2023 International Balkan Conference on Communications and Networking (BalkanCom) | 979-8-3503-3910-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/BalkanCom58402.2023.10167893

Multivariate and Multi-step Traffic Prediction for NextG Networks with SLA Violation Constraints

Evren Tuna[1,2] and Alkan Soysal[2,3]

[1]Department of Research & Development, ULAK Communications Inc., Ankara 06510, Turkey
[2]Department of Electrical and Electronics Engineering, Bahcesehir University, Istanbul 34353, Turkey
[3]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061, USA

***Abstract*—This paper focuses on predicting downlink (DL) traffic volume in mobile networks while minimizing overprovisioning and meeting a given service-level agreement (SLA) violation rate. We present a multivariate, multi-step, and SLA-driven approach that incorporates 20 different radio access network (RAN) features, a custom feature set based on peak traffic hours, and handover-based clustering to leverage the spatiotemporal effects. In addition, we propose a custom loss function that ensures the SLA violation rate constraint is satisfied while minimizing overprovisioning. We also perform multi-step prediction up to 24 steps ahead and evaluate performance under both single-step and multi-step prediction conditions. Our study makes several contributions, including the analysis of RAN features, the custom feature set design, a custom loss function, and a parametric method to satisfy SLA constraints.**

I. Introduction

Mobile networks are designed to support various vertical applications, each with unique requirements. To efficiently manage these networks, operators must deploy appropriate infrastructure and enhance their predictive capabilities using machine learning (ML) and an SLA-driven orchestration. This allows the networks to adapt dynamically to changing traffic demands. Therefore, it is crucial to develop a well-designed method for predicting cell-based traffic volume, which can be used to manage the allocation of network slices accurately.

The network traffic data in mobile networks depends on space and time due to interactions between base stations (BSs) and daily/weekly trends in mobile data usage. Several studies, including [1]–[5], have proposed using Convolutional Neural Network (CNN)-based models to leverage spatial dependencies. Among these studies, [1]–[3] have utilized a 3D-CNN structure borrowed from video processing applications. They assume that the inputs at a given time are in a matrix format, where each entry represents the aggregated traffic of BSs in a corresponding square grid area. On the other hand, [4], [5] have proposed using graph convolutional networks (GCNs) instead of a grid structure. In [4], the authors have combined Recurrent Neural Networks (RNNs) with GCNs for multi-step prediction. However, their proposed model performs worse than the vanilla LSTM for the special case of single-step prediction and about the same as ARIMA. In addition to GCNs, [5] has proposed using handover data to improve performance. The results in terms of Mean Square Error (MSE) and Mean Absolute Error (MAE) reported in [5] are 10-15% better than those of the vanilla LSTM results.

Several studies (e.g., [6]–[11]) have proposed RNN-based models that utilize the temporal dependency of network traffic data. However, in [9], [10], where the goal is to optimize resource allocation using predicted traffic, evaluating the performance of network traffic prediction is challenging. Meanwhile, [6], [7] have also utilized a grid structure similar to [1]–[3], while [8] has used a private dataset and cell clustering based on the similarity of time-series trends. In a previous work [11], we proposed RNN-based models that used handover data for single-step prediction. However, [11] does not guarantee an SLA violation rate.

Most literature uses standard loss functions such as MAE, MSE, and their variants to minimize prediction error [2]–[10]. These symmetrical loss functions treat SLA violation and overprovisioning the same and result in nearly identical SLA violation and overprovisioning rates. However, operators tend to favor overprovisioning to SLA violation due to the strict obligations of SLAs. Nearly 50% SLA violation resulting from the use of standard loss functions is unacceptable for operators. Therefore, designing and using application-specific and parameterized custom loss functions are crucial.

In this paper, our main goal is to predict DL traffic volume while minimizing overprovisioning and meeting a given SLA violation rate. Our approach is multivariate, multi-step, and SLA-driven, and we make the following contributions:

- We analyze 19 different RAN features, in addition to DL traffic volume, to assess their effect on prediction.
- We design a custom feature set based on peak traffic hours.
- We extract the spatiotemporal effect of high mobility using incoming and outgoing handover relationships between cells.
- We perform multi-step prediction up to 24 steps ahead.
- We propose a custom loss function that guarantees the SLA violation rate while minimizing overprovisioning.
- We evaluate our approach's performance under single-step and multi-step prediction scenarios.

II. Dataset and Methodology

This section discusses the dataset and methodology used in this study, which focuses on predicting DL traffic volume in

979-8-3503-3910-9/23/$31.00 ©2023 IEEE

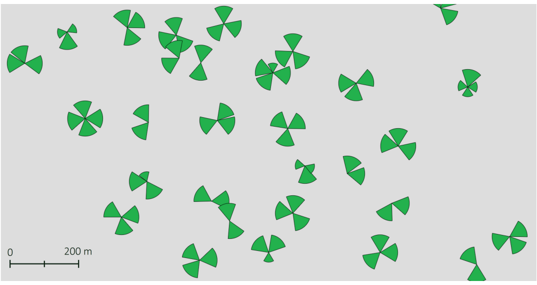

Fig. 1. Cells in a metropolitan area.

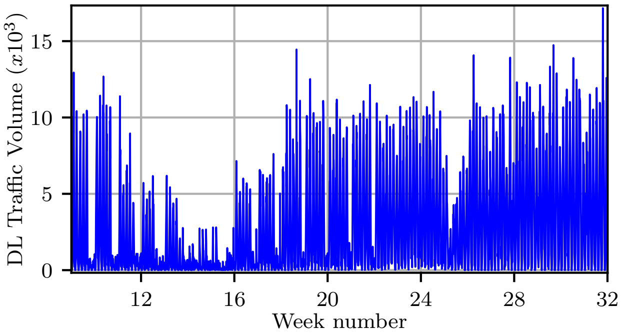

Fig. 2. Downlink traffic variation of the GU14 cell.

a live LTE RAN serving a highly dense urban area with high user mobility. Multiple BSs cover the area, each with varying numbers of sectors and carriers (see Fig. 1). Throughout this study, we denote a BS with a two-letter name such as GU, VO, etc. The cells within the relevant BS are represented using two digits, the first being the sector and the second being the carrier number. For example, GU14 refers to the first sector and fourth carrier of the GU base station. Therefore, a carrier in a sector of a BS is referred to as GU14, VO13, etc.

To improve the network's reliability and prevent SLA violations, the goal is to develop an ML-based prediction method that considers unexpected traffic spikes during peak traffic hours. For this purpose, this study uses a multivariate LSTM architecture, which has a higher learning capacity and overcomes the long-term dependency limitation of traditional RNN models.

The dataset contains hourly raw data measurements over 52 weeks. In Fig. 2, we show the DL traffic volume graph of the GU14 cell, serving a crowded city square. This cell has a higher traffic load than others, and the trend of the service demand changes throughout the measurement period. Thus, demand spikes occur unexpectedly.

Table I lists collected features with their measurement family name [12], [13]. The dataset used in this study contains 20 RAN features, with DL traffic volume (labeled F10) as the output feature to be predicted. We split the dataset into training, validation, and test sets with lengths of 40 weeks, 8 weeks, and 4 weeks, respectively. We normalize the feature sets using mean and standard deviation values calculated from the training set. Finally, we tune hyperparameters such as learning rate, epoch number, L2 regularization penalty, LSTM or layer number, and hidden unit number using a grid search.

TABLE I
RAN Features

Label	Name
F1	Num. of Initial E-RABs Attempted to Setup
F2	RACH Setup Succ. Rate
F3	Avg. RACH Timing Advance
F4	Num. of RRC Attempts
F5	Num. of S1 Signalling Establishment Attempt
F6	DL PDCP Cell Thr.
F7	UL PDCP Cell Thr.
F8	DL PDCP User Thr.
F9	UL PDCP User Thr.
F10	DL Traffic Volume
F11	UL Traffic Volume
F12	Avg. UL RSSI Weight PUCCH
F13	Avg. UL RSRP PUSCH
F14	Avg. UL RSRP PUCCH
F15	Avg. CQI
F16	Avg. Num. of Active Users in DL
F17	Avg. Num. of Active Users in UL
F18	Num. of Avg. Simultaneous RRC Connected Users
F19	DL PRB Utilisation
F20	UL PRB Utilisation

While training the model, we use the K-fold cross-validation technique to mitigate the changing traffic levels and trends. The number of cross-validation folds, K, is set to six, dividing the 12-month dataset into 2-month intervals. Each fold has different training and validation sets that are obtained by shifting the training dataset by 2 months.

III. Parametric Traffic Prediction

The core of our approach is a multivariate and multi-step prediction model for traffic volume, which utilizes an SLA-based weighted loss function. In this section, we will present our proposed weighted loss function, as well as our multivariate feature design and multi-step prediction structure.

A. Weighted Loss Function

We use x_t to denote the traffic volume measured at time t and $\hat{x}_t$ to denote its corresponding prediction. During test time, we consider SLA violations to occur when the prediction error is negative, i.e., $\hat{x}_t - x_t < 0$, and overprovisioning to occur when the prediction error is positive, i.e., $\hat{x}_t - x_t > 0$. The SLA violation rate is defined as the percentage of instances where an SLA is violated during test time. We also define overprovisioning volume as the average positive prediction error.

Standard loss functions such as MAE and MSE suffer from two significant disadvantages. Firstly, they impose the same penalty for both SLA violations and overprovisioning. Secondly, they cannot be parameterized with an SLA violation rate. Therefore, we propose a custom loss function that overcomes both of these disadvantages of standard loss functions.

The weighted Mean Absolute Error (wMAE) loss function can be written as

$$\mathcal{L}_{\mathrm{wMAE}}(\hat{x}-x, w) = \begin{cases} w|\hat{x}-x|, & \hat{x}-x \leq 0 \\ \hat{x}-x, & \hat{x}-x > 0 \end{cases} \quad (1)$$

where w is the weight parameter of the function. We aim to minimize the overprovisioning volume while adhering to a constraint on the SLA violation rate by selecting an appropriate w. To this end, we explore two different SLA violation rate scenarios, 3% and 5%, and employ a univariate LSTM model. For each of these scenarios, we perform a line search to determine the optimal weights, denoted as $w_{3\%}$, and $w_{5\%}$.

B. Multivariate Features

The aim of our research is to improve the performance of our model by including additional input features beyond the DL traffic volume. Firstly, we investigate the correlation between the DL traffic volume and the 19 other features monitored in the RAN. We expect that including highly correlated features in the input dataset will have a positive impact on the model's performance. Next, we introduce a custom feature set that emphasizes specific time periods during the day or specific days during the week. Finally, to address the spatiotemporal effect, we propose a feature clustering method, which incorporates the incoming and outgoing handover relationships.

To compare the performance of our multivariate model, we use the univariate LSTM model as a baseline. In the univariate model, the input consists of an array of the past 24 hours of DL traffic volume, and the output is the next instance of DL traffic volume. In contrast, in the multivariate model, the input consists of multiple arrays, each containing the past 24 hours of a feature.

1) RAN Features: We investigate the usefulness of additional input features beyond the DL traffic volume for predicting cell-based multivariate traffic volume. We consider 20 different RAN features, and their labels are presented in Table I. To identify the most correlated features, we calculate the Pearson correlation coefficient between the DL traffic volume feature (F10) and the other features. Fig. 3 shows an example of the correlation heatmap for the GU14 cell. We set a correlation threshold of 0.90 and evaluate the heatmaps of various cells to determine which features can be included in the input dataset. Based on our analysis, we find that the F16, F17, F18, and F19 features are highly correlated with F10 and should be included in the input dataset of our multivariate model. We denote this model as mvLSTM-RAN, which utilizes features F16 through F19 along with F10 to predict the future values of F10.

2) Peak Features: We aim to improve our model's performance by addressing the issue of abnormal peaks at busy hours in the F10 feature values. We observe that the F10 feature has a 24-hour period for all cells, with intervals of low and high demand for traffic volume. Moreover, F10 values tend to be higher during weekdays and lower during weekends in densely populated metropolitan areas. To capture these patterns, we introduce a custom Boolean feature vector, "peak days of the week," to differentiate between weekdays and weekends in the input dataset.

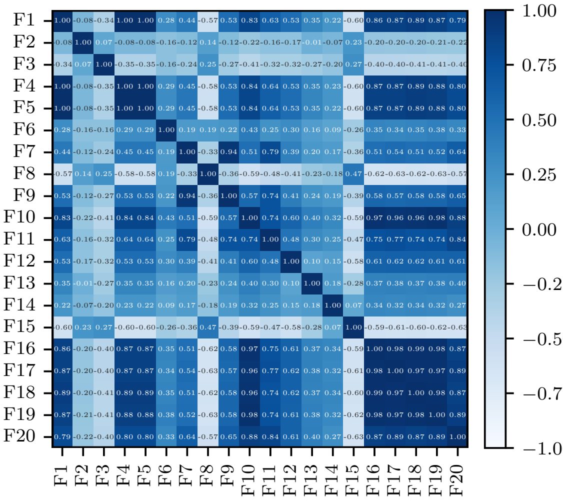

Fig. 3. Correlation heatmap of RAN features for GU14 cell.

However, our initial model fails to accurately predict F10 values during periods of abnormal peaks at busy hours. To address this, we develop a method to identify peak hours for each cell based on daily 24-hour periods and record the time the F10 value is at its maximum for each period. Using occurrence statistics, we label hours with occurrence values higher than a defined threshold as peak hours. This leads us to propose an additional custom Boolean feature vector, "peak hours of the day" to differentiate between peak hours from non-peak hours.

Combining both Boolean features, we introduce our next multivariate model, mvLSTM-peak, which incorporates the peak days of the week and peak hours of the day vectors in addition to F10. Notably, these additional features are derived from the statistics of F10 and do not require any further RAN measurements.

3) Handover Clustering Features: In this section, we investigate the impact of handovers on predicting traffic volume in different cells of the RAN. In densely populated areas with highly mobile users, there are dynamic changes in traffic volume demand as users move between cells.

One approach to exploit the correlation in the handover process is to add the features of all cells in the region to the input dataset to predict the traffic volume feature of the GU14 cell. However, this method significantly increases the problem's complexity and may reduce the model's performance. Another approach is to consider the traffic volume data of only a cluster of cells around the GU14 cell. However, we found that selecting cells in the cluster based only on proximity and/or intersection of coverage does not improve prediction performance, which highlights the limitations of the grid structure of CNN-based approaches.

We introduce the *Handover Clustering* method to exploit

TABLE II
THE RATES OF INCOMING AND OUTGOING HANDOVERS OF GU14

Incoming		**Outgoing**	
Cells	**Rate %**	**Cells**	**Rate %**
GU12	66.79	GU12	26.86
MS34	6.08	SY24	17.24
VO14	5.69	VO14	12.48
SY24	4.68	GU17	8.72
VO12	4.45	MS34	8.31
GU24	3.36	GU24	4.54
MS37	2.05	GU13	3.96
SY22	1.99	VO12	1.88
GU13	1.49	VO13	1.88
GU22	1.15	RE37	1.55

the correlation in cell-based incoming and outgoing handover statistics. This method considers only those cells that have a handover relationship with the target cell GU14. Table II presents the neighboring cells that have incoming and outgoing handover relations with GU14. Our analysis shows that 66.79% of incoming handover data to GU14 is from the GU12 cell. With Handover Clustering, we expand the input dataset by adding two feature vectors. These vectors are constructed by calculating the weighted averages of the DL traffic volumes of the neighboring cells, one vector for the incoming and the other for the outgoing handover relationship. We refer to our next multivariate model, mvLSTM-handover, which includes the two custom feature vectors formed through the Handover Clustering method.

In summary, we present four multivariate LSTM models for predicting traffic volume in the RAN. The first model, mvLSTM-RAN, incorporates additional RAN features, while the second model, mvLSTM-peak, focuses on the peak hours of the day and days of the week. The third model, mvLSTM-handover, incorporates the weighted averages of DL traffic volumes of cells in the handover cluster. Lastly, we propose the mvLSTM-all model, which combines all three multivariate models.

C. Multi-step Prediction

In this subsection, we examine the performance of our multivariate LSTM models in a multi-step prediction scenario. Specifically, we consider 1-hour, 2-hour, 4-hour, 8-hour, and 24-hour ahead predictions and analyze the performance degradation as we increase the prediction horizon.

In a multi-step prediction scenario, previously predicted values are used as valid past values for subsequent predictions. Therefore, the accuracy of the initial prediction is crucial in determining the performance of the subsequent predictions. To evaluate the performance degradation, we compare the results of the multi-step prediction with those of the single-step prediction.

We perform the multi-step prediction using all four of our proposed models: mvLSTM-RAN, mvLSTM-peak, mvLSTM-handover, and mvLSTM-all. For each model, we conduct experiments on the same test dataset for 1-hour, 2-hour, 4-hour, 8-hour and 24-hour ahead predictions.

TABLE III
1-HOUR AHEAD PREDICTION PERFORMANCE FOR THE GU14 CELL

Models	**3%**		**5%**	
	Loss	**Volume**	**Loss**	**Volume**
univariate LSTM	0.50	42.61	0.44	36.92
mvLSTM-RAN	0.49	42.74	0.43	37.65
mvLSTM-peak	0.46	39.23	0.42	34.75
mvLSTM-handover	**0.44**	**38.08**	**0.39**	**31.28**
mvLSTM-all	0.48	39.55	0.41	33.23

IV. RESULTS

In this section, we provide the results of our DL traffic volume prediction for the GU14 cell. Initially, we demonstrate the change in the SLA-based loss and the overprovisioning volume required for 3% and 5% SLAs. We then compare the performance outcomes of single-step and multi-step predictions.

A. Different SLA Violation Rate Percentages

In this section, we present the performance results of DL traffic volume prediction for the GU14 cell under 3% and 5% SLAs. Table III presents the single-step performance results for the GU14 cell under both SLA conditions.

The results show that the test loss and accompanying overprovisioning volume of the models decrease as the SLA percentage increases. The weight for 3% SLA is higher than 5%, causing the loss function to penalize SLA violation cases more severely under the 3% SLA condition. Consequently, the model avoids violating SLA more, leading to increased test loss and overprovisioning volume due to frequent overestimation of the actual values.

For both SLA conditions, the mvLSTM-handover model performs better than the other models. Specifically, the mvLSTM-handover model yields 12% and 11.36% lower test losses than univariate LSTM for 3% and 5% SLA conditions, respectively.

However, the use of additional RAN features in mvLSTM-RAN and mvLSTM-all models resulted in more test loss and overprovisioning compared to mvLSTM-peak and mvLSTM-handover models. Fig. 4 illustrates that mvLSTM-RAN and mvLSTM-all models overestimate more than mvLSTM-peak and mvLSTM-handover models during the late afternoon hours for 3% SLA. This leads to an increase in the test loss and overprovisioning volume for mvLSTM-RAN and mvLSTM-all models more than mvLSTM-peak and mvLSTM-handover models.

B. Multi-step Prediction

In this section, we present the results of the multi-step prediction for the GU14 cell under the 5% SLA condition. Table IV shows the multi-step performance results for the GU14 cell. We observe that the accuracy of the initial prediction significantly affects the subsequent predictions, and the performance degrades as we increase the prediction horizon. As in the 1-hour ahead prediction, the mvLSTM-handover model has the lowest test loss and overprovisioning volume in

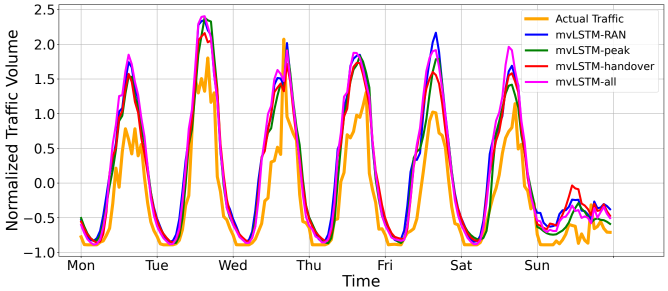

Fig. 4. Actual and predicted values for GU14 cell under 3% SLA.

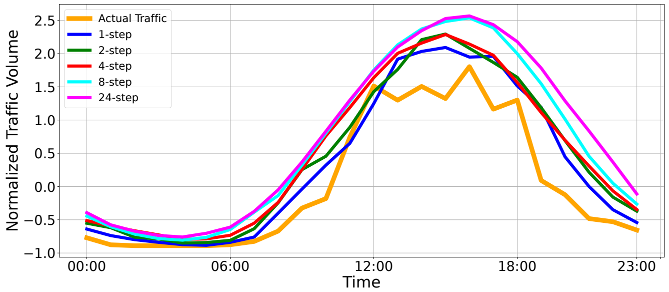

Fig. 5. Multi-step prediction using the mvLSTM-handover model.

the 2-hour ahead prediction. As the number of prediction steps increases, the performance of all models degrades, resulting in SLA violation when traffic demand increases and overprovisioning when it decreases. The test loss for the mvLSTM-handover model's 2-hour ahead prediction is 18% higher than that of the 1-hour ahead prediction. In multi-step prediction, increasing the number of steps leads to the model doing more overprovisioning during periods of increasing and decreasing traffic demand. Specifically, during the evening hours, when the traffic demand begins to decrease, we observe a lag of up to 3 hours in 24-hour predictions. The time graph in Fig. 5 illustrates the delay in the model predictions as the number of steps increases.

V. Conclusion

In this paper, we present four multivariate LSTM models for predicting DL traffic volume for a particular cell. We compare the performance of these models under two SLA conditions, 3% and 5%, for both single-step and multi-step predictions.

Our results show that the mvLSTM-handover model outperformed the other models under both SLA conditions for single-step prediction. On the other hand, the use of additional RAN features in mvLSTM-RAN and mvLSTM-all models do not provide better performance compared to mvLSTM-peak and mvLSTM-handover models.

In the multi-step prediction scenario, the performance of all models degraded as the number of steps increased. Similar to the 1-hour ahead prediction, we observe that the mvLSTM-handover model yields the lowest test loss value compared to other models in the 2-hour ahead prediction. However, when the number of steps is 4 or more, only considering the handover relationship makes it difficult to capture the changes in the neighbor cells of GU14.

In conclusion, our study demonstrates the effectiveness of multivariate LSTM models for DL traffic volume prediction in cellular networks. Furthermore, our findings suggest that incorporating handover clustering with weighted averages of DL traffic volumes of neighboring cells can significantly improve the prediction performance under different SLA conditions. The study recommends that operators set thresholds for degradation compared to single-step prediction and adjust their time horizon accordingly.

TABLE IV
Multi-step prediction test loss results for GU14 cell

Models	1-hour	2-hour	4-hour	8-hour	24-hour
univariate LSTM	0.44	0.61	0.70	0.87	0.91
mvLSTM-RAN	0.43	0.49	0.56	0.77	1.00
mvLSTM-peak	0.42	0.66	**0.48**	**0.65**	**0.82**
mvLSTM-handover	**0.39**	**0.46**	0.57	0.89	1.07
mvLSTM-all	0.41	0.54	0.59	0.69	0.95

References

[1] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Pérez, "Deepcog: Optimizing resource provisioning in network slicing with AI-based capacity forecasting," *IEEE Journal on Selected Areas in Communications*, vol. 38, pp. 361–376, February 2020.

[2] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Communications Letters*, vol. 22, pp. 1656–1659, August 2018.

[3] C. Zhang and P. Patras, "Long-term mobile traffic forecasting using deep spatio-temporal neural networks," in *ACM Mobicom*, June 2018, p. 231–240.

[4] L. Fang, X. Cheng, H. Wang, and L. Yang, "Mobile demand forecasting via deep graph-sequence spatiotemporal modeling in cellular networks," *IEEE Internet of Things Journal*, vol. 5, pp. 3091–3101, August 2018.

[5] S. Zhao, X. Jiang, G. Jacobson, R. Jana, W.-L. Hsu, R. Rustamov, M. Talasila, S. A. Aftab, Y. Chen, and C. Borcea, "Cellular network traffic prediction incorporating handover: A graph convolutional approach," in *IEEE SECON*, Como, Italy, June 2020.

[6] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *IEEE INFOCOM*, Atlanta, GA, USA, May 2017.

[7] L. Chen, D. Yang, D. Zhang, C. Wang, J. Li, and T. M. T. Nguyen, "Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization," *Journal of Network and Computer Applications*, vol. 121, pp. 59–69, 2018.

[8] J. Feng, X. Chen, R. Gao, M. Zeng, and Y. Li, "DeepTP: An end-to-end neural network for mobile cellular traffic prediction," *IEEE Network*, vol. 32, pp. 108–115, November 2018.

[9] I. Alawe, A. Ksentini, Y. Hadjadj-Aoul, and P. Bertin, "Improving traffic forecasting for 5G core network scalability: A machine learning approach," *IEEE Network*, vol. 32, pp. 42–49, November 2018.

[10] H. Chergui and C. Verikoukis, "Offline SLA-constrained deep learning for 5G networks reliable and dynamic end-to-end slicing," *IEEE Journal on Selected Areas in Communications*, vol. 38, pp. 350–360, February 2020.

[11] E. Tuna and A. Soysal, "Multivariate Spatio-temporal Cellular Traffic Prediction with Handover Based Clustering," in *CISS*, Mar. 2022.

[12] "LTE; Telecommunication management; Performance Management (PM); Performance measurements E-UTRAN," ETSI, Technical Specification (TS) 132.425, 11 2020, version 16.5.0.

[13] "GSM; UMTS; LTE; Telecommunication management; Performance Management (PM); Concept and requirements," ETSI, Technical Specification (TS) 132.401, 08 2020, version 16.0.0.