Int. j. inf. tecnol. (August 2023) 15(6):2987–2993
https://doi.org/10.1007/s41870-023-01358-9

ORIGINAL ARTICLE

Downlink throughput prediction using machine learning models on 4G-LTE networks

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

Received: 21 November 2022 / Accepted: 15 May 2023 / Published online: 1 July 2023
© The Author(s), under exclusive licence to Bharati Vidyapeeth's Institute of Computer Applications and Management 2023

Abstract With the enormous evolution of the smartphone, especially with the appearance of the fourth generation (4G) cellular networks, the demand for high-speed data rate, low latency, and video streaming have been increased. This rising demand for network utilization has demonstrated the need for more service improvement. Furthermore, with rising demand and complexity, traditional network management techniques are inadequate, necessitating an autonomous calibration to reduce system parameter usage and processing time. Therefore, real network monitoring and performance analysis should be applied by utilizing various models. Because Downlink Throughput (DL-Throughput) holds significant importance factors for network performance, DL-Throughput prediction can be used to evaluate the quality of cellular networks. Various Machine Learning (ML) models utilized Long-Term Evolution (LTE) data measurements for the prediction process. In this article, the selected ML models Support Vector Regression (SVR), Linear Regression (LR), K Nearest Neighbors (KNN), and Decision Tree Regression (DTR) have been used for forecasting DL-Throughput from three different cellular network operators in an urban area. The parameters with high correlation on throughput and are used as feature selection with ML are the GPS coordinates, RSRP, RSRQ, SINR, and RSSI. The statistical analysis has been utilized to determine the accuracy of the ML models. As a result, the KNN and DTR obtain the best accuracy in the three operators compared with other ML models. For instance, the accuracy for R^2 of DTR is 99%, 93%, and 98% with operator 1 (OPR1), operator 2 (OPR2), and Operator 3 (OPR3), respectively.

Keywords Machine learning · Downlink thourghput prediction · 4G-LTE

1 Introduction

The advances in mobile network technology over the past decade have provided users with boundless services and opportunities. Today smartphone subscribers, especially with the launching of 4G-LTE networks and developing 5G, depend on their phones for leisure and work activities such as smooth video streaming and seamless online gaming [1–4]. Thus, it maximized the load on the mobile network and caused a spike in network traffic. Mobile carriers continually seek solutions to meet the growing demand by emerging new source management and load balancing system and providing higher data rates, low latency and reliability. After November 2019, COVID-19 started appearing in China, and the pandemic began quickly worldwide. Therefore, all work, business, and education are transferred electronically from home in all countries. Accordingly, the pandemic significantly impacted the network data traffic, which increased the demand for improving and guaranteeing the best Quality

✉ 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 TeleWorld Solutions, Chantilly, VA, USA

7 Department of Computer Engineering and Science, Florida Institute of Technology, Melbourne, FL, USA

of Service (QoS) for users, especially download and upload throughput [5].

A predictive methodology for source allocation and network management is a new concept for enhancing network QoS and adopting network scaling challenges. The basic concept is to forecast network connectivity oscillations before they occur to the users and by forecasting these oscillations in advance, the operators can take precautions to mitigate users' QoS necessities. For example, users can pre-purchase more resources to pre-buffer video content in anticipation of future bandwidth will drop for that user [2, 6, 7]. Through the proliferation of bandwidth-hungry applications, streaming on social media platforms and media publishing increase, and the development of connected autonomous vehicles (CAV) expands, the necessity for and advantages of such an aggressive allocation plan will grow. Among the several measures of network quality, predicting bandwidth is the best convenient for guiding predictive network tasks.

However, bandwidth prediction is insignificant because it depends on the received signal strength and other measurements, including contextual data such as geographic location, time, and landscape [8–10]. Hence, a major contribution of this work is the development of various approaches to model and forecast network throughput. The approaches utilize real-time measurements gathered from live mobile networks. It has been done via estimating various ML models to predict the downlink throughput. Various assessment statistics are used to evaluate the algorithm model's effectiveness. Our analysis procedure utilized the following ML models (SVR, LR, KNN, and DTR). Figure 1 shows the learning process and the implementing the algorithm. For more understanding, a comprehensive analysis of 4G LTE network for DL-Throughput.

2 Related work

Some researchers studied throughput prediction with advanced telecommunication, especially in the 4G-LTE and 5G eras which the applications demand high throughput. In [11], the authors investigated the throughput prediction for the network by developing an interface called "PROTEUS" for real-time networks. By using a regression tree model, "PROTEUS" predicated the DL-Throughput based on the last 20 s of network measurements. Meanwhile, the researchers in [12] studied the 4G-LTE throughput prediction using the "QXDM toolset" for gathering data measurement for cellular networks and applied a machine learning framework to predicate DL-Throughput. It predicated link bandwidth by using the data of the lower layer. Also, the authors in [13] used the random forest model to predict the DL Throughput. Their forecast was based on the collected data with additional information for the cellular network; for instance, the average throughput for the cells, the rate of successful connection, and the average number of subscribers by the cell. They found the additional data helps to improve the accuracy of the prediction.

Furthermore, the authors in [14] investigated the prediction of DL-Throughput using the stable LTE data measurement collected from highways, local roads, and pedestrian lanes. This study used two cellular operators, and ML models obtained errors rate around 4% to 17%. Although the study has good DL-Throughput forecasting accuracy, it still needs improvement to reduce error oscillation. In [15], the researchers studied the DL-Throughput prediction using ML models to provide reliable data connection because mobility considerably fluctuates network performance. While in [8], the authors perform DL-Throughput prediction using ML models and time series forecasting such as Arima and neural networks based on LTE data measurements from real cellular networks.

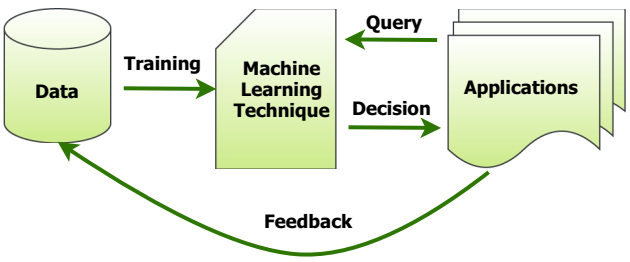

Fig. 1 Learning process

3 Data preprocessing

3.1 Dataset

The 4G-LTE data measurement was collected from three telecommunication operators in an urban area in Najaf city, Iraq. The data was collected from 377 serving cells, and the long route is around 19 Km, as shown in Fig. 2. The collected data was made by driving test, which is the familiar method for collecting and testing cellular networks. The measured data were taken at various times for multiple days, between 9 am to 9 pm, and recorded every two seconds.

Many network parameters have been included in the dataset, including GPS location in the form of (Latitude/longitude) for each point; Reference Signal Received Power (RSRP) is utilized as a major parameter for signal level in the LTE network and determined as a power of reference signal over full bandwidth and narrowband, Reference Signal Receive Quality (RSRQ) is utilized as a major for quality signal and Received Signal Strength Indicator (RSSI) is used as average power over the whole bandwidth. Furthermore, the Signal to Interference and Noise Ratio (SINR) is the ratio

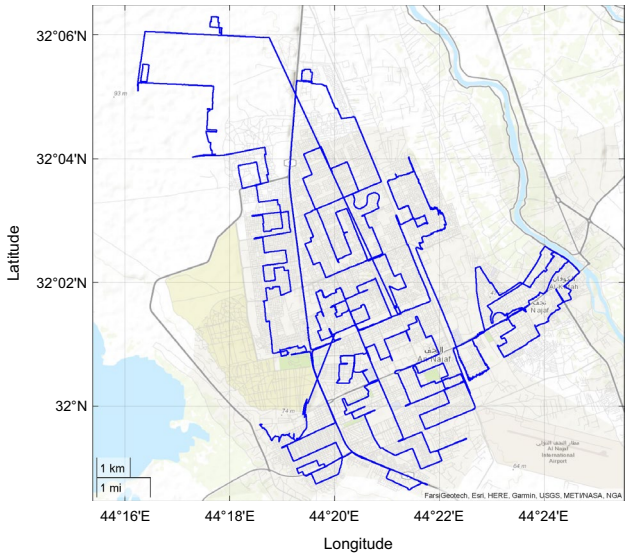

Fig. 2 Driving test rout in Najaf city

between signal to interference from other cells and noise. In addition, DL-Throughput is one of the major parameters for transmitting data in the downlink in (Mbit/s) [16–19].

3.2 Data binning

ML models cannot be directly applied to raw measurement data because the raw data has some outliers, noisy, and missing points. Therefore, several preprocessing were carried out on data to ensure accuracy and effectiveness. Data binning is one of the useful methods to decrease the noise in the data raw due to the car's speed during the driving test. In addition, data binning improves the smoothness of the data and increases the generalizability of the model. The area was divided into bins; each bin dimension was 5 x 5 m, and took the average of each parameter (RSRP, RSRQ, SINR, RSSI, DL-Throughput) was taken.

4 Machine learning models for downlink throughput prediction

ML is a branch of artificial intelligence that has grown rapidly over the past 2 decades (during the past 2 decades) [20, 21]. ML algorithms build a method for constructing or developing models that represent the relationship between data. These models are built by utilizing all the available variables to satisfy numerous hypotheses for prediction or classification. In this work, some popular models have been employed for training and testing the real-time measured data to determine the optimal predictive model, such as SVR, LR, KNN, and DTR.

4.1 Support vector regression (SVR)

It is one of the supervised ML models and is derived from the support vector machine approach that utilized for regression analysis. It utilizes a kernel function method for regression. This technique produces a linear regression model after mapping the given data points into high dimensional feature space. The hyperplane considers an important factor (line) in the SVR model, which is used to fit the given data, while the boundary lines are utilized to bind the points for prediction. The notation ϵ represents the gap from the hyperplane to the boundary. Thus, the SVR model aims to perform the optimal value of the ϵ to ensure that the support vector falls within that boundary line [22]. Equation 1 depicts the linear function:

$$f(x) = wx + b \tag{1}$$

Where $f(x)$ represents the desired variable, w represents the weight coefficient, x represents the input variable, and b represents the bias. Then, the kernal function in this model attempts to increase the flatness of the function by decreasing the squared sum of the weight coefficients w as shown in Eq. 2.

$$\text{minimize} \frac{|\ w\ |^2}{2} \tag{2}$$

And Eq. 3 ensures the restriction that all residuals should be less than ϵ.

$$\forall_n\ :\ |y_n - (wx_n + b)| \leq \epsilon \tag{3}$$

4.2 Linear regression (LR)

LR is one of the most common predictive models in ML. This model aims to identify a linear relationship between the independent and dependent variables [23]. Equation 4 depicts LR line.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \tag{4}$$

The purpose of the LR model is to identify the optimal values for the coefficients β_0 and β_1. The cost function works to determine the regression coefficients β_0 and β_1 and give the best results. The value of the cost function is denoted by Eq. 5.

$$\text{minimize}(1/n \sum_{i=1}^{n} (y_i - \bar{y}_i)^2) \tag{5}$$

4.3 K-nearest neighbors (KNN)

The KNN model is a supervised learning model utilized to solve commonly machine learning problems for instance

classification, regression, and pattern recognition. This model predicts the desired target using the similarities among different data attributes [24–26]. To find similar attributes, this model determines the distance based on Manhattan distance measure as denoted by Eq. 6.

$$D_{KNN}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{6}$$

For each input attribute, the KNN model locates the k nearest neighboring locations utilizing distance measure in Eq. 6. Then, the prediction of the desired value is based on the value of its nearby neighbors. For validation, RMSPE is combined with cross-validation to determine the optimal k value. In regression, the desired value prediction is the mean of its nearest k neighbors data points. Equation 7 indicates the RMSPE formula.

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \tag{7}$$

4.4 Decision tree regression (DTR)

DTR is a supervised learning model in ML algorithms. DTR can be utilized for both classification and regression. In this model, data will be separated based on the parameters. A DTR utilizes a tree structure to generate rules that make predictions based on these rules; it is constructed by splitting the source set into subsets. Each decision tree contains nodes, branches, and leaves where nodes for features, branches for decisions, and leaves for outcomes. The objective of this model constructs a tree of features based on the input dataset and utilize it to generate a distinct outcome at every leaf. DTR is constructed using a technique named binary recursive partitioning. This technique is repeated by partitions and distributes the input dataset into branches. In the model, partitioning continues until the user-specified reduction threshold is reached [27].

5 Prediction results and discussion

Three datasets used in DL-Throughput modeling have been gathered specifically in city of Najaf, Iraq. After data preparation, The collected datasets were divided into training and testing sets to facilitate the DL-Throughput modeling process. The datasets are randomly shuffled as training and testing sets to eliminate data bias and generate a more accurate model. The training set has been utilized to create and optimize the learning model. To avoid biasing in the predictions after constructing the model, the testing set has been utilized to assess the accuracy of the model. In this research, the gathered dataset was divided randomly into 80% for training and 20% for testing and validation using cross-validation to efficiently analyze the performance of the ML models

Two well-known metrics were utilized to evaluate the learning model's performance: the R^2 and the Root Mean Square Errors (RMSE). The R^2 represents the proportion of dependent variable variance described by independent variables. It quantifies the strength of the model's association with the dependent variable. The RMSE is a commonly used metric for measuring the error of a learning model in predicting quantitative data by indicating the difference between the actual and the model-predicted values. Also, RMSE utilizes to identify the model with the best prediction accuracy on the data by comparing the prediction errors of learning models.

All the preprocessing and training of the ML models were performed on a Linux system machine with 8 GB RAM and Intel Core i7 Processor. Various packages have been utilized with Python3, such as Matplotlib, Pandas, Scikit-Learn, and Numpy [28–30].

As mentioned previously, the R^2 and the RMSE were utilized to evaluate the performance of the various ML models. Table 1 depicts the R^2 and the RMSE values of each ML model.

KNN and DTR provide the most accurate predictions among all ML models, as depicted in Table 1.

The best prediction accuracy of KNN and DTR is because both models employ classification to avoid model over-fitting and enhance the accuracy. In addition, by selecting a few samples and a random set of features to build each model, the KNN and DTR models reduce the correlation between distinct samples and the variance in the predictions of the

Table 1 R^2 and RMSE [Mbps] Analysis of ML algorithms for OPR1, OPR2, and OPR3

Model	OPR1		OPR2		OPR3	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
SVR	0.30	19.653	0.11	25.098	0.60	10.228
LR	0.558	16.072	0.548	18.063	0.832	6.997
KNN	0.94	5.836	0.93	7.083	0.97	2.990
DTR	0.996	1.484	0.935	6.638	0.984	2.074

model. On the other hand, the LR and SVR models did not perform satisfactorily because they developed mathematical models. Both models had the lowest prediction accuracy of R^2 and RMSE values, as depicted in Table 1.

Figures 3 and 4 illustrate the performance of each ML model by showing the actual and predicted values of the DL-Throughput.

6 Conclusion

This paper applies ML models to LTE data measurements for the DL Throughput prediction process. We discovered the advantages of using different ML models in the prediction process, for instance (SVR, LR, KNN, and DTR). The LTE data measurements are gathered from the new recently

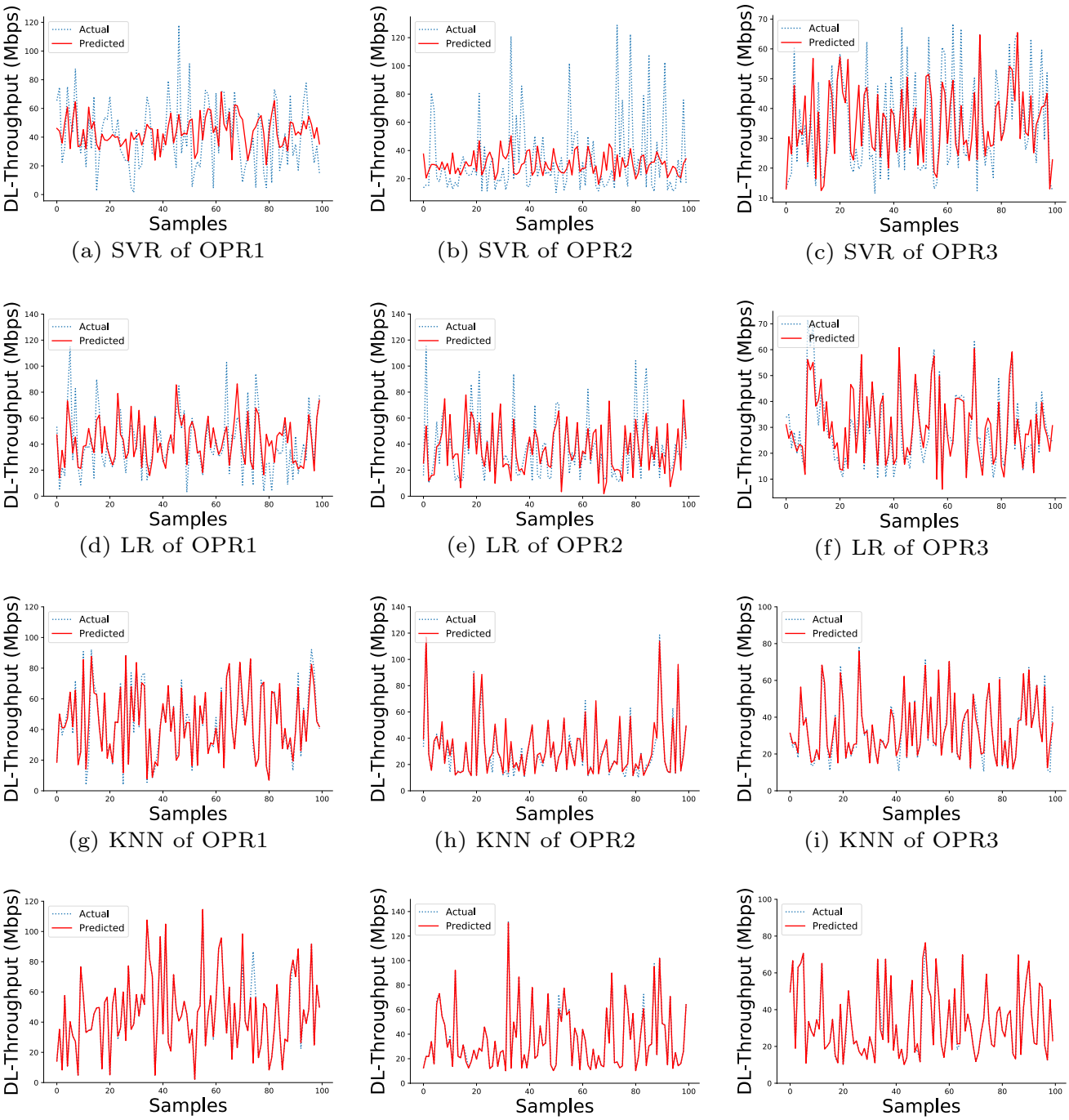

Fig. 3 ML models performance on test set at OPR1, OPR2, and OPR3

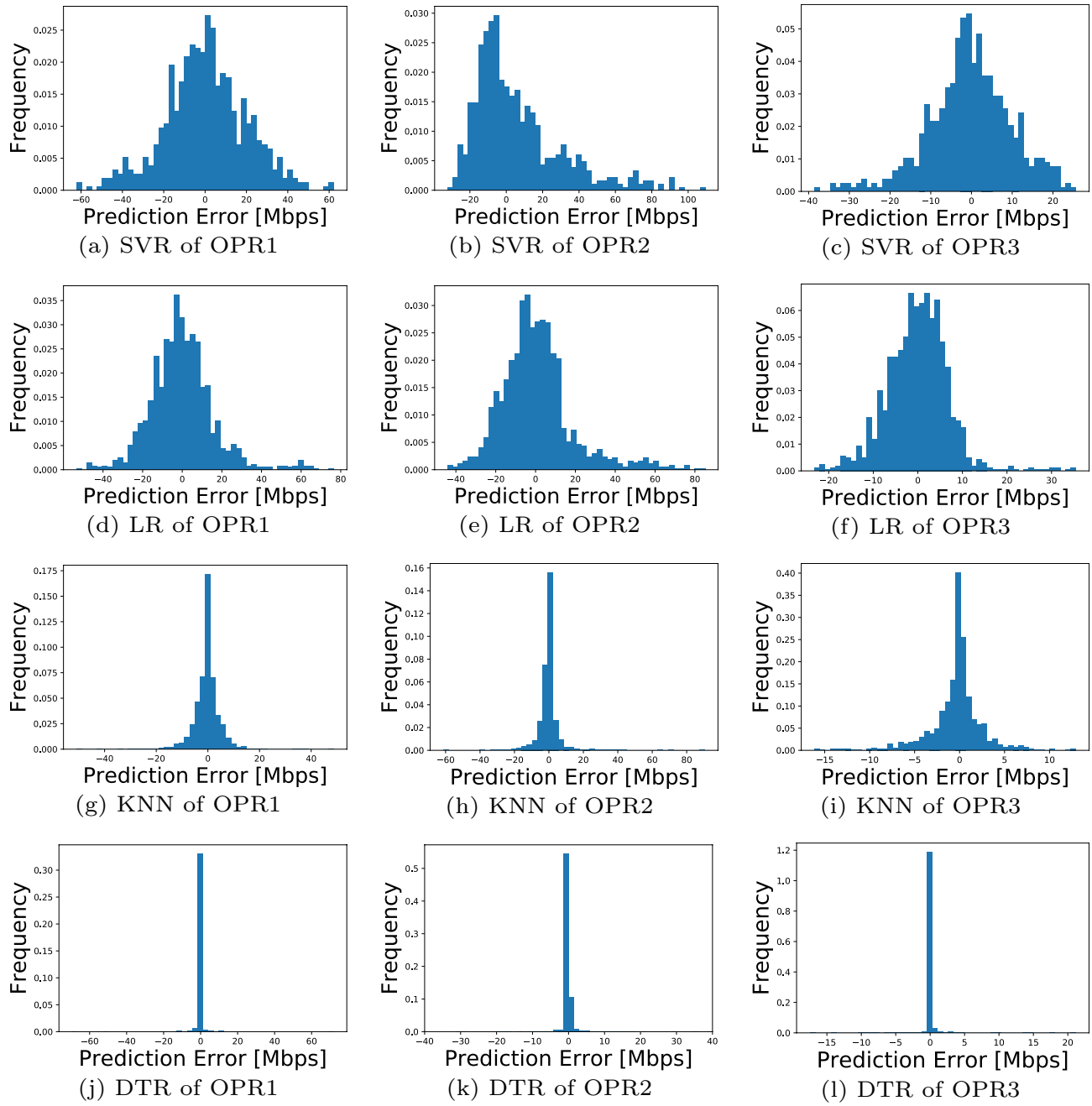

Fig. 4 Prediction models comparison at OPR1, OPR2, and OPR3

deployed LTE networks in an urban area in Najaf city from the three famous operators in this city, OPR1, OPR2, and OPR3. Various feature selections have been utilized with ML models to enhance the accuracy of the predictions, such as GPS coordinates, RSRP, RSRQ, SINR, and RSSI. Furthermore, to choose the best ML model for throughput prediction, its advantages and disadvantages were investigated along with a comparison study utilizing various assessment criteria. The determination coefficient R^2 and RMSE are utilized as evaluation metrics. The highest accuracy for the predictions is KNN and DTR; it observed R^2 is 94%, 93%, and 97% for KNN. And, R^2 is 99%, 93%, and 98% for DTR. Both models have achieved the highest accuracy due to the ease of understanding and accessibility of various real data worldwide.

For future work, various area environments (urban, suburban, and dense city) are intended to investigate with various ML models and compare the performance. Moreover, we are planning to obtain the cell configuration

information as additional data, which will significantly enhance the accuracy of the prediction.

References

1. 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
2. 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
3. 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 14:2423–2428
4. 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 14:23–2405
5. Eyceyurt E, Egi Y, Zec J (2022) Machine-learning-based uplink throughput prediction from physical layer measurements. Electronics 11(8):1227
6. Eyceyurt E, Zec J (2020) Uplink throughput prediction in cellular mobile networks. Int J Electron Commun Eng 14(6):149–153
7. Alsabah R, Aljshamee M, Abduljabbar AM, Al-Sabbagh A (2021) An insight into internet sector in Iraq. Int J Electr Comput Eng 11(6):2088–8708
8. 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
9. Egi Y, Otero CE (2019) Machine-learning and 3d point-cloud based signal power path loss model for the deployment of wireless communication systems. IEEE Access 7:42507–42517
10. Alsabah R, Al-Sabbagh A, Zec J (2017) Calibration of rapidscat scatterometer. In: 2017 IEEE Microwaves, Radar and Remote Sensing Symposium (MRRS), IEEE, pp 249–252
11. Xu Q, Mehrotra S, Mao Z, Li J (2013) Proteus: network performance forecast for real-time, interactive mobile applications. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, pp 347–360
12. Jin R (2015) Enhancing upper-level performance from below: performance measurement and optimization in LTE networks.
13. Samba A, Busnel Y, Blanc A, Dooze P, Simon G (2017) Instantaneous throughput prediction in cellular networks: Which information is needed? In: 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), IEEE, pp 624–627
14. 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
15. Jomrich F, Herzberger A, Meuser T, Richerzhagen B, Steinmetz R, Wille C (2018) Cellular bandwidth prediction for highly automated driving-evaluation of machine learning approaches based on real-world data. In: VEHITS, pp 121–132
16. 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
17. Shakir Z, Zec J, Kostanic I (2019) LTE geolocation based on measurement reports and timing advance. Future Inf Commun Conf. Springer, pp 1165–1175
18. 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 15:1347–1355
19. 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
20. Olukan TA, Chiou Y-C, Chiu CH, Lai C-Y, Santos S, Chiesa M (2020) Predicting the suitability of lateritic soil type for low cost sustainable housing with image recognition and machine learning techniques. J Build Eng 29:101175
21. Verma KK, Singh BM, Dixit A (2022) A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. Int J Inf Technol 14(1):397–410
22. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V (1996) Support vector regression machines. Adv Neural Inf Process Syst 9:155–161
23. Flaih A, Abdalmuhsen A, Abdulah E, Ramaswamy S (2010) Gross product simulation with pooling of linear and nonlinear regression models. Enterp Organ Model Simul EOMAS 2010:69
24. Kramer O (2013) K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors. Springer, Berlin, pp 13–23
25. Itoo F, Singh S et al (2021) Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int J Inf Technol 13(4):1503–1511
26. Nayakwadi N, Fatima R (2021) Automatic handover execution technique using machine learning algorithm for heterogeneous wireless networks. Int J Inf Technol 13(4):1431–1439
27. Wang F, Wang Q, Nie F, Li Z, Yu W, Ren F (2020) A linear multivariate binary decision tree classifier based on k-means splitting. Pattern Recogn 107:107521
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
29. McKinney W et al (2010) Data structures for statistical computing in python. Proc Python Sci Conf 445:51–56
30. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(03):90–95

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.