Machine Learning Based Throughput Prediction for 4G/5G Mobile Networks Using Large Scale Dataset

Hamzah Bahar, Tarteel Sider, and Ali Jamoos
Department of Electronic and Communication Engineering
Al-Quds University
Jerusalem, Palestine
hbahar@staff.alquds.edu, tsider@staff.alquds.edu, ali.jamoos@staff.alquds.edu

***Abstract*—As cellular networks evolve from 4G LTE to 5G, throughput prediction becomes increasingly important for understanding user experience and supporting data-driven network optimization. However, real active-measurement datasets are often sparse, noisy, and technology-dependent, making throughput modeling a challenging regression problem. This paper presents a machine-learning-based framework for predicting downlink and uplink throughput in both 4G and 5G networks using large-scale active measurement dataset. The original dataset contains 367,104 samples and 81 parameters, including radio indicators, scheduling features, modulation statistics, transport block size information, and contextual variables. Due to the high sparsity of the throughput targets, the data were divided into four independent regression tasks: 4G downlink, 4G uplink, 5G downlink, and 5G uplink. After preprocessing, feature ranking, and outlier removal, several regression models were trained and evaluated using 10-fold cross-validation in MATLAB Regression Learner. The results show that tuned Bagged Trees achieved the best performance across all four datasets, reaching R^2 values of 0.9017, 0.9605, 0.8436, and 0.8767 for 4G DL, 4G UL, 5G DL, and 5G UL, respectively. These results demonstrate that target-specific ensemble learning can effectively model the nonlinear relationship between radio/scheduling parameters and measured cellular throughput.**

***Index Terms*—4G LTE, 5G, throughput prediction, machine learning, Large Scale Dataset, Regression Learner, Bagged Trees, RMSE, MAE, coefficient of determination.**

I. Introduction

Mobile communication networks have become a central infrastructure for modern digital services. The evolution from Long-Term Evolution (LTE) to fifth-generation (5G) systems has enabled higher data rates, wider service coverage, and support for applications such as high-definition video streaming, cloud services, Internet of Things (IoT), vehicular communication, and latency-sensitive applications [1]–[4]. However, this evolution has also increased the complexity of network operation and optimization. As the number of users, services, radio parameters, and deployment scenarios grows, maintaining reliable Quality of Service (QoS) becomes increasingly challenging for mobile network operators.

Network operators commonly rely on Key Performance Indicators (KPIs) to evaluate coverage, signal quality, capacity, and Quality of Experience (QoE) [5]. Among these indicators, throughput is one of the most important metrics because it directly reflects the data rate experienced by users. Throughput is affected by several radio and network parameters, including signal strength, signal quality, interference, resource allocation, modulation and coding, and carrier aggregation. Since these parameters interact in a nonlinear manner, throughput prediction cannot be accurately described using simple analytical relationships in many real-world scenarios.

Several measurement-based studies have investigated the behavior of cellular KPIs and their impact on network performance. Experimental LTE studies have shown that parameters such as Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), Signal-to-Interference-plus-Noise Ratio (SINR), and Received Signal Strength Indicator (RSSI) are strongly related to coverage quality and user-perceived performance [6]–[8]. Propagation and measurement-based performance studies have also confirmed that radio channel conditions have a direct effect on achievable data rates and network reliability [9]. These studies highlight the importance of using real measurement data when developing throughput prediction models.

Machine learning (ML) has become a promising approach for modeling cellular network performance because it can learn complex relationships directly from measured data. Earlier studies explored data-driven link and throughput prediction using network measurements and contextual information [10]. Elsherbiny et al. [2] investigated 4G LTE throughput modeling and prediction using machine learning techniques, while Al-Thaedan et al. [11] applied several ML models for 4G LTE/5G downlink throughput prediction. For uplink throughput, Eyceyurt and Zec [12] and Eyceyurt et al. [13] studied the use of physical-layer measurements to predict cellular uplink throughput. These works show that supervised learning can provide effective throughput estimation when relevant radio and network features are available.

In addition to conventional ML models, deep learning methods have also been investigated for throughput forecasting. Na et al. [14] proposed a Long Short-Term Memory (LSTM)-based approach for LTE throughput prediction, showing the potential of sequence-based models for time-dependent network behavior. More recent work has extended ML-based prediction to 5G and beyond-5G networks, where KPI prediction is increasingly important for automation, resource management, and service assurance [3], [4]. Similarly, Yuliana et al. [15] demonstrated the role of feature analysis and machine learning in 5G coverage prediction, further supporting the use of data-driven models in modern cellular networks.

The availability of large-scale measurement datasets has further encouraged research on ML-based cellular performance prediction. Kousias et al. introduced a large-scale dataset containing measurements from operational 4G LTE, NB-IoT, and 5G Non-Standalone (NSA) networks [16]. Such datasets are valuable because they contain real-world network behavior across different technologies, radio conditions, and measurement scenarios. They also allow researchers to evaluate prediction models under practical conditions rather than relying only on simulations or limited drive-test campaigns.

Despite the progress in this area, many existing studies focus on a single radio access technology, a single link direction, or a limited set of radio indicators. In particular, fewer studies provide a unified comparison of ML models for both 4G and 5G throughput prediction while also considering both downlink (DL) and uplink (UL) directions. This distinction is important because downlink and uplink throughput depend on different scheduling behavior, resource allocation, and link conditions. Therefore, a target-specific modeling approach is needed to better understand the performance of ML algorithms across different technologies and transmission directions.

In this work, a machine-learning-based framework is developed for predicting DL and UL throughput in both 4G LTE and 5G networks using active measurement data from a large-scale operational cellular network dataset [16] [17]. The original dataset contains 367,104 samples and includes radio signal indicators, scheduling information, modulation statistics, transport block size measurements, location information, and contextual network parameters. Due to the sparsity of the throughput target variables and the differences between technologies and link directions, the data are divided into four independent regression tasks: 4G DL, 4G UL, 5G DL, and 5G UL throughput prediction. Multiple regression models are trained and evaluated using 10-fold cross-validation in MATLAB Regression Learner. Model performance is assessed using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The results show that tuned Bagged Trees or the so-called Random Forest provide the best overall prediction performance across the four throughput prediction tasks.

II. Methodology

This work follows a supervised machine learning methodology for throughput prediction using active cellular measurement data. The overall workflow is shown in Fig. 1. The proposed framework consists of four main stages: raw dataset preparation, preprocessing, target-specific dataset generation, and regression model training and evaluation. The original dataset was obtained from a large-scale operational cellular network measurement dataset that includes 4G LTE, NB-IoT, and 5G NSA measurements [16]. In this study, only the active throughput measurement records related to 4G and 5G throughput prediction were considered.

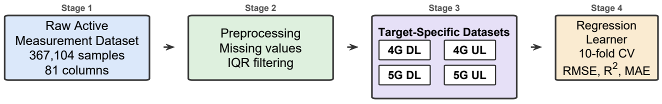

Fig. 1. Proposed methodology for 4G and 5G downlink/uplink throughput prediction using active measurement data.

The original active measurement dataset contains 367,104 samples and 81 columns. These columns include radio-frequency indicators, scheduling-related parameters, modulation information, location information, contextual variables, and measured throughput values. The four target variables considered in this work are LTE Physical Downlink Shared Channel (PDSCH) throughput, LTE Physical Uplink Shared Channel (PUSCH) throughput, 5G PDSCH throughput, and 5G PUSCH throughput.

Since the target throughput values are highly sparse, with more than 92% missing values in each target column, the original dataset was not treated as a single regression problem. Instead, it was divided into four separate datasets: 4G DL, 4G UL, 5G DL, and 5G UL. This separation was necessary because each target corresponds to a different radio access technology and link direction. It also prevents mixing technology-specific predictors and allows each model to use only the features relevant to its corresponding throughput target.

A. Data Preprocessing

The raw dataset could not be directly used for machine learning because it contained missing values, non-numeric missing markers, sparse measurement records, and extreme values. Missing markers such as "?" were first converted to missing values. Numeric variables were imported as double-precision values, while contextual variables such as operator, scenario, and user equipment mode were treated as categorical predictors.

Since the dataset is a sparse active-measurement log, not all radio and scheduling predictors were available in the same row as the measured throughput. Therefore, missing predictor values were replaced by the most recent previous non-missing value within the same campaign, operator, scenario, and user equipment mode group. This forward-filling step was applied only to predictor variables to align recent radio and scheduling conditions with the measured throughput sample. The target throughput columns were never forward-filled or imputed. Rows with missing target throughput values were removed because supervised learning requires measured response labels. The throughput targets were then converted from Kbps to Mbps and renamed as Target_Mbps.

Technology-specific predictor sets were constructed for each dataset. The 4G datasets included LTE-related predictors such as Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), Signal-to-Interference-plus-Noise Ratio (SINR), LTE Physical Cell Identity (PCI), carrier aggregation information, Modulation and Coding Scheme (MCS), modulation rates,

TABLE I
SUMMARY OF THE GENERATED REGRESSION DATASETS

Dataset	Target Variable	Measured Rows	Final Rows	Predictors
4G DL	LTE PDSCH Throughput	26924	14295	22
4G UL	LTE PUSCH Throughput	16653	11331	21
5G DL	5G PDSCH Throughput	16216	5646	22
5G UL	5G PUSCH Throughput	17403	7761	22

TABLE II
EFFECT OF IQR-BASED OUTLIER REMOVAL

Dataset	Rows Before IQR	Rows Removed by IQR	Rows After IQR
4G DL	26924	12629	14295
4G UL	16653	5322	11331
5G DL	16216	10570	5646
5G UL	17403	9642	7761

resource block statistics, and Transport Block Size (TBS) statistics. These radio and KPI-related measurements are commonly used in cellular performance analysis and throughput prediction studies [6], [7], [11]. The 5G datasets used only 5G-related predictors, including Synchronization Signal RSRP (SS-RSRP), Synchronization Signal RSRQ (SS-RSRQ), Synchronization Signal SINR (SS-SINR), 5G PCI, serving Synchronization Signal Block (SSB) information, number of beams, 5G MCS, resource block statistics, TBS statistics, and modulation rates. LTE-specific radio indicators were excluded from the 5G datasets.

Outlier removal was performed using the interquartile range (IQR) method. For a numeric variable x, the first quartile Q_1, third quartile Q_3, and interquartile range are defined as

$$IQR = Q_3 - Q_1. \quad (1)$$

A sample was considered an outlier if it was outside the following range:

$$[Q_1 - 1.5IQR,\ Q_3 + 1.5IQR]. \quad (2)$$

In this work, IQR filtering was applied to the target and numeric predictor space before model training. Although this reduced the number of available samples, this preprocessing strategy produced better validation performance in terms of RMSE and R^2 in the Regression Learner experiments. Therefore, it was adopted for the final model comparison. After outlier removal, remaining missing numeric predictor values were replaced using the median of the corresponding feature, while missing categorical values were assigned to a separate missing category.

Table I summarizes the generated datasets after preprocessing.

Table II shows the effect of IQR-based outlier removal on the four target-specific datasets.

B. Feature Ranking and Selection

After preprocessing, feature ranking was performed separately for each target-specific dataset using the Feature Selection tool in MATLAB Regression Learner. The minimum redundancy maximum relevance (MRMR) algorithm was used to rank the predictors according to their relevance to the response variable while reducing redundancy among them. In this work, MRMR was used mainly as an interpretability tool rather than as a strict dimensionality-reduction step.

The MRMR results showed a consistent trend across the 4G and 5G DL/UL datasets: scheduling, modulation, and resource-allocation features were generally ranked higher than standalone radio-quality indicators. For the 4G DL dataset, LTE PDSCH average Transport Block Size (TBS) was the highest-ranked predictor, as shown in Fig. 2.

This ranking is consistent with fundamental communication theory. On the one hand, Shannon's capacity equation shows that the maximum reliable data rate of a noisy channel is given by

$$C = B \log_2(1 + SINR), \quad (3)$$

where C is the channel capacity, B is the signal bandwidth resource, and $SINR$ is the signal-to-interference-plus-noise ratio.

On the other hand, the Nyquist formula shows that the maximum data rate of an ideal noiseless channel depends on signal bandwidth B and modulation level M

$$C = 2B \log_2(M), \quad (4)$$

These equations explain why TBS is highly relevant for throughput prediction. Indeed, in 4G LTE and 5G systems, TBS represents the number of information bits scheduled for transmission in a transport block. It is directly related to the bandwidth resources allocated, represented by the number of resource blocks. In addition, the modulation level M in (4) is directly related to the selected Modulation and Coding Scheme (MCS). Therefore, larger TBS and higher MCS indicate that more bits are scheduled over the available radio resources, which naturally leads to higher achievable throughput.

For the 4G DL dataset, the remaining high-ranked predictors included LTE PDSCH minimum TBS, LTE PDSCH 16QAM rate, LTE UL MCS index, and LTE PDSCH maximum TBS. Radio-quality indicators such as RSRQ, SINR, and RSRP were also relevant, but ranked lower than the dominant scheduling-related features. A similar pattern was observed in the 5G datasets; for example, in the 5G UL dataset, the highest-ranked predictors included 5G average PUSCH TBS, PUSCH resource-block statistics, and 5G PUSCH MCS, followed by radio-quality indicators such as SS-RSRP and SS-SINR.

Although some cross-link predictors, such as LTE UL MCS index in the 4G DL dataset, appeared among the high-ranked features, they were treated as correlated indicators of the overall link and scheduling conditions rather than direct causal variables. Finally, experiments with reduced MRMR-based feature subsets produced worse RMSE, R^2, and MAE values. Therefore, all predictors within each target-specific dataset were retained in the final models, since even lower-ranked predictors can contribute useful information through nonlinear interactions in ensemble models such as Bagged Trees.

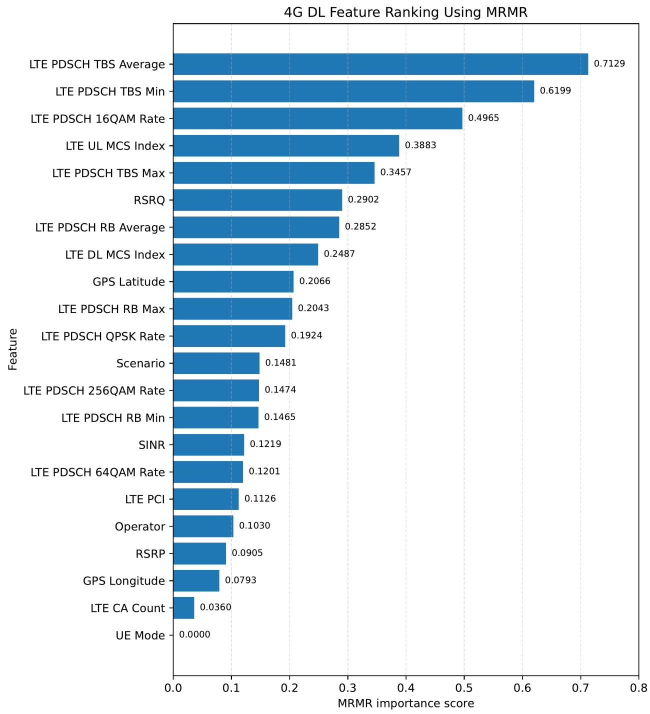

Fig. 2. Representative MRMR feature-ranking results for the 4G downlink dataset obtained using the Feature Selection tool in MATLAB Regression Learner.

C. Machine Learning Training and Evaluation

The four cleaned datasets were imported separately into the MATLAB Regression Learner App. Several regression models were trained and evaluated using 10-fold cross-validation, including linear regression, interaction linear regression, regression trees, boosted trees, default bagged trees, and tuned bagged trees. Hyperparameter tuning was applied only to the Bagged Trees model by varying the minimum leaf size, number of learners, and number of predictors sampled at each split.

The models were evaluated using Root Mean Square Error (RMSE), coefficient of determination (R^2), and Mean Absolute Error (MAE), which are commonly used in throughput prediction studies [2], [11], [13]. These metrics are defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (7)$$

where y_i is the measured throughput, $\hat{y}_i$ is the predicted throughput, $\bar{y}$ is the mean measured throughput, and n is the number of samples. RMSE was used as the main selection metric because it penalizes large prediction errors more strongly, while MAE was used to measure the average absolute prediction error.

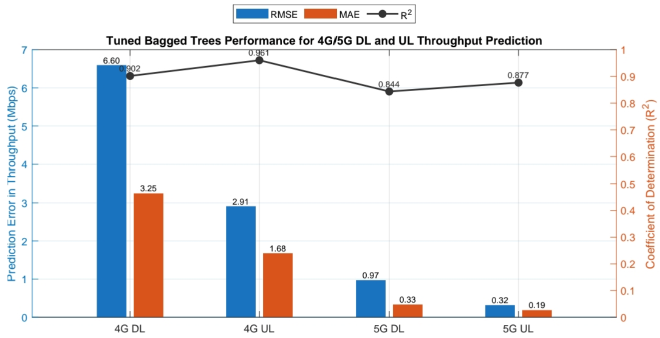

Fig. 3. Performance of the tuned Bagged Trees model for 4G and 5G downlink/uplink throughput prediction.

TABLE III
BEST 10-FOLD CROSS-VALIDATION RESULTS

Dataset	Best Model	RMSE (Mbps)	R^2	MAE (Mbps)
4G DL	Tuned Bagged Trees	6.5967	0.9017	3.2481
4G UL	Tuned Bagged Trees	2.9107	0.9605	1.6785
5G DL	Tuned Bagged Trees	0.9724	0.8436	0.3334
5G UL	Tuned Bagged Trees	0.3173	0.8767	0.1852

III. RESULTS AND DISCUSSION

The performance of the best tuned Bagged Trees model for each dataset is shown in Fig. 3. The figure compares RMSE, MAE, and R^2 for 4G DL, 4G UL, 5G DL, and 5G UL throughput prediction. The detailed numerical results are also summarized in Table III.

For the 4G DL dataset, the tuned Bagged Trees model achieved an RMSE of 6.5967 Mbps, an R^2 of 0.9017, and an MAE of 3.2481 Mbps. This indicates that the model was able to explain approximately 90% of the variance in the measured LTE downlink throughput. The best 4G DL configuration used a minimum leaf size of 1, 100 learners, and 15 predictors sampled at each split. The use of a small leaf size suggests that detailed tree partitions were useful for modeling the nonlinear behavior of downlink throughput.

For the 4G UL dataset, the best model achieved an RMSE of 2.9107 Mbps, an R^2 of 0.9605, and an MAE of 1.6785 Mbps. This was the highest R^2 among all four prediction tasks, which indicates that the selected 4G uplink predictors captured most of the variation in LTE PUSCH throughput. The best 4G UL configuration used a minimum leaf size of 5, 100 learners, and 10 predictors sampled at each split.

For the 5G DL dataset, the tuned Bagged Trees model achieved an RMSE of 0.9724 Mbps, an R^2 of 0.8436, and an MAE of 0.3334 Mbps. Although the absolute error values are small, the R^2 value is lower than the 4G models. This may be related to the smaller final number of samples after preprocessing, the high sparsity of the original 5G PDSCH throughput target, or additional 5G-specific factors not fully represented by the available predictors.

For the 5G UL dataset, the best model achieved an RMSE of 0.3173 Mbps, an R^2 of 0.8767, and an MAE

TABLE IV
Best Tuned Bagged Trees Hyperparameters

Dataset	Minimum Leaf Size	Number of Learners	Sampled Predictors
4G DL	1	100	15
4G UL	5	100	10
5G DL	1	100	15
5G UL	5	100	10

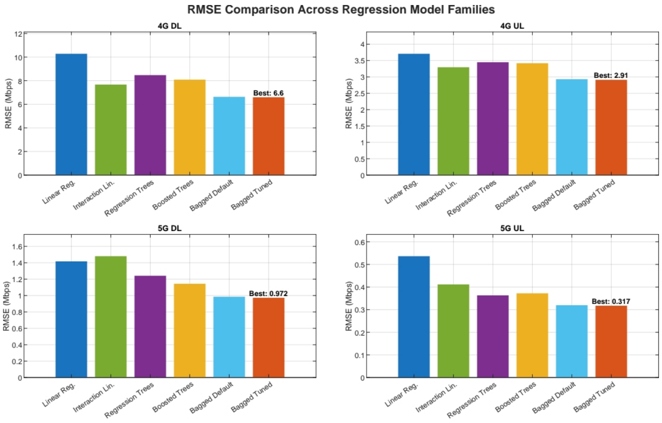

Fig. 4. RMSE comparison across regression model families for 4G and 5G downlink/uplink throughput prediction.

of 0.1852 Mbps. Similar to the 4G UL case, the best configuration used a minimum leaf size of 5, 100 learners, and 10 predictors sampled at each split. This suggests that the uplink models benefited from smoother trees and fewer sampled predictors compared with the downlink models.

Table IV summarizes the best tuned Bagged Trees hyperparameters. The results show that the downlink models achieved their best performance using a minimum leaf size of 1 and 15 sampled predictors, while the uplink models achieved their best performance using a minimum leaf size of 5 and 10 sampled predictors. Therefore, the optimum Bagged Trees configuration depends on both the radio access technology and the link direction.

Fig. 4 compares the RMSE values of the main model families for each dataset. The tuned Bagged Trees model achieved the lowest RMSE in all four cases. This confirms that the hyperparameter tuning step improved the prediction accuracy compared with the default Bagged Trees model and the other regression models.

Fig. 5 compares the R^2 values across the same model families. The tuned Bagged Trees model achieved the highest R^2 for all four datasets, which is consistent with the RMSE results. The improvement is especially clear when comparing tuned Bagged Trees with linear regression models, showing that throughput prediction requires nonlinear modeling.

The model-family comparison shows that linear regression models were generally less accurate than tree-based ensembles. This is expected because cellular throughput is influenced by nonlinear interactions among signal quality, Modulation and Coding Scheme (MCS), modulation rates, allocated resource blocks, and Transport Block Size (TBS). Regression trees improved over linear models in several cases, but single-tree models were less accurate than ensemble methods. Boosted Trees also performed

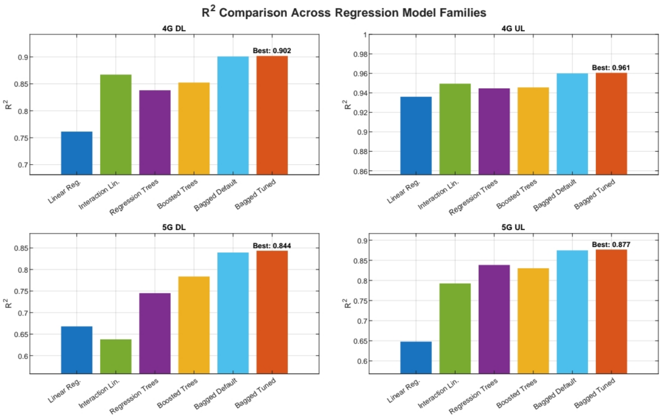

Fig. 5. Coefficient of determination comparison across regression model families for 4G and 5G downlink/uplink throughput prediction.

well, but they did not outperform the tuned Bagged Trees models in any of the four datasets.

Overall, the results demonstrate that separating the original dataset into four target-specific datasets was an effective modeling strategy. Each dataset had different best hyperparameters and different validation performance, confirming that 4G, 5G, downlink, and uplink throughput prediction should not be treated as a single regression problem. The tuned Bagged Trees ensemble was selected as the optimum model because it achieved the lowest RMSE and highest R^2 across all four target datasets. This result is consistent with previous throughput prediction studies, where tree-based and ensemble learning models showed strong performance for modeling nonlinear relationships between radio measurements and cellular throughput [2], [11], [13].

IV. Future Work

This work presented a unified machine learning framework for predicting 4G and 5G throughput in both downlink and uplink directions using active measurement data. Although the proposed preprocessing and target-specific modeling approach produced strong validation results, the study can be extended in several directions.

First, future studies can expand the 5G measurement set by collecting more samples across different operators, mobility conditions, traffic loads, frequency bands, and deployment scenarios. This is particularly important because 5G performance is highly dependent on deployment configuration, spectrum allocation, beam behavior, and user location. A larger and more diverse 5G dataset would allow the models to better capture the variability of 5G behavior, especially for downlink throughput, where the final number of retained samples was more limited than in the 4G cases.

Second, future work can investigate neural-network-based models as an extension to the current regression-based framework. Feedforward neural networks can be trained using the same target-specific datasets to evaluate whether they can capture additional nonlinear relationships between radio, scheduling, and throughput parameters. Moreover, if temporally ordered measurement sequences are constructed from the active measurement logs, sequence-based models such as Long Short-Term Memory (LSTM) networks can be explored to study how recent radio and network conditions affect throughput evolution over time.

Finally, future work can also include a deeper feature analysis for the 5G datasets. In this work, feature ranking showed that scheduling and resource-allocation parameters were highly relevant to throughput prediction. Extending this analysis with larger 5G datasets may provide clearer insight into the role of 5G-specific features such as SS-RSRP, SS-SINR, beam-related parameters, PUSCH/PDSCH resource allocation, and Transport Block Size (TBS). This would help improve both the prediction accuracy and the interpretability of machine-learning-based throughput models.

V. Conclusion

This paper presented a machine-learning-based framework for throughput prediction in 4G LTE and 5G networks using active measurement data. Instead of treating the dataset as a single prediction problem, the work separated it into four target-specific regression tasks: 4G downlink, 4G uplink, 5G downlink, and 5G uplink. This separation was essential because each target belongs to a different radio access technology and link direction, with different missing-value patterns and relevant predictors.

The raw dataset contained 367,104 samples and 81 columns; however, not all columns were used together in a single model. After selecting the appropriate throughput target, four target-specific datasets were generated for 4G DL, 4G UL, 5G DL, and 5G UL prediction. Each dataset contained only the predictors relevant to its corresponding technology and link direction. A complete preprocessing pipeline was then applied, including missing-value handling, predictor forward-filling, throughput conversion from Kbps to Mbps, IQR-based outlier removal, and median imputation. Feature ranking using the MRMR algorithm was performed on the final predictor set of each dataset. For example, the 4G DL dataset contained 22 predictors, and MRMR showed that scheduling-related parameters such as TBS, MCS, modulation rates, and resource block statistics were highly relevant to throughput prediction. However, reducing the selected predictor set based only on MRMR ranking degraded the validation performance; therefore, all predictors within each target-specific dataset were retained in the final models.

Several regression algorithms were evaluated using 10-fold cross-validation in MATLAB Regression Learner. The results showed that tuned Bagged Trees consistently achieved the best performance across all four datasets. The model obtained R^2 values of 0.9017, 0.9605, 0.8436, and 0.8767 for 4G DL, 4G UL, 5G DL, and 5G UL, respectively. These results confirm that ensemble tree models are effective for capturing the nonlinear relationship between radio conditions, scheduling behavior, modulation parameters, and measured throughput.

Overall, the study demonstrates that active measurement data can be used to build accurate and interpretable throughput prediction models when the data are properly cleaned, divided according to technology and link direction, and evaluated using suitable machine learning techniques. The proposed framework provides a practical basis for data-driven cellular network performance analysis and can be extended in future work using larger 5G datasets and neural-network-based models.

References

[1] M. Hassan, *Wireless and Mobile Networking*, 1st ed. CRC Press, 2022. [Online]. Available: https://doi.org/10.1201/9781003042600

[2] H. Elsherbiny, H. M. Abbas, H. Abou-Zeid, H. S. Hassanein, and A. Noureldin, "4G LTE network throughput modelling and prediction," in *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, 2020.

[3] N. P. Tran, O. Delgado, B. Jaumard, and F. Bishay, "ML KPI prediction in 5G and B5G networks," in *EuCNC and 6G Summit*, 2023.

[4] B. Sliwa, H. Schippers, and C. Wietfeld, "Machine learning-enabled data rate prediction for 5g nsa vehicle-to-cloud communications," *arXiv preprint*, 2021. [Online]. Available: http://arxiv.org/abs/2109.04117

[5] S. Nassereldin, R. Sawalmeh, M. Barakat, N. Shanan, and A. Jamoos, "Performance evaluation of palestinian mobile networks based on crowdsourcing measurements," in *Proceedings of the IEEE ICEEE*, 2024, pp. 351–355.

[6] A. L. Imoize, K. Orolu, and A. A. A. Atayero, "Analysis of key performance indicators of a 4g lte network based on experimental data obtained from a densely populated smart city," *Data in Brief*, vol. 29, 2020.

[7] Z. Shakir, A. Y. Mjhool, A. Al-Thaedan, A. Al-Sabbagh, and R. Alsabah, "Key performance indicators analysis for 4G-LTE cellular networks based on real measurements," *International Journal of Information Technology*, vol. 15, no. 3, pp. 1347–1355, 2023.

[8] A. A. Oje and S. O. Edeki, "KPI measurement on the 4G network within the university of ilorin," in *Journal of Physics: Conference Series*, 2021.

[9] Z. Shakir, A. Al-Thaedan, R. Alsabah, A. Al-Sabbagh, M. E. M. Salah, and J. Zec, "Performance evaluation for RF propagation models based on data measurement for lte networks," *International Journal of Information Technology*, vol. 14, no. 5, pp. 2423–2428, 2022.

[10] T. Pögel and L. Wolf, "Optimization of vehicular applications and communication properties with connectivity maps," in *Proceedings of the IEEE 40th Conference on Local Computer Networks (LCN)*, Oct. 2015, pp. 870–877.

[11] A. Al-Thaedan, Z. Shakir, A. Y. Mjhool, R. Alsabah, A. Al-Sabbagh, F. Nembhard, and M. Salah, "A machine learning framework for predicting downlink throughput in 4G-LTE/5G cellular networks," *International Journal of Information Technology (Singapore)*, vol. 16, pp. 651–657, 2 2024.

[12] E. Eyceyurt and J. Zec, "Uplink throughput prediction in mobile networks," *International Journal of Electronics and Communication Engineering*, pp. 149–153, 2020.

[13] E. Eyceyurt, Y. Egi, and J. Zec, "Machine-learning-based uplink throughput prediction from physical layer measurements," *Electronics*, vol. 11, no. 8, 2022.

[14] H. Na, Y. Shin, D. Lee, and J. Lee, "LSTM-based throughput prediction for lte networks," *ICT Express*, vol. 9, no. 2, pp. 247–252, 2023.

[15] H. Yuliana, Iskandar, and Hendrawan, "Comparative analysis of machine learning algorithms for 5G coverage prediction: Identification of dominant feature parameters and prediction accuracy," *IEEE Access*, vol. 12, pp. 18 939–18 956, 2024.

[16] K. Kousias, M. Rajiullah, G. Caso, U. Ali, O. Alay, A. Brunstrom, L. D. Nardis, M. Neri, and M.-G. D. Benedetto, "A large-scale dataset of 4G, NB-IoT, and 5G non-standalone network measurements," IEEE Dataport, 2022. [Online]. Available: https://dx.doi.org/10.21227/7a8s-nt68

[17] K. Kousias, M. Rajiullah, G. Caso, U. Ali, O. Alay, A. Brunstrom, L. De Nardis, M. Neri, and M.-G. Di Benedetto, "A large-scale dataset of 4G, NB-IoT, and 5G non-standalone network measurements," *IEEE Communications Magazine*, vol. 62, no. 5, pp. 44–49, 2024.