Int. j. inf. tecnol. (February 2024) 16(2):651–657
https://doi.org/10.1007/s41870-023-01678-w

ORIGINAL RESEARCH

A machine learning framework for predicting downlink throughput in 4G-LTE/5G cellular networks

Abbas Al-Thaedan[1] · Zaenab Shakir[1] · Ahmed Yaseen Mjhool[2] · Ruaa Alsabah[3] · Ali Al-Sabbagh[4,5] · Fitzroy Nembhard[6] · Monera Salah[7]

Received: 25 August 2023 / Accepted: 28 November 2023 / Published online: 2 January 2024

Abstract The current and next generations of cellular networks produce a massive amount of data. With this vast parameter increase, cellular communication networks have grown incredibly complicated. In addition, these cellular networks are unmanaged with conventional techniques, and a more advanced design and optimization methodology that depends on Machine Learning (ML) models is necessary. This work proposes a framework model for predicting downlink throughput (DL-Throughput) using ML models in fourth and fifth generations (4G/5G) cellular networks. The important parameters are selected from data measurements based on the correlation coefficient. The critical and effective parameters such as Reference Signal Received Power (RSRP), Signal to Interference and Noise Ratio(SINR), Received Signal Strength Indicator (RSSI), and Reference Signal Receive Quality (RSRQ) have been applied for the training model to predict the DL-Throughput in cellular networks. The prediction accuracy of the determination coefficient ranges between 89% and 96% from three different operators.

Keywords Machine learning · 4G-LTE · 5G · Framework · Correlation

1 Introduction

With the launch of 4G and the development of the 5G for mobile cellular technology, mobile broadband usage has increased rapidly with an increase of subscribers. As a result, it is crucial to preserve Quality of Service (QoS), which guarantees steady networks and high data rates. Thus, the throughput in wireless cellular systems have increased drastically, starting with the second generation (2G) to the current 5G and beyond. The need for a high and reliable data rate comes from the applications used by subscribers, including audio, video streaming, video games, and social media. In addition, mobile cellular operators sought to improve the QoS, which increased the demand for higher data rates and is anticipated to reach 77.5 exabytes monthly in 2023 [1–4].

Cellular network operators have complete priority in choosing the serving cell in LTE networks that need spectrum growth to prevent network bottlenecks. Furthermore, to prevent the cost of adding additional infrastructure hardware that is not fully utilized, such as new sectors, increasing the number of Multiple Input Multiple Output (MIMO) antennas or utilizing spectrum sharing solutions that could impact the user Key Performance Indicator (KPI) with inappropriate planning has been proposed. Therefore, cellular operators consider DL-Throughput one of the most crucial key performance parameters for LTE network planning [5–8].

✉ Abbas Al-Thaedan
abbas.khlaf@mu.edu.iq

[1] Al-Muthanna University, Samawah, Iraq

[2] Information Technology Research and Development Center (ITRDC), University of Kufa, Najaf, Iraq

[3] Information Technology Department, College of Information Technology and Computer Sciences, University of Kerbala, Kerbala, Iraq

[4] Information Technology Department, Al-Taff University College, Kerbala, Iraq

[5] Ministry of Communication, ITPC, Kerbala, Iraq

[6] L3Harris Institute for Assured Information, Florida Institute of Technology, Melbourne, FL, USA

[7] TeleWorld Solutions, Chantilly, VA, USA

Utilizing the DL-Throughput is essential for network dimensioning. However, the measured DL-Throughput vary from the real throughput because of fluctuation due to the mobility of the mobile device. Therefore, DL-Throughput prediction is based on network performance and does not consider each base station's distinctive qualities. Nowadays, cellular networks contain too many parameters that are not easy for humans to manage by conventional techniques. Even the most skilled engineers are unable to effectively process all the data produced within a network. Therefore, ML models are increasingly utilized with big data technology to develop and optimize cellular networks. One obvious ML use case is modeling DL-Throughput since measurement data throughput has been the key metric recently [9–11].

This work models and predicts the DL-Throughput by utilizing LTE data measurements to train supervised ML models. The quality of DL-Throughput is a critical element for video streaming and gaming. A proprietary dataset is utilized to train and test the model. Feature selection is also applied by choosing the most significant parameters and removing the weak and redundant ones, resulting in an improved model.

The rest of the paper is organized as follows: Sect. 2 presents a background, review and analysis of the state of art methods followed by the methods and materials in Sect. 3. Section 4 reports the results and discussions. The paper concludes in Sect. 5.

2 Related work

Cellular network operators consider throughput the most crucial parameter when assessing network quality in recent technologies, especially with 4G, 5G, and beyond. Consequently, throughput modeling is crucial in cellular network deployment and optimization. The throughput modeling has been done through various approaches; the throughput data can be gathered and predicated utilizing extrapolation of the previous behavior based on various factors. Also, throughput is modeled analytically using calculation-based approaches as a function of numerous variables. In addition, various ML models have been used for throughput prediction.

Many previous studies related to throughput prediction, such as the authors in [12], investigated the accuracy of cellular bandwidth in LTE networks by developing a framework based on ML models in the US called “LinkForecasting". It utilized the history of throughput and other information for realistic prediction. On the other hand, the researchers at [13] develop time-series and ML models that employ neural networks and various linear regression techniques to forecast DL-Throughput hourly. In addition, the authors in [14] applied the Recurrent Neural Networks (RNN) to calculate DL-Throughput, and this approach reduced the prediction error by 29% compared to conventional prediction approaches. In [15], the researchers utilized an improved RNN method called Long Short Term Memory (LSTM) to forecast DL-Throughput in LTE networks. It is utilized with time-sensitive services, such as video streaming, to decrease the delay. Also, the authors in [16] investigated the throughput prediction via time series modeling via ARIMA and exponential smoothing models in the LTE network. They utilized the models in the area to prove that the ARIMA model works better on weekdays. On the other hand, the exponential smoothing model proved better on predictions based on weekends.

Additionally, in [17], the authors studied the DL-Throughput prediction using ML models and deep learning to obtain better accuracy than traditional statistical models. They applied the seasonal ARIMA model and RNN to enhance the accuracy of RMSE for the prediction process. Finally, the authors in [18] studied the measurement of real-world scenarios from urban, suburban, rural, and dense city areas. They established ML models that utilize radio parameters to predict the end-user throughput and obtained an accuracy of around 93% and 84% in LTE and 5G cellular networks, respectively.

3 Materials and methods

In this research, a framework for predicting DL-Throughput by using ML models is established as (i) data collection, (ii) data pre-processing, and (iii) training, testing, and evaluation of ML models. The multi-stage diagram of the DL-Throughput predictive framework is shown in Fig. 1. The proposed approach requires the actual collection of data for training and validation. Then, data binning and feature selection based on feature correlation is applied. Finally, the ML models are applied for predicting DL-Throughput in cellular networks, and the results compare with actual data for validation to select the optimum ML model.

3.1 Data collection

Collection of DL Throughput data was carried out in the urban environment of Najaf province, Iraq, from three mobile network operators. Radio Frequency (RF) driving test was used to collect the parameters of LTE data at a different route with an average total of 10Km. An RF scanner was attached to a car, and used to log the measurements while driving at different times during weekday during both rush hour and regular hours [19–22]. All the data was collected under the same conditions and weather. The datasets which were used for feature selection to predict DL-Throughput by applying the selected ML models are as follows:

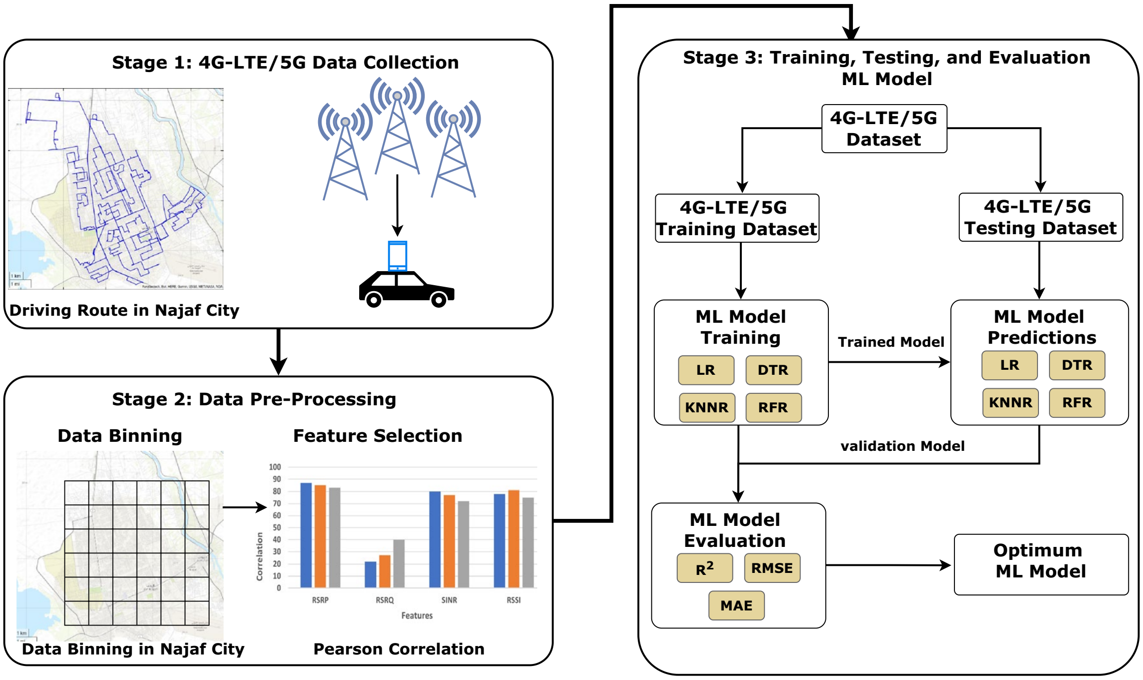


Fig. 1 Framework for DL-Throughput prediction based on ML models

- GPS location: It represents the recorded point in the latitude/longitude geographic coordinate system.
- RSRP: It is the average received power at signal reference and the range between (−44 and −140).
- RSRQ: It describes the quality of the signal and the range between (−3 and −20).
- RSSI: It describes the signal level across the total bandwidth, including serving cell, interference, and noise.
- SINR: It is indicated and used in coding and modulation and is the fraction of the serving cell to the total interference and noise.

A small sample of the recorded LTE data measurements can be seen in Table 1 some parameters in the driving test are disabled due to power consumption.

3.2 Data preprocessing

After the data collection stage, the raw data cannot be used by ML methods due to missing points, outliers, and noise. Therefore, preprocessing techniques have been applied to the measured data to reduce the challenges and guarantee the accuracy and effectiveness of the prediction stage. One of the best techniques is data binning to reduce the fluctuation and provide smoothness for the data measurement. The route area is divided into geographical bins, each assigned 5*m* x 5*m*. Then, the features such as (RSRP, SINR, RSRQ, and DL-Throughput) were applied to the average for each bin.

Feature selection is an essential part of the preprocessing stage for data compression, which involves splitting the datasets into more useful and manageable groups. Selecting appropriate features is crucial to choose suitable and practical information for the ML training stage and excluding the unrelated data points. Various methods for feature selection are utilized, and Pearson correlation has

Table 1 Sample of the data collection

Time stamp	Latitude	Longitude	RSRP (dBm)	RSRQ (dB)	SINR (dB)	RSSI (dBm)	Throughput (Mbps)
09:09:37.000	32.016719	44.330895	– 81.1	– 7.5	17.8	– 52.8	15.708
09:09:39.000	32.016696	44.330957	– 80.7	– 11.2	17.3	– 49.2	74.378
09:09:41.000	32.016662	44.330810	– 82.2	– 11.7	18.1	– 50.1	77.841
09:09:43.000	32.016658	44.330812	– 83.8	– 11.8	16.4	– 51.5	65.246
09:09:45.000	32.016615	44.330701	– 84.6	– 11.9	16.0	– 52.1	64.512

been applied in this work to test the parameters' dependency among the vast dataset. Correlation coefficients range between −1 to +1. The values near +1 specify linear positive correlation, near the zeros specifies weak correlation, and near −1 indicates the correlation in the opposite direction. The formula for Pearson correlation coefficient is shown in Eq. 1:

$$COR = \frac{\sum \left(x_i - \hat{x}\right)\left(y_i - \hat{y}\right)}{\sqrt{\sum \left(x_i - \hat{x}\right)^2 \sum \left(y_i - \hat{y}\right)^2}} \tag{1}$$

Where x_i are values of the x-variable in a sample, $\hat{x}$ is the mean of the values of the x-variable, y_i are values of the y-variable in a sample, and $\hat{y}$ is the mean of the values of the y-variable.

3.3 Training, testing, and evaluating the ML model

After data preprocessing (stage 2), as demonstrated in Fig. 1, the collected dataset was utilized in DL-Throughput modeling. The dataset was first split into training and testing sets. Next, the dataset was randomly shuffled into training and testing sets to produce an accurate model and eliminate data bias. Linear Regression (LR), K-Nearest Neighbours Regression(KNNR), Decision Tree Regression(DTR), and Random Forest Regression (RFR) are some common ML models that have been used for training and testing real-time measured data to select the optimal predictive model. Consequently, these three ML algorithms were applied to the dataset and performance metrics evaluated for each one.

Ideally, a training dataset is utilized to generate and optimize the ML model, while a testing dataset is used to evaluate the accuracy of a ML model. Multiple performance metrics can be used to quantify the performance and select the optimum ML model. These metrics include coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), which can be calculated as shown in Eqs. 2, 3, and 4, respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2} \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2} \tag{3}$$

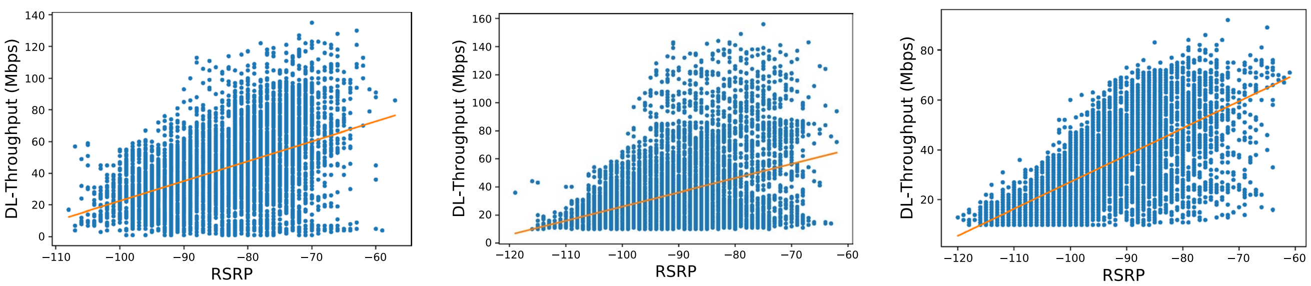
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left|y_i - \hat{y}_i\right| \tag{4}$$

Where y_i, $\hat{y}_i$, and $\hat{y}_i$ represent observed, predicted, and the mean of observed values, respectively.

4 Results and discussion

Prior to the training and prediction phases, a correlation evaluation was employed to assess the significance of each parameter for DL-Throughput. The first feature is the RSRP measurement which is the key indicator for coverage area in LTE networks. RSRP measurements were taken from a steady power reference signal and is not impacted by interference. Therefore, a high level of RSRP indicates the mobile device is near the serving cell, which is the advantage for DL-Throughput. Consequently, RSRP is correlated with DL-Throughput, and Fig. 2 shows the behavior of RSRP measurement with DL-Throughput.

The Scatter figure illustrates the anticipated rise in DL-Throughput with rising RSRP levels. Lower DL-Throughput is related to the RSRP level below −85dBm, whereas higher DL-Throughput is mostly seen at the topside of RSRP. Thus, RSRP measurement is one of the most effective features in predicting the process. Secondly, RSRQ is another essential 4G radiometric. It measures the ratio of total power in the downlink, containing all mobile devices in serving and



(a)RSRP of OPR1 (b)RSRP of OPR2 (c)RSRP of OPR3

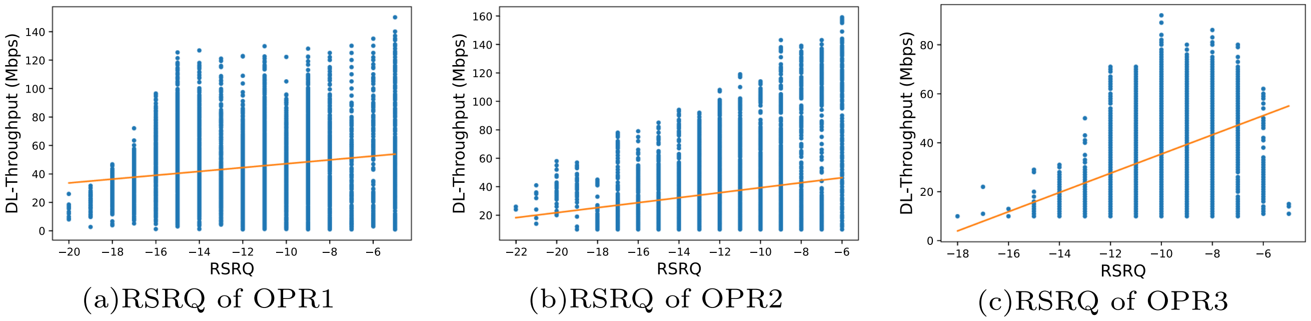
Fig. 2 Scatter plots for RSRP of OPR1, OPR2, and OPR3

neighboring cells, to the received downlink power from the reference signal of the serving cell. RSRQ values with a –20dB threshold denote significant interference and a –3dB threshold denotes pure conditions. Figure 3 presents the relation between the RSRQ and DL-Throughput. The concentration of scatter points of the RSRQ display less correlation than the RSRQ.

Thirdly, the SINR parameter is also highly correlated to the DL-Throughput. The SINR levels are between 0, the lowest level to 30dB, the highest level Fig. 4 presents the correlation between DL-Throughput and SINR. As shown in Fig. 4, the concentration of low DL-Throughput is on the low SINR, and data with high DL-Throughput is intense on the high SINR area.

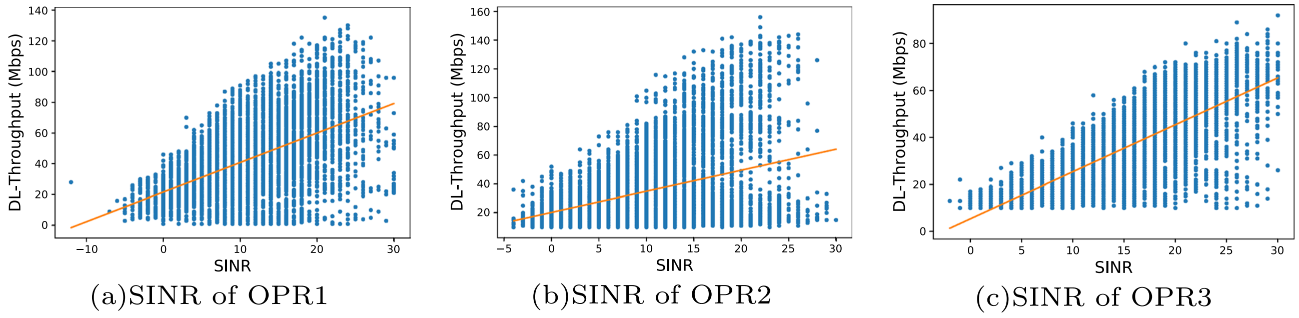
Fourthly, the RSSI parameter is another metric that indicates the total received power, including interference and thermal noise. Figure 5 demonstrates the correlation behavior between DL-Throughput and RSSI. Again, the scatter of the data measurements presents a high correlation similar to SINR and RSRP. The correlation analysis for the feature selection despited in Table 2.

The measurement data is split into training and testing sets. During the training step, 80% of the collected data are assigned and fed into the Random Forest Regression (RFR)



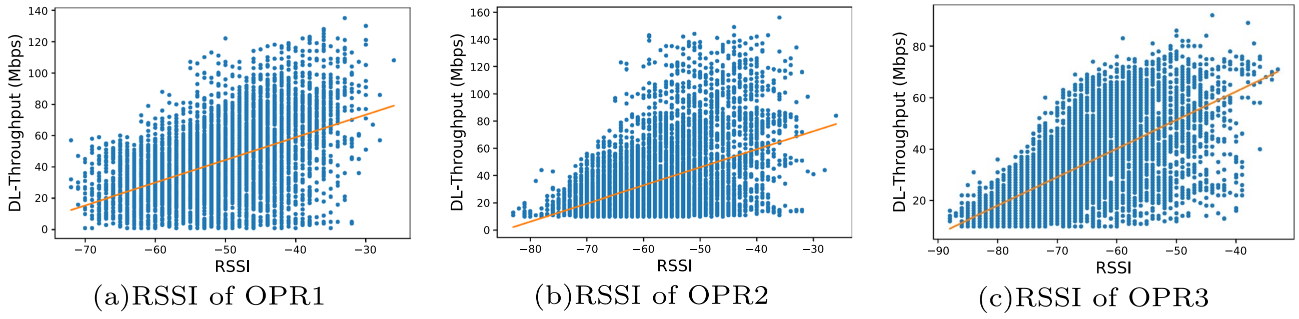
(a)RSRQ of OPR1
(b)RSRQ of OPR2
(c)RSRQ of OPR3

Fig. 3 Scatter plots for RSRQ of OPR1, OPR2, and OPR3



(a)SINR of OPR1
(b)SINR of OPR2
(c)SINR of OPR3

Fig. 4 Scatter plots for SINR of OPR1, OPR2, and OPR3



(a)RSSI of OPR1
(b)RSSI of OPR2
(c)RSSI of OPR3

Fig. 5 Scatter plots for RSSI of OPR1, OPR2, and OPR3

Table 2 Correlation coefficient between feature selection and DL-Throughput

Correlation with DL-throughput	Features selection			
	RSRP	RSRQ	SINR	RSSI
OP1	87	22	80	78
OP2	85	27	77	81
OP3	83	40	72	75

model. The rest of the data is used as a testing set, which evaluates the accuracy of the prediction. The RFR model is selected among other models to predict the DL-Throughput and validate with measured DL-Throughput. Table 3 illustrates the R^2, RMSE, and MAE as a validation step, and the R^2 ranges 93%, 89%, and 96% of OP1, OP2, and OP3, respectively. The performance metric indicates good accuracy prediction with these selected features.

Finally, after training the ML models and testing their performance to predict the DL-Throughput feature,the significance of each feature is examined in order to arrange features according to their influence on prediction performance. Figure 6 captures the importance of each feature and confirms that the RSRP has the most impact on the DL-Throughput prediction.

5 Conclusion

This paper presents a prediction framework built using a set of machine learning models, which are applicable to cellular networks based on 4G-LTE, 5G generations, and beyond. Features from a real-world data set, such as RSRP, RSSI, RSRQ, and SINR, have been used to predict the DL-Throughput for cellular networks. The data measurements have been collected from three well-known cellular communication operators. The correlation coefficient for the selected features shows that RSRP is the most correlated feature with DL-Throughput followed by SINR and RSSI, with RSRQ being weakly correlated with it. In addition, R^2 reached 89% - 96% for the accuracy prediction of DL-Throughput.

Future work will investigate uplink through prediction with ML approaches and deep learning in 5G cellular networks, which is effective with autonomous vehicles (AV). Further, we intend to collect more data measurements and increase the feature set to contain various parameters to enhance the prediction accuracy.

Table 3 Performance metrics for DL-Throughput using RFR model

Operators	Evaluation metrics		
	R^2	MAE	RMSE
OP1	0.93	3.72	6.14
OP2	0.89	4.36	9.01
OP3	0.96	2.03	3.21

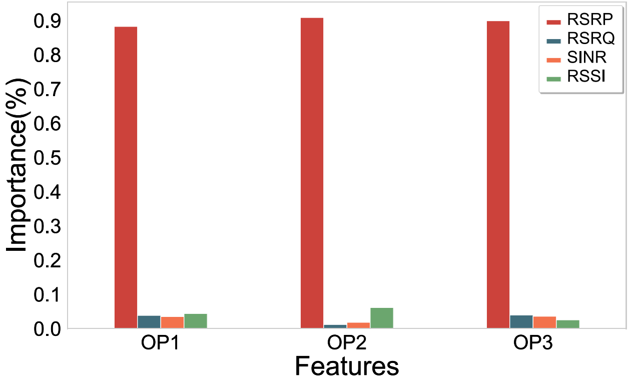


Fig. 6 Features Importance

Data availability The data supporting the findings of this study are not currently available due to future work. We anticipate making the data openly accessible after finishing our works. In the meantime, for any inquiries regarding the data or requests for access, please contact corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Eyceyurt E, Zec J (2020) Uplink throughput prediction in cellular mobile networks. Int J Electron Commun Eng 14(6):149–153
2. Shakir ZD, Zec J, Kostanic I, Al-Thaedan A, Salah MEM (2023) User equipment geolocation depended on long-term evolution signal-level measurements and timing advance. Int J Electr Comput Eng 13(2):1560
3. Kim Y, Kim Y, Oh J, Ji H, Yeo J, Choi S, Ryu H, Noh H, Kim T, Sun F et al (2019) New radio (nr) and its evolution toward 5G-advanced. IEEE Wirel Commun 26(3):2–7
4. Shakir Z, Al-Thaedan A, Alsabah R, Salah M, AlSabbagh A, Zec J (2023) Performance analysis for a suitable propagation model in outdoor with 2.5 GHz band. Bull Electr Eng Inform 12(3):1478–1485
5. Imoize AL, Orolu K, Atayero AA-A (2020) Analysis of key performance indicators of a 4G LTE network based on experimental data obtained from a densely populated smart city. Data Brief 29:105304
6. Rajarajeswarie B, Sandanalakshmi R (2022) Machine learning based hybrid precoder with user scheduling technique for maximizing sum rate in downlink MU-MIMO system. Int J Inf Technol 14(5):2399–405
7. Shakir Z, Mjhool AY, Al-Thaedan A, Al-Sabbagh A, Alsabah R (2023) Key performance indicators analysis for 4 G-LTE cellular networks based on real measurements. Int J Inf Technol 15(3):1347–55

8. Eyceyurt E, Egi Y, Zec J (2022) Machine-learning-based uplink throughput prediction from physical layer measurements. Electronics 11(8):1227
9. Elsherbiny H, Abbas HM, Abou-zeid H, Hassanein HS, Noureldin A (2020) 4G LTE network throughput modelling and prediction. In: GLOBECOM 2020-2020 IEEE Global Communications Conference, IEEE. pp 1–6
10. Abou-Zeid H, Hassanein HS, Valentin S (2014) Energy-efficient adaptive video transmission: exploiting rate predictions in wireless networks. IEEE Trans Veh Technol 63(5):2013–2026
11. AbdulRaheem M, Oladipo ID, Imoize AL, Awotunde JB, Lee C-C, Balogun GB, Adeoti JO (2023) Machine learning assisted snort and zeek in detecting DDoS attacks in software-defined networking. Int J Inf Technol. https://doi.org/10.1007/s41870-023-01469-3
12. Yue C, Jin R, Suh K, Qin Y, Wang B, Wei W (2017) Linkforecast: cellular link bandwidth prediction in LTE networks. IEEE Trans Mob Comput 17(7):1582–1594
13. Lee D, Lee D, Choi M, Lee J (2020) Prediction of network throughput using arima. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE. pp 1–5
14. Wei B, Okano M, Kanai K, Kawakami W, Katto J (2018) Throughput prediction using recurrent neural network model. In: 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), IEEE. pp 107–108
15. Na H, Shin Y, Lee D, Lee J (2021) LSTM-based throughput prediction for LTE networks. ICT Express 19(2):247–52
16. Dong X, Fan W, Gu J (2015) Predicting LTE throughput using traffic time series. ZTE Commun 13(4):61–64
17. Mostafa A, Elattar MA, Ismail T (2022) Downlink throughput prediction in LTE cellular networks using time series forecasting. In: 2022 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom), IEEE. pp 1–4
18. Minovski D, Ogren N, Ahlund C, Mitra K (2021) Throughput prediction using machine learning in LTE and 5G networks. IEEE Trans Mob Comput 22(1):1825–1840
19. Shakir Z, Zec J, Kostanic I (2020) LTE geolocation based on measurement reports and timing advance. In: Advances in information and communication: proceedings of the 2019 Future of Information and Communication Conference (FICC), Vol. 2. Springer. pp 1165–1175
20. Shakir Z, Zec J, Kostanic I (2018) Measurement-based geolocation in lte cellular networks. In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), IEEE. pp 852–856
21. Al-Thaedan A, Shakir Z, Mjhool AY, Alsabah R, Al-Sabbagh A, Salah M, Zec J (2023) Downlink throughput prediction using machine learning models on 4G-LTE networks. Int J Inf Technol 15(6):2987–93
22. Shakir Z, Al-Thaedan A, Alsabah R, Al-Sabbagh A, Salah MEM, Zec J (2022) Performance evaluation for RF propagation models based on data measurement for LTE networks. Int J Inf Technol 14(5):2423–2428