as Fig. 8.

The objective of this clustering to group the eNodeB and benchmark the performance which is done in later part of the paper.

E. MULTIVARIATE DEEP LEARNING ALGORITHMS FOR TIME SERIES TRAFFIC PREDICTION

1) *Multivariate Time-Series Traffic Prediction:*
Actual Cellular Network traffic is not only related to the previous data trend. Several factors may affect

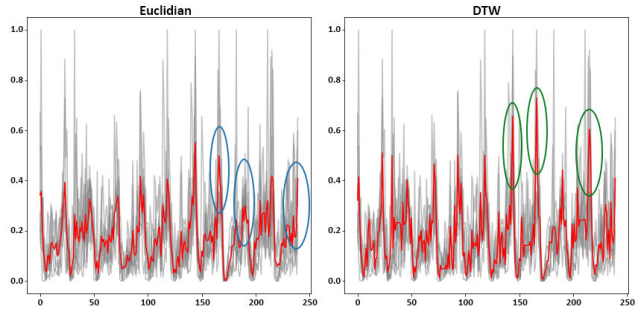


FIGURE 8. DTW-based cluster represented traffic trend line able to capture each spike or pattern of long traffic trend (circular green marked), whereas Euclidean based clustered trend line missed those details.

the data traffic volume of a particular eNodeB. For instance, traffic may fall drastically if an eNodeB is down for a higher time than regular. As well as any social and religious event in a particular area that causes more people to gather under one or more eNodeBs in that location is also the reason for increasing traffic. So, considering those real network dimensioning challenges, we considered Multivariate input-based time-series traffic prediction [25].
If we consider both the factors as affecting elements for traffic forecasting. In other words, denoting the related variables by $x_{1,t}, x_{2,t}, \cdots x_{k,t}$ and at the end of t time traffic $T_{1,t}$ can be represent as equation (10)

$$T_{1,t} = f1\left(x_{1,t}, x_{2,t}, \cdots, x_{k,t}, x_{1,t-1}, x_{2,t-1}, \cdots, x_{k,t-1} \cdots\right) \tag{10}$$

After forecasting traffic $T_{1,t}$, *next t+1* time traffic will be dependent on all previous stage variables. Considering this logic equation (10) can be written as below for t+k time predicted traffic $T_{k,t+k}$:

$$T_{k,t+k} = f_k\left(x_{1,t}, x_{2,t}, \cdots, x_{k,t}, x_{1,t-1}, x_{2,t-1}, \cdots, x_{k,t-1} \cdots\right) \tag{11}$$

As per the working principle of multivariate time series analysis, where different variables are dependent on their previous value as well as other feature or variables [25]. Like univariate time series, the major objective of multivariate time series prediction is to get the data forecast. But multivariate function enables more accurate results with the help of other associate parameters, which we include in this research work.
As represented in the deep neural network model architecture in Fig.9, after collecting raw time-series input data, key features are extracted from dataset. Later on, we generated time series from this feature extracted data by using the sliding window technique algorithm. The sliding window technique works on N-1 historical time series data [26]. The working principle is after feature extraction, *ts function* generate the time series data (as shown in Fig.9) and explained by algorithm 1.

Algorithm 1 Time-Series Generation With Sliding Window Technique

Data:
A: array of traffic and feature[1]
p: number of days in past as sliding window
f: number of total features
Result: return array of X and target Y

initialization;
$x, y \leftarrow 0$;
for $i \leftarrow p$ **to** $length(A)$ **do**
  $append(A[i-p:i, 0:f])toX$
  $append(A[i:i+1, 0])toY$
end
return X, Y

2) *Fusion Strategy in Recurrent Neural Networks:*
This research work introduced a fusion strategy in multiple Recurrent Neural Networks for building a data traffic forecasting prediction model. We have proposed Long Short-Term Memory (LSTM), BiLSTM, and GRU algorithm-driven multivariate Deep Neural networks for multiple parallel time-series predictions, as shown in Fig. 9. Based on eNodeB-wise performance of the RNN model, we proposed the fusion strategy.
Long Short-Term Memory (LSTM)
The most widely used recurrent neural network (RNN) algorithm processes the inputs individually. However, the performance of RNN degraded over long time series or sequences as it has no memory in architecture [27].
We have decided to use Long Short-Term Memory (LSTM) instead of the traditional RNN model to overcome this memory issue. Long Short-Term Memory is also an advanced version of the recurrent neural network (RNN) model; the architecture allows the model of chronological sequences and their long-range dependencies more precisely than conventional RNNs. LSTMs was also designed to deal with the long-term dependency problem faced by standard RNNs.
As per LSTM architecture for computing forecasting traffic or any sort of time series data starts with the calculation of output value from previous time data and presents input series data which is enabled as an input of forget gate [28], [29].

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{12}$$

[1]Algorithm 1 involves the use of multivariate input data (consisting of time-series historical traffic, downtime, and user counts) represented as X, and output data (traffic at t+1) represented as Y. The training data consists of 79% of the total data, with the remaining 21% used as test data. The algorithm employs a sliding window technique, in which both X and Y are shifted from one window to the next with respect to array A[i..]. This can be seen in the third part of Fig. 9.

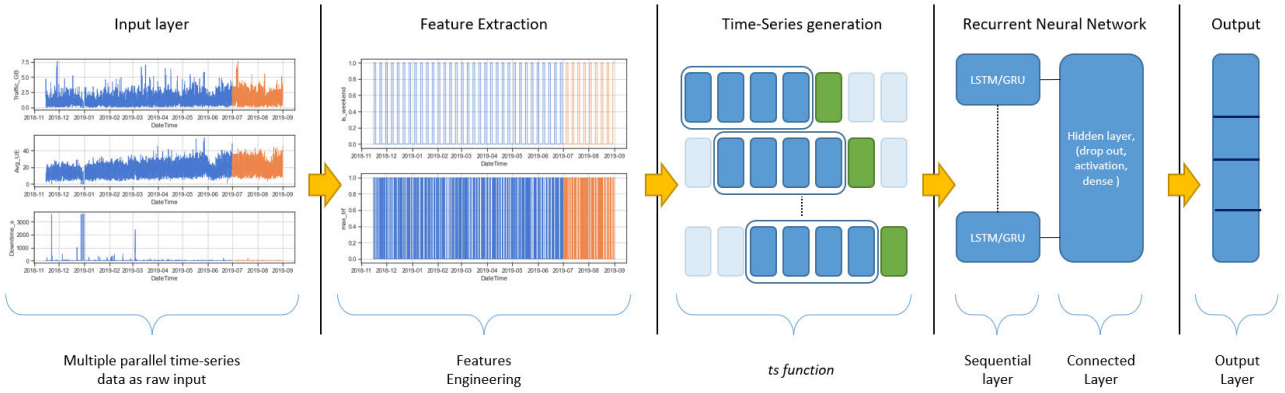


FIGURE 9. Architecture of the proposed multivariate Deep Neural Network for multiple parallel time-series prediction.

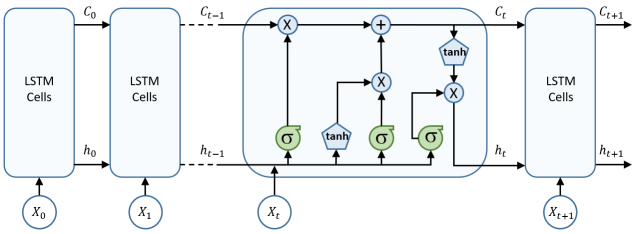


FIGURE 10. LSTM architecture for predicting future traffic.

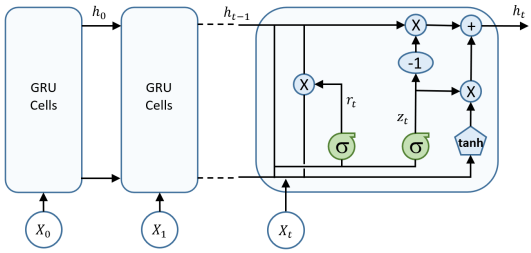


FIGURE 11. GRU architecture for predicting future traffic.

Here h_{t-1} is the output value of the previous time, as well as x_t, denotes the input value of the present time. f_t is the output gate which value rage is (0,1). The weight of the forget gate is represented as W_f, where b_i is the bias of that forget gate. In addition of that, input to input gate, output value and condition of candidate cell at input gate can also be calculated through output value of previous time and the input value of present time, which can be calculated through the below equations –

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \quad (13)$$

$$\tilde{C}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right) \quad (14)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (15)$$

$$O_t = \sigma\left(W_0 \cdot [h_{t-1}, x_t] + b_o\right) \quad (16)$$

$$h_t = O_t * \tanh\left(C_t\right) \quad (17)$$

In these equations (14),(15) and (17), C_t is the cell state of the candidate cell in t time, which value ranges (0,1). O_t denotes the output gate, i_t is the input gate, and h_t is the hidden layers in the cell. Here, x_t is the cellular network data traffic. The bias of the network indicates by b function.

This LSTM is used as a sequential layer for building traffic forecasting model. This LSTM architecture is modified from [25] and [30]. From the above equation (12),(13),(14) and (16) information transfer is based on dot product outcome. If the dot product result is zero, it means information is not transferred [27]. Information will transfer, in case of dot product outcome is one.

Bidirectional LSTMs (BiLSTM)

In this research Bidirectional Long Short-Term Memory (BiLSTM) is also used to create fusion strategy. While training a system, the BiLSTM model utilized input data in both directions, which means at first in right to left, then left to right. This process of twice the operation of LSTM improves the performance and accuracy of the BiLSTM model by removing a long-term dependency [31]. Due to this in some cases model perform better than LSTM, which is discussed later part of this paper.

Gated Recurrent Unit (GRU)

In this research for fusion model building GRU is also used. As GRU is a comparatively recent RNN model introduced by Kyunghyun Cho et al. in 2014, which has almost similar architecture compared to LSTM. But GRU models are more convenient and more straightforward for training and implementation.

Typical GRU model architecture in Fig. 11. This GRU architecture is modified from [32]. GRU neural network architecture reduced computational due to its existence of update and reset gates, which also enables remembering the cell's long-term states [33]. The reset gate in GRU model works similarly to LSTM forget gate. In GRU, hidden state output at time t can be

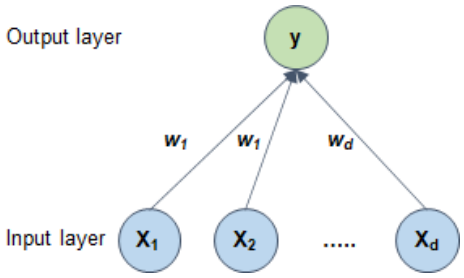


FIGURE 12. Single layer regression with deep neural network.

calculated as below general expression:

$$h_t = f\left(h_{t-1}, x_t\right) \tag{18}$$

In equation (18), h_{t-1} is the hidden state status in $t-1$ time and x_t input time series value at t time. For explaining to the GRU NN model as shown in architecture (Fig. 11) below equation can be used –

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \tag{19}$$
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \tag{20}$$
$$\tilde{h_t} = \tanh\left(W_{\tilde{h}}\left[r_t * h_{t-1}, x_t\right]\right) \tag{21}$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h_t} \tag{22}$$
$$y_t = \sigma\left(W_o \cdot h_t\right) \tag{23}$$

In these equations (19), (20) and (23), Sigmoid function is represented as σ, which output is (0,1). r_t is the updated, which works for determining stored information quantity from one movement to another. Reset gate z_t determines the status of information of the last state, whether the information is kept or erased. The parameter which needs to train are denoted as W_r, W_z, W_h, W_o [34], [31], [35], and [33].

F. REGRESSION WITH DEEP NEURAL NETWORK

The regression technique enables solving the task of the critical problem of predicting continuous value based on input [36], [37]. In this research, we focused on utilization prediction after getting the predicted traffic based on the deep learning model. In this case, Deep Regression can predict utilization from eNodeB-wise forecast traffic (As shown in System Model, Fig. 1) from the equation (15)

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b \tag{24}$$

Here, w is the weight of input traffic x_1 to x_d, and b is known as bias or offset. Weight determines the influence of features in the model [36], [37] and [38].

V. PERFORMANCE EVALUATION

Model Evaluation Criteria:

The mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and squared correlation (R2) metrics are that considered as evaluation criteria. The equations are used to determine the difference between actual and predicted data. [39] The formula is describing as below in equations (25):

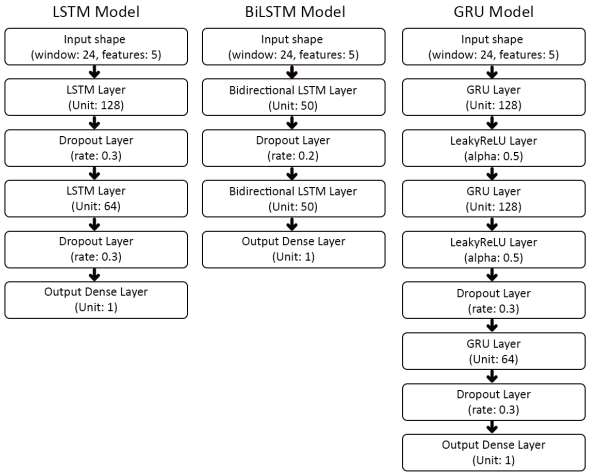


FIGURE 13. Sequential model for LSTM, BiLSTM and GRU.

$$MSE = \frac{1}{N}\sum_{k=1}^{n}\left(y_t - x_t\right)^2$$
$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{n}\left(y_t - x_t\right)^2}$$
$$MAE = \frac{1}{N}\sum_{k=1}^{n}\left|y_t - x_t\right|$$
$$R^2 = 1 - \frac{\sum_{k=1}^{n}\left(y_t - x_t\right)^2}{\sum_{k=1}^{n}\left(\bar{y} - x_t\right)^2} \tag{25}$$

Experiment results

The virtual environment setup is done with TensorFlow, scikit learn, and some standard python libraries like pandas, seaborn, etc. the system requirement is designed with OS windows10, processor Ryzen 5 3600, RAM 32GB, GPU RTX 3070. The parameter configuration for creating the model and taking Multivariate input below structure shows the best evaluation score.

For each model epochs size: 100, batch size: 128, Adam optimizers with learning rate: 0.001 is considered. For optimized and efficient training, some callbacks are used like *EarlyStopping*: To stop training when a monitored metric has stopped improving, *ModelCheckpoint*: To save the Keras model or model weights at some frequency and *ReduceLROnPlateau*: To reduce the learning rate when a metric has stopped improving.

Fig. 13 proposed multivariate LSTM, BiLSTM and GRU model for multiple time-series prediction with 5 features and 24-time steps for the prediction process All site 890 nodes train through the designed model and make predictions for the next 62 days. Evaluating the model's overall accuracy Evaluation Criteria are combined and shown in Fig. 14. This plot is used to

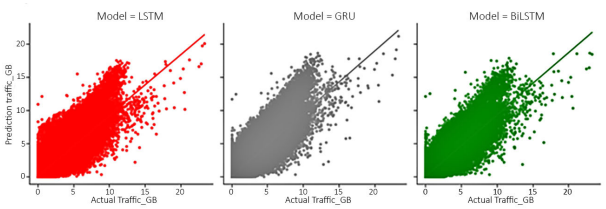


FIGURE 14. Regression plots of the models at the training phase.

TABLE 1. Performance of the model in the testing phase.

Model	MSE	MAE	RMSE	R^2
LSTM	0.4478	0.4355	0.6692	0.7635
GRU	0.4461	0.4346	0.6679	0.7644
BiLSTM	0.3922	0.4158	0.6262	0.7929

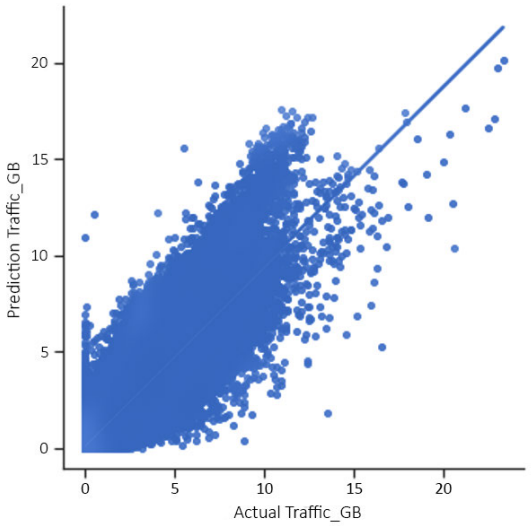


FIGURE 15. Regression plots of the Fusion Model.

find the predicted and actual values relationship. Also, Table 1 presents the testing results of the proposed model.

The observation from each node shows that out of 890 nodes in 65% nodes, BiLSTM provides a high score, 24% in LSTM and 20% in GRU. As different node traffic patterns are different, the model forecast precision was different. In this scenario, a fusion model was approached where the best model was selected based on training accuracy and minimal loss. The prediction R^2 score was 0.8034 for over all system from the fusion model. The Fig. 15 and Table 2 show that the proposed model's experimental results in the testing phase were optimal. Comparative analysis between the descriptive statistics of actual and predicted values of all node indices as outlined in Table 1 indicates that both data sets retain considerably similar values. A sample eNodeB's Traffic patterns with different model and R^2 value is also shown in the Fig. 16. The result confirms

TABLE 2. Performance of the fusion model.

Fusion Model	
MSE	0.3723
MAE	0.4025
RMSE	0.6101
R-Square	0.8034

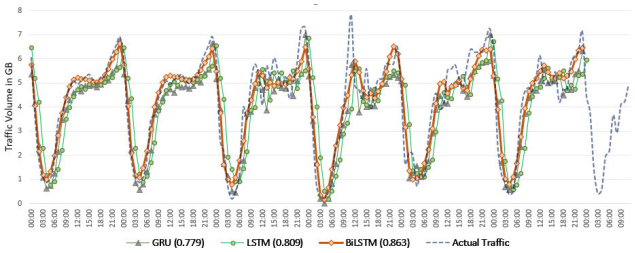


FIGURE 16. Actual vs. Predicted traffic based on different Model.

that both actual and predicted data set have similar basic nature and, therefore, can be concluded that the proposed multivariate fusion model is capable of time-series traffic forecasting.

VI. EXPERIMENTAL OUTCOME

With this system design and trained model, we have predicted the next two-month traffic as well as forecast utilization of that particular eNodeBs of the cluster. In this Table 3, $\sum Vol_{Actual}$: Pre-actual Total traffic (GB) in the last 2-month, $\sum Vol_{Predict}$: Post-predicted Total traffic (GB) in the Next 2-month, *Count* ($PRBU_{Actual}$): count of sample utilization > 70% in the pre time frame, *Count* ($PRBU_{Predict}$): count of sample utilization > 70% in post time frame and % indicate the % change of the pre and post traffic. Based on that Table 3 shows the cluster-wise total traffic for the last 2-month time frame next 2-month time frame. Also, the table outlined the count of sample utilization > 70% within the respective time frame.

VII. OPTIMIZED LTE RADIO PARAMETER ESTIMATION

Several LTE parameters are mainly contributing to the quality of services. Before proposing an algorithm for LTE QoS parameter estimation, we need to discuss those key parameters

1) PHYSICAL RESOURCE BLOCK

A resource block (RB) is the smallest unit consisting of resource elements in the LTE air interface. One physical resource block (PRB) spans 12 sub-carriers; each sub-carrier is 15-kHz spacing corresponding to 180 kHz [39], [40]. When the scheduling algorithm uses full-PRB transmission, the smallest time-frequency resource that can be scheduled to a device is one PRB pair mapped over two slots RB; in the normal CP length total 14 OFDM symbols corresponds to over 12 sub-carriers as illustrated in Fig. 17.

TABLE 3. Experimental outcome table.

Cluster	Vol_{Act}	Vol_{Pred}	%	U_{Act}	U_{Pred}
1	234193	242947	3.74	1015	1697
2	2215	2276	2.77	86	14
3	1092	1126	3.17	0	0
4	88575	86776	-2.03	64	197
5	1565	1599	2.2	0	0
6	464	461	-0.62	0	0
7	1219	1218	-0.04	0	0
8	719	852	18.51	0	0
9	1067	1123	5.25	0	0
10	424	355	-16.31	0	0
11	128599	140779	9.47	8202	8846
12	103936	111979	7.74	11570	14646
13	3485	3630	4.18	2	0
14	133977	138778	3.58	1470	2053
15	174525	182734	4.7	2704	3105
16	1085	1791	65.13	0	0
17	1703	1699	-0.23	0	0
18	1068	1301	21.85	0	0
19	1343	1326	-1.23	0	0
20	2461	2308	-6.22	9	9
21	1574	1579	0.33	0	0
22	178059	182311	2.39	329	756
23	23017	23647	2.74	1193	547
24	321	400	24.44	0	0
25	919	896	-2.56	0	0
26	853	847	-0.63	1	0
27	118642	128411	8.23	3146	4626
28	1739	2164	24.42	0	0
29	113749	122865	8.01	5747	5883
30	1189	963	-19.04	2	0
31	1496	1520	1.61	0	0
32	3473	3697	6.45	105	150
33	2015	2060	2.23	8	0
34	815	758	-6.95	0	3
35	106649	108547	1.78	703	657
36	521	473	-9.25	0	0

- 1 RB = 12(Sub-carriers) × 7 (Symbols) = 84 Resource Elements. (For Normal CP: 7 symbols)
- 1 RB = 12(Sub-carriers) × 6 (Symbols) = 72 Resource Elements. (For Extended CP: 6 symbols)

2) CHANNEL QUALITY INDICATOR (CQI)

Channel Quality Indicator (CQI) Indicates the quality of the carriers and reports sent from the UE to eNodeB. The LTE contains 15 different CQI values ranging from 1 to 15, and depending on the reports network transmit data with different transport block size [39], [41].

3) BLOCK ERROR RATE

The ratio of erroneous blocks and the total number of blocks indicate the Block error rate (BLER). The technique used to detect errors in the transport block is CRC. If the calculation does not give the desired results, the receiver will request HARQ NACK for re-transmission. To ensure service quality, 90% successful transmission at the receiver end means the typical BLER target should be 10% [42]. If the BLER target is failed to achieve, then more re-transmission might be required, which causes more radio resource consumption. Optimized and precious resource scheduling techniques are required to maintain the targeted QoS benchmark of maximizing throughput, experience fairness among the different users, and reducing reduced Block Error Ratio [43].

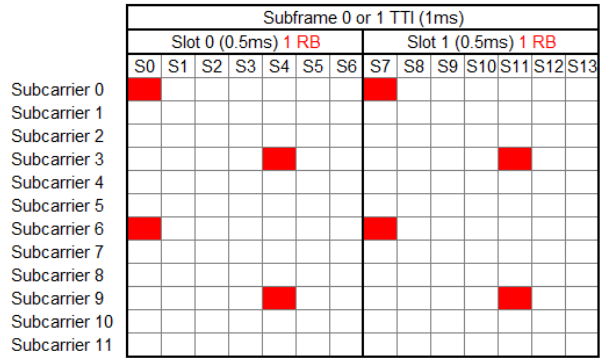


FIGURE 17. Physical resource block the LTE air interface.

A. PROPOSED ALGORITHM FOR QOS PARAMETER ESTIMATION

We have proposed an algorithm for radio parameter estimation based on predicted PRB utilization. As per the algorithm in the initial step, each eNodeB capacity requirement (Fig. 18) is estimated based on the assessment of 60 days of predicted PRB utilization. If 70% sample of 60 days eNodeB busy hour (NBH) is greater than or equal to 80% PRB utilization(X). It is recommended for hard capacity expansion with the implementation of Multibeam Cell Split Solution [44], New Spectrum addition [45], and planning and deployment of a new node [45], [46]. New node expansion always triggers higher capital expenditures. So, it should be the last option for maintaining QoS [47] if the PRB utilization (X) lies between $70\% \leq X < 80\%$, three soft step-up solutions will trigger according to the estimated parameter to reduce the PRB utilization and maintain the QoS. Actual PRB Utilization will be assessed during the soft step-up solution. Later in this paper, we have discussed why 70% and 80% are considered thresholds for triggering action points for Optimized QoS parameter estimation

Step up Solution 1: Adjustment of CQI Switch:

Downlink CQI adjustment, interactively compensates for inaccurate CQIs reported by UEs, optimizes MCS (Modulation and Coding Scheme) selection, and increases throughput [41]. If the network has moderate or heavy loads in the downlink, the downlink user-perceived rate will increase by 1% to 3% after CQI adjustment Fig. 18(a) [26].

Script 1 (CQI Switch adjustment)

MOD CELLALGOSWITCH: LocalCellId=0, CqiAdjAlgoSwitch = DlVarIBLERtargetSwitch-1 CqiAdjAlgoSwitch = DlEnVarIblerTargetSwitch-1;

`DlVarIBLERtargetSwitch`: In Adaptive configuration, The downlink target initial block error rate (IBLER) is

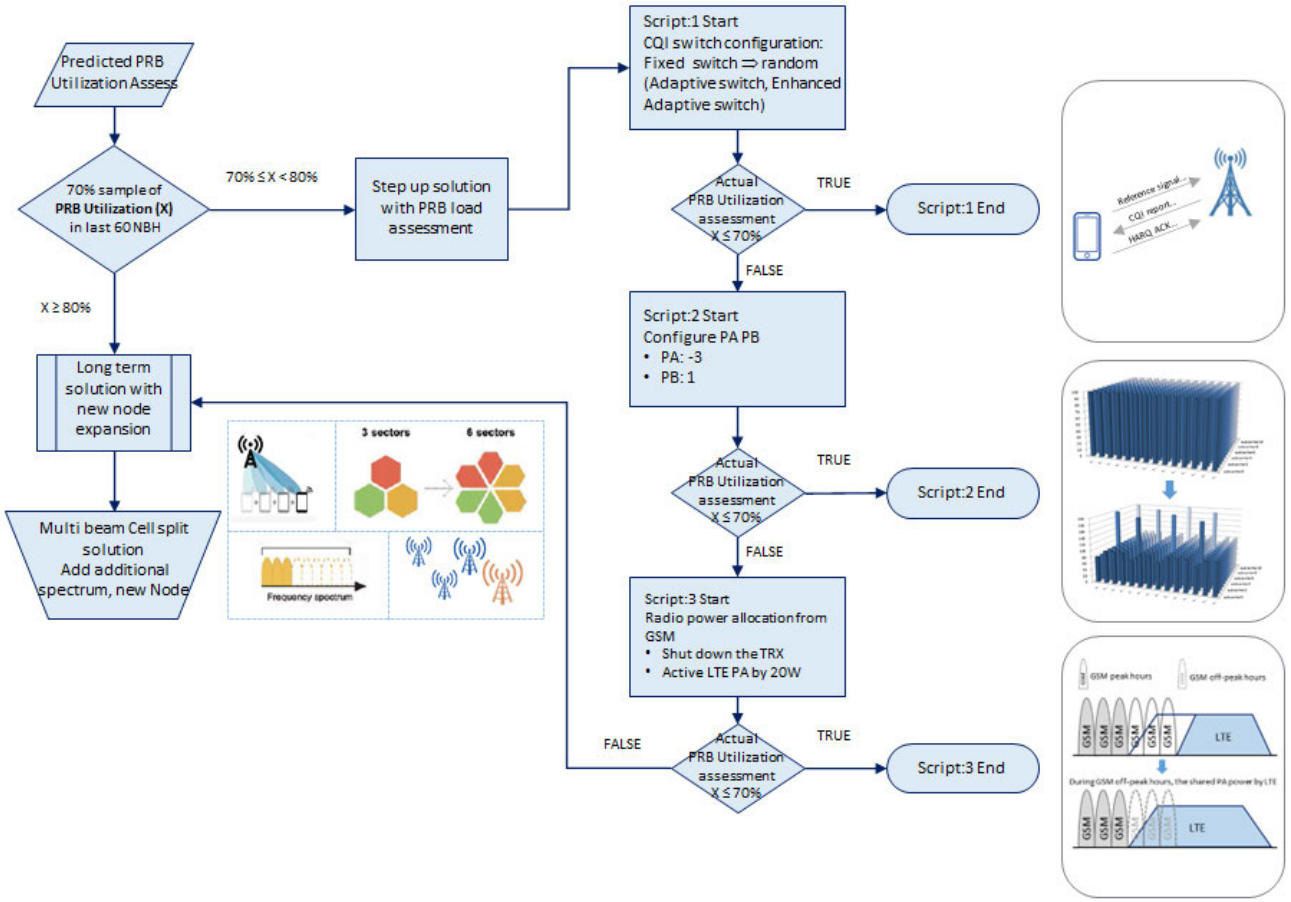


FIGURE 18. Algorithm for predicted PRB Utilization based QoS Parameter Estimation, (a) CQI switch adjustment, (b) Resource block power allocation, (c) Radio power allocation for LTE from GSM.

adaptively adjusted from a fixed configuration based on the Transport Block Size (TBS) to improve spectral efficiency. The higher the block size, the higher the throughput. In this scenario, the eNodeB adjusts the target IBLER to 10% for UEs with large packet services at non-edge locations and 30% for UEs with small packet services at edge locations.

`DlEnVarIblerTargetSwitch`: In Enhanced Adaptive configuration, the downlink target IBLER is adaptively adjusted from a fixed configuration based on TBS and as well as CQI fluctuation. In this scenario, the eNodeB adjusts the target IBLER to 5% and 10% for slightly fluctuated CQI values. With heavily fluctuated CQI values, the target IBLER is 10% and 30% [48].

Step Up Solution 2: P_A and P_B Power Allocation:

4G LTE RS RE Power (RSRP power) boosting depends on ρ_B/ρ_A parameters Fig. 18(b). this value is determined by many parameters, such as the max power of the RRU channel (P_{max}), the number of RBs of the cell Nrb, P_A, P_B etc. Below describes the impact.

- The higher P_A Implies, the lower RS power (ERS), the smaller cell radius and the higher throughput can get.
- With changing P_A, P_B need to change as well to make full usage of power.

TABLE 4. The cell-specific parameter for P_B.

P_A	ERS	ρ_B/ρ_A	P_B
0	EA	5/4	0
-3	$2*EA$	1	1
-1.77	$1.5*EA$	3/4	2
-6	$4*EA$	1/2	3

- The max value of ERS is determined by (P_{max}) and P_A.
- To configure RS power, first determine P_A, then RS and, P_B is determined following the Table 4 (for 2T cells).

The following definition from 3GPP. 36.213 protocol:

$TypeA$: the PDSCH OFDM symbol without RS
$TypeB$: the PDSCH OFDM symbol with RS
EA: the power of one element in $TypeA$, in W
EB: the power of one element in $TypeB$, in W
ERS: the power of $referenceSignal$, in W
RS: $referenceSignal = 10log\,(ERS * 1000)$, in dbm
P_{max}: the max power of RRU channel, in W
Nrb: the number of RBs in the cell
$\rho_A = EA/ERS$
$\rho_B = EB/ERS$
$P_A = 10log\,(\rho_A) = 10log\,(EA/ERS)$
$\rho_B/\rho_A = EB/EA$

TABLE 5. The cell-specific ratio ρ_B/ρ_A.

	ρ_B/ρ_A	
P_A	One Antenna Port	Two and Four Antenna Ports
0	1	5/4
1	4/5	1
2	3/5	3/4
3	2/5	1/2

Script 2 (P_A and P_B Power Allocation):

MOD CELLDLPCPDSCHPA: LocalCellId=0, PaPcOff=-3 dB; MOD PDSCHCFG: LocalCellId=0, Pb=1 `PaPcOff`: Indicates the PA to be used when PA adjustment for PDSCH power control is disabled, DL ICIC is disabled, and the even power distribution is used for the PDSCH [49].

`Pb`: Indicates the Energy Per Resource Element (EPRE) scaling factor index on the PDSCH. The value of this parameter and the antenna port control this scaling factor.

After executing script 2, PRB utilization (X) is expected to reduce by 70%. But if it does not happen next script will be executed.

Step Up Solution 3: Dynamic Radio Power Allocation from GSM:

As per definition, Dynamic Cell Power Off is a BSC feature [49] that enables power dynamically off or on the TRXs (in GSM Cell) based on the traffic demand of the co-coverage cell within a certain time frame. When LTE load is high, some radio power allocates from existing GSM PA to LTE PA through GSM TRX shutdown Fig. 18(c). In this way, the LTE network can be boosted up to 20W PA power.

Dynamic Power Allocation:

SET GCELLDYNTURNOFF: IDTYPE=BYID, CELLID=0, TURNOFFENABLE=ENABLE, SAMECVGCELLIDTYPE=BYID, SAMECVGCELLID=1, TURNOFFCELLSTRTIME= [Time PRB Util>70%], TURNOFFCELLSTPTIME= [Time PRB Util<70%] `TURNOFFENABLE`: to enable the Dynamic Cell Power

`SAMECVGCELLIDTYPE`: Index type of a co-coverage cell. If the coverage area of a cell is under the coverage area of another cell, the cell can be disabled during off-peak hours. In this situation, another cell is considered as the co-coverage cell of the cell.

`TURNOFFCELLSTRTIME`: Start time for dynamically disabling a cell.

`TURNOFFCELLSTPTIME`: End time for dynamically disabling a cell.

After executing script 3, PRB utilization still persists above 70% then we have no other option rather implement a long-term solution with node expansion.

B. IDENTIFYING QOS BREAKDOWN POINT FROM PRB UTILIZATION

The main objective of this sub-section is to identify the quality of service or user throughput (User_TP) breakdown point from the PRB utilization graph. This part is significant

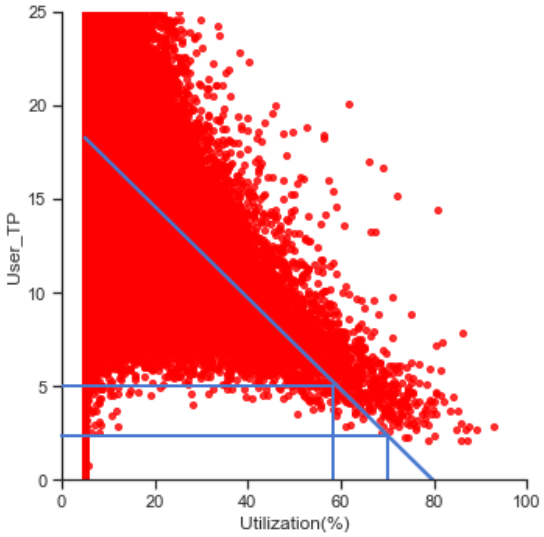


FIGURE 19. User Throughput (User_TP) vs. PRB Utilization Graph.

to point out in different networks; based on the identified threshold point network engineers can decide whether they will move for step-up or long-term solutions (Fig. 19). We have collected 890 eNodeB 120 days hourly PRB Utilization and User Throughput (User_TP) data for regression plot from a live operator network data. We found that when PRB Utilization hits 70% and above, user throughput degraded by 50% (from 5 Mbps to 2.5 Mbps). 50% degradation means severely impacted user experiences. Responsible engineers must immediately trigger step-up solutions to reduce PRB utilization and enhance user experience. Similarly, when PRB utilization goes above 80%, we found that almost zero user throughput means no QoS. This threshold point will vary network-wise. We have considered reference live from an LTE network configured in 1800 Mhz and per eNodeB frequency bandwidth 10 Mhz.

Limitations: Apart from that, the proposed model has a few limitations too. The model will perform well in any normal day scenario. However, the model's performance may slightly degrade during any social gathering, when the number of users of a certain place will increase massively compared to a typical day. As well as, while designing the algorithm for soft parameter tuning, we primarily focused on LTE capacity enhancement. Due to this dynamic resource-sharing technique from GSM to LTE, GSM networks may rarely face resource constraints. This issue can be solved through optimization techniques. In addition, computational power needs to enhance with respect to increasing the number of eNodeB.

VIII. CONCLUSION

This research innovatively devises a fusion model in a combination of three deep learning algorithms for the most granular level cellular network traffic prediction as a solution of NP-hard optimization of user throughput maximization.

In addition, a DTW-based self-organized map (SOM) makes cluster mapping of different eNodeB time series data easy. Besides that, considering the reference LTE network radio configuration, the QoS breakdown threshold point is also determined by correlating to the PRB Utilization graph, which is 70%.

The accuracy of the proposed model is increased by $6.6 - 7.0\%$ using the Fusion Strategy in RNN and maintains excellent R^2 score i.e., 0.8034, which represents a very precise prediction of network traffic volume. In the next level, a rigorous parameter estimation algorithm was proposed for triggering a dynamic capacity step-up solution two months in advance with optimized radio power allocation based on predicted LTE network traffic and PRB utilization. The proposed algorithm for network capacity optimization is another novelty of this research, that would help network engineers to plan and execute soft parameters before the quality of service (QoS) degrades compared to the benchmark. Thus, customers will be less sufferer from capacity expansion lead time from the MNO side.

One drawback of this model is that, the proposed deep learning model may require additional computational resources to predict traffic for a large number of eNodeB setups or public gathering events. Apart from that, lower technology (GSM) get less priority while designing the model to allocate more radio resources on demanding higher technology (LTE). Additional optimization effort (i.e., traffic shifting to neighbor cell) may easily resolve GSM rare case resource allocation issue (if any).

In the future, we will address the prediction of traffic peaks during social events of a particular geographical area or eNodeB serving area by adopting a Restricted Boltzmann Machines (RBM) with Conditional Random Fields (CRFs). In addition to that, we will also focus on smooth dynamic resource allocation in heterogeneous complex networks system, including GSM, LTE, 5G, and beyond technology of a particular period based on forecasted traffic and customer demand.

REFERENCES

[1] (Jan. 2022). *Ericsson Mobility Report*. [Online]. Available: https://www.ericsson.com/49d3a0/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-june-2022.pdf

[2] K. B. A. Delaporte. (Sep. 2021). *The State of Mobile Internet Connectivity 2021*. [Online]. Available: https://www.gsma.com/r/wp-content/uploads/2021/09/The-State-of-Mobile-Internet-Connectivity-Report-2021.pdf

[3] J. L. Bejarano-Luque, M. Toril, M. Fernandez-Navarro, C. Gijon, and S. Luna-Ramirez, "A deep-learning model for estimating the impact of social events on traffic demand on a cell basis," *IEEE Access*, vol. 9, pp. 71673–71686, 2021.

[4] W. Shen, H. Zhang, S. Guo, and C. Zhang, "Time-wise attention aided convolutional neural network for data-driven cellular traffic prediction," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1747–1751, Aug. 2021.

[5] 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.

[6] 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.

[7] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, "Data for development: The D4D challenge on mobile phone data," 2012, *arXiv:1210.0137*.

[8] L. Feng, W. Li, Y. Lin, L. Zhu, S. Guo, and Z. Zhen, "Joint computation offloading and URLLC resource allocation for collaborative MEC assisted cellular-V2X networks," *IEEE Access*, vol. 8, pp. 24914–24926, 2020.

[9] D. Babicz, A. Tihanyi, M. Koller, C. Rekeczky, and A. Horváth, "Simulation of an analogue circuit solving NP-hard optimization problems," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.

[10] Y. Fang, S. Ergut, and P. Patras, "SDGNet: A handover-aware spatiotemporal graph neural network for mobile traffic forecasting," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 582–586, Mar. 2022, doi: 10.1109/LCOMM.2022.3141238.

[11] 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. Serv. Comput.*, vol. 9, no. 5, pp. 796–805, Sep./Oct. 2016, doi: 10.1109/TSC.2016.2599878.

[12] A. Kirmaz, D. S. Michalopoulos, I. Balan, and W. Gerstacker, "Mobile network traffic forecasting using artificial neural networks," in *Proc. 28th Int. Symp. Modeling, Anal., Simulation Comput. Telecommun. Syst. (MASCOTS)*, Nov. 2020, pp. 1–7, doi: 10.1109/MASCOTS50786.2020.9285949.

[13] F. Sun, P. Wang, J. Zhao, N. Xu, J. Zeng, J. Tao, K. Song, C. Deng, J. C. S. Lui, and X. Guan, "Mobile data traffic prediction by exploiting time-evolving user mobility patterns," *IEEE Trans. Mobile Comput.*, vol. 21, no. 12, pp. 4456–4470, Dec. 2022, doi: 10.1109/TMC.2021.3079117.

[14] L. Lo Schiavo, M. Fiore, M. Gramaglia, A. Banchs, and X. Costa-Perez, "Forecasting for network management with joint statistical modelling and machine learning," in *Proc. IEEE 23rd Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2022, pp. 60–69, doi: 10.1109/WoWMoM54355.2022.00028.

[15] Q. Yu, H. Wang, T. Li, D. Jin, X. Wang, L. Zhu, J. Feng, and C. Deng, "Network traffic overload prediction with temporal graph attention convolutional networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2022, pp. 885–890, doi: 10.1109/ICCWorkshops53468.2022.9814643.

[16] H. S. Jang, H. Lee, H. Kwon, and S. Park, "Deep learning-based prediction of resource block usage rate for spectrum saturation diagnosis," *IEEE Access*, vol. 9, pp. 59703–59714, 2021, doi: 10.1109/ACCESS.2021.3073670.

[17] M. M. Hasan, S. Kwon, and J.-H. Na, "Adaptive mobility load balancing algorithm for LTE small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2205–2217, Apr. 2018, doi: 10.1109/TWC.2018.2789902.

[18] F. W. Alsaade and M. H. Al-Adhaileh, "Cellular traffic prediction based on an intelligent model," *Mobile Inf. Syst.*, vol. 2021, pp. 1–15, Jul. 2021.

[19] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Inf. Fusion*, vol. 42, pp. 146–157, Jul. 2018.

[20] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger, and M. Manic, "Explainable unsupervised machine learning for cyber-physical systems," *IEEE Access*, vol. 9, pp. 131824–131843, 2021.

[21] K. Li, K. Sward, H. Deng, J. Morrison, R. Habre, M. Franklin, Y.-Y. Chiang, J. L. Ambite, J. P. Wilson, and S. P. Eckel, "Using dynamic time warping self-organizing maps to characterize diurnal patterns in environmental exposures," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, Dec. 2021.

[22] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.

[23] L. Yang, Z. Ouyang, and Y. Shi, "A modified clustering method based on self-organizing maps and its applications," *Proc. Comput. Sci.*, vol. 9, pp. 1371–1379, Jan. 2012.

[24] H. R. Medeiros, F. D. B. de Oliveira, H. F. Bassani, and A. F. R. Araujo, "Dynamic topology and relevance learning SOM-based algorithm for image clustering tasks," *Comput. Vis. Image Understand.*, vol. 179, pp. 19–30, Feb. 2019.

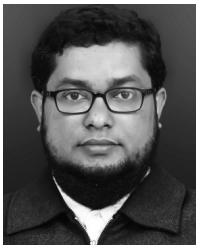
[25] H. Widiputra, A. Mailangkay, and E. Gautama, "Multivariate CNN-LSTM model for multiple parallel financial time-series prediction," *Complexity*, vol. 2021, pp. 1–14, Oct. 2021.

[26] H. S. Hota, R. Handa, and A. K. Shrivas, "Time series data prediction using sliding window based RBF neural network," *Int. J. Comput. Intell. Res.*, vol. 13, no. 5, pp. 1145–1156, 2017.

[27] X. Wang and D. Liang, "LSTM-based alarm prediction in the mobile communication network," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 561–567.
[28] X. Yuan, L. Li, and Y. Wang, "Nonlinear dynamic soft sensor modeling with supervised long short-term memory network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 3168–3176, May 2019.
[29] A. Yadav, C. K. Jha, and A. Sharan, "Optimizing LSTM for time series prediction in Indian stock market," *Proc. Comput. Sci.*, vol. 167, pp. 2091–2100, Jan. 2020.
[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[31] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 3285–3292.
[32] N. Li, L. Hu, Z.-L. Deng, T. Su, and J.-W. Liu, "Research on GRU neural network satellite traffic prediction based on transfer learning," *Wireless Pers. Commun.*, vol. 118, no. 1, pp. 815–827, May 2021.
[33] J. Yuan, H. Wang, C. Lin, D. Liu, and D. Yu, "A novel GRU-RNN network model for dynamic path planning of mobile robot," *IEEE Access*, vol. 7, pp. 15140–15151, 2019.
[34] N. Tavakoli, "Modeling genome data using bidirectional LSTM," in *Proc. IEEE 43rd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Jul. 2019, pp. 183–188.
[35] J. Kim and N. Moon, "BiLSTM model based on multivariate time series data in multiple field for forecasting trading area," *J. Ambient Intell. Hum. Comput.*, pp. 1–10, Jul. 2019. [Online]. Available: https://link.springer.com/article/10.1007/s12652-019-01398-9#citeas
[36] S. Lathuiliere, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2065–2081, Sep. 2020.
[37] A. Masood, T.-V. Nguyen, and S. Cho, "Deep regression model for videos popularity prediction in mobile edge caching networks," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2021, pp. 291–294.
[38] D. Rügamer, C. Kolb, C. Fritz, F. Pfisterer, P. Kopper, B. Bischl, R. Shen, C. Bukas, L. B. D. A. E. Sousa, D. Thalmeier, P. Baumann, L. Kook, N. Klein, and C. L. Müller, "Deepregression: A flexible neural network framework for semi-structured deep distributional regression," 2021, *arXiv:2104.02705*.
[39] D. Chmieliauskas and D. Guršnys, "LTE cell traffic grow and congestion forecasting," in *Proc. Open Conf. Electr., Electron. Inf. Sci. (eStream)*, Apr. 2019, pp. 1–5.
[40] M.-C. Nguyen, H. Nguyen, D.-H. Nguyen, E. Georgeaux, P. Mege, and L. Martinod, "Adaptive physical resource block design for enhancing voice capacity over LTE network in PMR context," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–5.
[41] H. Al-Zayadi, O. Lavriv, M. Klymash, and A.-S. Mushtaq, "Increase throughput by expectation channel quality indicator," in *Proc. 1st Int. Sci.-Practical Conf. Problems Infocommun. Sci. Technol.*, Oct. 2014, pp. 120–121.
[42] I. Toyoda, F. Nuno, Y. Shimizu, and M. Umehira, "Proposal of 5/25-GHz dual band OFDM-based wireless LAN for high-capacity broadband communications," in *Proc. IEEE 16th Int. Symp. Pers., Indoor Mobile Radio Commun.*, vol. 3, Sep. 2005, pp. 2104–2108.
[43] M. B. Shahab, M. A. Wahla, and M. T. Mushtaq, "Downlink resource scheduling technique for maximized throughput with improved fairness and reduced BLER in LTE," in *Proc. 38th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2015, pp. 163–167.
[44] M. Caretti, M. Crozzoli, G. M. Dell'Aera, and A. Orlando, "Cell splitting based on active antennas: Performance assessment for LTE system," in *Proc. IEEE Wireless Microw. Technol. Conf. (WAMICON)*, Apr. 2012, pp. 1–5.
[45] J. Xiao, R. Q. Hu, Y. Qian, L. Gong, and B. Wang, "Expanding LTE network spectrum with cognitive radios: From concept to implementation," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 12–19, Apr. 2013.
[46] N. Morozs, T. Clarke, and D. Grace, "Intelligent secondary LTE spectrum sharing in high capacity cognitive cellular systems," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Sep. 2015, pp. 1–2.
[47] A. Mostafa, M. A. Elattar, and T. Ismail, "Downlink throughput prediction in LTE cellular networks using time series forecasting," in *Proc. Int. Conf. Broadband Commun. Next Gener. Netw. Multimedia Appl. (CoBCom)*, Jul. 2022, pp. 1–4, doi: 10.1109/CoBCom55489.2022.9880654.
[48] *Huawei eRAN Channel State Management Feature Parameter (Feature ID: LBFD-00101501 CQI Adjustment)*, Huawei, Shenzhen, China, 2020.
[49] *Huawei Power Control Feature Parameter (Feature ID: LBFD- 002016 Dynamic Downlink Power Allocation)*, Huawei, Shenzhen, China, 2020.

SYED TAUHIDUN NABI (Graduate Student Member, IEEE) received the Bachelor of Science (B.Sc.) degree in electrical and electronic engineering from American International University–Bangladesh (AIUB), and the Master of Science (M.Sc) in computer science and engineering from BRAC University, Bangladesh. He is currently pursuing the Ph.D. degree in computer science with Virginia Tech (formally, Virginia Polytechnic Institute and State University), USA. He has accumulated substantial industry experience as a Network Planning and Deployment Engineer, having worked for eight years at two major Mobile Network Operators (MNO) in Bangladesh. His research interests span information and communication technologies for development (ICTD), mobile network architecture, machine learning in networking, and human-computer interaction (HCI).

MD. RASHIDUL ISLAM received the Bachelor of Science (B.Sc.) degree in information and communication engineering from East West University, in 2011. He is currently pursuing the Master of Science (M.Sc.) degree in computer science and engineering (CSE) with BRAC University. He is an experienced telecom professional with expertise in different vendors' cellular systems for more than ten years. He has expertise in radio network planning, solution design and optimization, also have experience in predictive coverage and capacity analysis for LTE and network big data analysis. His research interests include mobile network capacity management, traffic modeling, artificial intelligence, and machine learning.

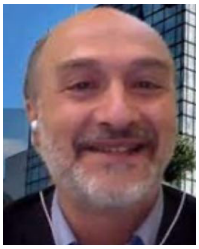


MD. GOLAM RABIUL ALAM (Member, IEEE) received the B.S. degree in computer science and engineering and the M.S. degree in information technology, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2017. He also worked as a Postdoctoral Researcher with the Computer Science and Engineering Department, Kyung Hee University, from March 2017 to February 2018. He is currently a Full Professor of computer science and engineering with the Department of Computer Science and Engineering, BRAC University, Bangladesh. His research interests include healthcare informatics, mobile cloud and edge computing, ambient intelligence, and persuasive technology. He is a member of the IEEE IES, CES, CS, SPS, CIS, KIISE, and IEEE ComSoc. He received several best paper awards at prestigious conferences.

MOHAMMAD MEHEDI HASSAN (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in February 2011. He is currently a Full Professor with the Department of Information Systems, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia. He has authored or coauthored around more than 180 publications, including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, books, and book chapters. Recently, his four publications have been recognized as the ESI Highly Cited Papers. His research interests include cloud computing, edge computing, the Internet of Things, body sensor networks, big data, deep learning, mobile cloud, smart computing, wireless sensor networks, 5G networks, and social networks. He has served as the chair and a technical program committee member for international conferences.

SALMAN A. ALQAHTANI (Member, IEEE) is currently a Professor with the Department of Computer Engineering, King Saud University, Riyadh. His current research interests include 5G networks, broadband wireless communications, radio resource management for 4G and beyond networks (call admission control, packet scheduling and radio resource sharing techniques), cognitive and cooperative wireless networking, small cell and heterogeneous networks, self-organizing networks, SDN/NFV, 5G network slicing, smart grid, intelligent IoT solutions for smart cities, dynamic spectrum access, co-existence issues on heterogeneous networks in 5G, industry 4.0 issues, the Internet of Everything, mobile edge and fog computing, and cyber sovereignty. In addition, his interests also include performance evaluation and analysis of high speed packet switched networks, system model and simulations, and integration of heterogeneous wireless networks. His main focus is on the design and optimization of 5G MAC layers, closed-form mathematical performance analysis, energy-efficiency, and resource allocation and sharing strategies. He has authored two scientific books and authored/coauthored around 76 journals and conference papers in the topic of his research interests. He serves as a reviewer for several national and international journals.

GIANLUCA ALOI (Member, IEEE) received the Ph.D. degree in systems engineering and computer science from the DEIS Department, University of Calabria, in 2003. In 2004, he joined the University of Calabria, where he is currently an Assistant Professor of telecommunications with the Department of Informatics, Modeling, Electronics and System Engineering. His main research interests include spontaneous and reconfigurable wireless networks, cognitive and opportunistic networks, sensor and self-organizing wireless networks, and the Internet of Things technologies.



GIANCARLO FORTINO (Fellow, IEEE) received the Ph.D. degree in systems and computer engineering from the University of Calabria (Unical), Italy, in 2000. He is currently a Full Professor of computer engineering at the Department of Informatics, Modeling, Electronics, and Systems, University of Calabria (Unical). He is also Guest Professor at the Wuhan University of Technology, Wuhan, China; an High-End Expert at HUST, China; and a Senior Research Fellow at the Italian National Research Council ICAR Institute. He is the Director of the SPEME Laboratory, Unical, as well as the Co-Chair of joint laboratories on the IoT established between Unical and WUT and SMU Chinese universities, respectively. He is the Co-Founder and the CEO of SenSysCal S.r.l., a Unical spinoff focused on innovative IoT systems, and cofounder of BigTech S.r.l., a startup focused on AI-driven systems and Big Data. He is Highly Cited Researcher 2002 to 2022 by Clarivate in Computer Science. He is the author of over 600 papers in international journals, conferences, and books. His research interests include wearable computing systems, the Internet of Things, and Cyber-security. He is also a member of the IEEE SMCS BoG and co-chair of the SMCS TC on IWCD. He is the Chair of the IEEE SMCS Italian Chapter. He is the (Founding) Series Editor of IEEE Press Book Series on Human-Machine Systems and the Springer Internet of Things series, and he is AE of premier IEEE TRANSACTIONS, such as IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, IEEE SYSTEMS JOURNAL, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, *Information Fusion*, *JNCA*, and *EAAI*.

Open Access funding provided by 'Università della Calabria' within the CRUI CARE Agreement