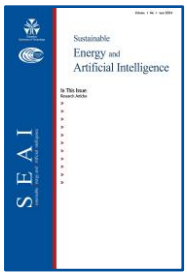
Hamedan University of Technology
Iranian Association of Electrical and Electronics Engineers

Sustainable Energy and Artificial Intelligence

Online ISSN: 3060-8015

Journal Homepage: www.enai.hut.ac.ir



Distance-Aware Machine Learning Approach for Data Rate Prediction in Cellular Networks

Pouya Deabae Shishavan[1], Siavash Rajabi[2*], Reza Shahbazian[3]

[1] *Department of Electrical and Computer Engineering, University of Guilan, Rasht, Iran*
[2] *Department of Electrical Engineering, Hamedan University of Technology, Hamedan, Iran*
[3] *Department of humanities University of Palermo Italy*
* *Corresponding Author: siavash.rajabi@hut.ac.ir*

Article Info

Article type:
Original Article

Article history:
Received 2025-08-30;
Revised 2025-11-24;
Accepted 2025-11-24.

How to cite this article:
Deabae, P., Rajabi, S. and Shahbazian, R. (2025). Distance-Aware Machine Learning Approach for Data Rate Prediction in Cellular Networks. *Sustainable Energy and Artificial Intelligence*, 2(1), 25-35. DOI: 10.61882/seai.2508-1034

Abstract

Accurate estimation of cellular network throughput is essential for effective network management and user experience optimization. This study conducts a comprehensive comparative analysis of four prominent machine learning (ML) models (i.e., Random Forest Regressor (RFR), Gaussian Process Regressor (GPR), K-Nearest Neighbors (KNN), and Support Vector Regressor) for predicting downlink and uplink data rates based exclusively on the user-base station distance. Evaluation metrics, including coefficient of determination (R^2 score), Mean Squared Error (MSE), Mean Absolute Error (MAE), and computational runtime, were employed to assess model performance. Results indicate that ensemble-based (RFR) and probabilistic kernel-based (GPR) approaches outperform instance-based (KNN) and margin-based (SVR) methods in terms of predictive accuracy and error minimization. Furthermore, RFR achieves a favorable balance between accuracy and computational efficiency, making it a practical choice for real-time throughput prediction. Beyond performance estimation, the proposed approach offers potential benefits for energy optimization, enabling more efficient resource allocation both at the user device and base station (or network infrastructure) level. The findings suggest that incorporating additional network parameters and signal quality features could further improve model effectiveness in capturing the complex dynamics of cellular communications. This study employs a public real-world 4G LTE dataset comprising 135 traces (≈15 minutes each) collected under a static mobility model. The proposed distance-only prediction framework highlights that accurate throughput estimation can be achieved using a single distance feature, showing that lightweight models can remain competitive for real-time deployment.

Keywords: *Random Forest Regressor, Gaussian Process Regressor, K-Nearest Neighbors, Support Vector Regressor, cellular networks*.

1. Introduction

In recent years, there has been a significant surge in both the number of mobile devices and the prevalence of data-intensive services such as video streaming, online gaming, and cloud-based applications. This growth has placed considerable strain on cellular networks, requiring them to provide optimal performance. Several variables, including the distance between the mobile device and its serving cell base station, weather conditions, interference, and device movement,

impact the network's ability to deliver dependable and rapid data transfer. Gaining insight into the influence of these phenomena on network performance is essential for enhancing the architecture and administration of cellular networks, particularly as we approach the extensive implementation of 5G and future generations [1,2].

A key obstacle in optimizing cellular networks is the prediction of data rates using network factors, including signal strength and the device's proximity to the network base station. Conventionally, network performance has been estimated using empirical approaches and theoretical models. Nevertheless, the emergence of machine learning (ML) methods presents a chance to build data-driven models to accurately predict key performance indicators (KPIs) such as downlink and uplink data rates [3].

This work investigates the application of ML methods in forecasting downlink and uplink data rates by considering the position of a device in relation to its serving cell base station. More precisely, we utilize four commonly used ML algorithms; Random Forest Regression (RFR), Gaussian Process Regression (GPR), K-Nearest Neighbors (KNN), and Support Vector Regression (SVR). The main question is which of these techniques is highly suitable for this particular regression job and provides complementary advantages. In this paper, we specifically examine stationary or almost stationary devices, assuming that the device speed is either zero or restricted to walking speed. The aforementioned assumption streamlines the modeling process while offering significant insights into the functioning of networks in both indoor and outdoor settings. Prior studies typically rely on rich radio features (RSSI, RSRP, RSRQ, SNR, CQI) and often utilize deep-learning architectures. In contrast, we isolate the predictive value of distance alone (a readily available and privacy-lean feature) and benchmark both accuracy and runtime across four classical regressors to inform real-time, resource-constrained use cases.

The main contributions of this paper are as follows:

- Utilizing a single, simple feature (the user's distance from the base station) for predicting downlink and uplink data rates in cellular networks, minimizing the need for complex data collection.
- Conducting a comprehensive comparison of four ML models: RFR, GPR, KNN, and SVR, evaluated with multiple metrics including R^2 score, mean squared error (MSE), and mean absolute error (MAE).
- Providing an in-depth analysis of both prediction accuracy and computational runtime, offering a practical perspective on model suitability for resource-constrained and real-time wireless network applications.
- Emphasizing the importance of selecting models robust to noise and nonlinearities for wireless throughput prediction, and suggesting future inclusion of richer features such as RSSI and CQI to enhance accuracy.

The subsequent sections of the paper are structured as follows. In Section II, the relevant literature and the significance of precise KPIs prediction in cellular networks are explored. section III provides a detailed description of the dataset and the preprocessing procedures used. Section IV provides an explanation of the ML models employed in this work. The results of our experiments are presented in Section V, followed by limitations and future work highlighted in Section VI. Finally, the paper is concluded in Section VII with a discussion of important findings.

2. Related Works

Predicting throughput and other KPIs in cellular networks remains a critical challenge due to the dynamic and complex nature of wireless communication environments. Traditional models primarily relied on statistical or physics-based approaches, leveraging radio channel parameters such as Received Signal Strength Indicator (RSSI), Channel Quality Indicator (CQI), and Signal-to-Noise Ratio (SNR) to estimate data rates [4,5,6]. While these approaches offer interpretability and domain insights, they often fail to fully capture the non-linear and temporal variabilities inherent in modern networks.

The advent of ML has significantly advanced throughput prediction research, enabling models to learn complex relationships from empirical data. Ensemble learning techniques, notably RFR, have gained popularity because of their capability to handle high-dimensional and noisy data, robustness to overfitting, and interpretability [7], [8]. The authors in [7] investigates various machine learning algorithms for predicting network throughput, evaluating their performance across different network conditions to enhance LTE and 5G system efficiency. Furthermore, [8] proposed a lightweight, comprehensive evaluation method based on the Random Forest algorithm to assess wireless user perception, demonstrating the effectiveness of RFR in delivering accurate and

computationally efficient user experience evaluations.

In parallel, GPR has been employed for its probabilistic nature, which not only produces accurate predictions but also quantifies uncertainty, a crucial feature for adaptive network control. These methods have proven effective in various scenarios, including heterogeneous networks and mobile edge computing contexts, where uncertainty and nonlinearity abound [9], [10]. A scalable Gaussian Process framework optimized with ADMM for efficient large-scale wireless traffic prediction in C-RAN architectures, balancing prediction accuracy and computational cost while outperforming existing methods is introduced in [9]. The authors in [10] propose a hybrid classification and regression model using signal strength measurements to accurately predict throughput variations in 5G networks, enhancing real-time network adaptation for improved QoS and QoE.

Instance-based learning algorithms such as KNN remain popular due to their simplicity and adaptability in diverse application scenarios, especially when working with limited or noisy data. Recent advancements have enhanced the traditional KNN algorithm to address practical challenges. For example, the KWAP-KNN variant introduces kernel-weighted methods to improve neighbor selection, significantly reducing errors in complex indoor positioning environments affected by multipath effects and signal attenuation, thereby increasing accuracy and robustness even in obstacle-dense scenarios [11]. Similarly, in heterogeneous wireless networks, improved KNN-based decision algorithms have been developed to support vertical handover processes by efficiently handling multiple decision attributes with limited samples and reduced complexity. These improvements result in better throughput and lower drop rates, ultimately enhancing system resource utilization [12]. Such innovations demonstrate KNN's continued relevance and potential for adaptation in network-related prediction and decision-making tasks.

On the other hand, SVR, as a kernel-based learning method, continues to be utilized for its strong theoretical foundation and capability to handle nonlinear regression tasks effectively. In wireless multihop networks, SVR has demonstrated superior performance in predicting k-coverage probabilities under complex scenarios involving boundary effects, outperforming models such as Gaussian Regression Neural Networks (GRNN) and Random Forest, with robust metrics indicating its scalability and accuracy for real-world applications like urban sensor networks and disaster management [13]. However, challenges such as the curse of dimensionality and limited communication in large-scale wireless sensor networks (WSNs) affect SVR's performance. To address these, ensemble SVR approaches have been proposed, improving robustness against measurement noise and enhancing localization accuracy compared to conventional SVR models [14]. These advancements highlight SVR's adaptability when combined with ensemble methods but also underscore the need for careful handling of hyperparameters and data characteristics to optimize results in complex network environments.

Recent literature has seen a surge in employing deep learning architectures for various wireless network prediction and optimization tasks, leveraging temporal and spatial correlations from large-scale datasets. For example, bidirectional long short term memory (BiLSTM)-based frameworks have been used for mobile traffic flow prediction in 5G networks with hyperparameter optimization to improve peak-range estimation accuracy [15], while hybrid CNN-regression models enhanced with transfer learning have demonstrated robust performance in RSRP prediction across heterogeneous cell sites [16]. Deep reinforcement learning has also been applied for mobility load balancing, dynamically adjusting cell individual offsets to improve throughput and reduce blocking rates [17]. Similarly, CNN-based architectures have been explored for cellular network positioning by treating localization as an object recognition task in geographic space [18], and attention-aided residual convolutional networks (RAConv) have been proposed for accurate spatial-temporal cellular traffic forecasting [19]. These studies collectively show the capacity of deep models to capture complex, dynamic patterns in wireless environments.

However, despite their predictive power, such models often incur significant computational overhead and require large volumes of labeled data, making them less practical for real-time deployment in resource-constrained scenarios. This limitation underscores the need for lightweight yet accurate alternatives. In our work, we build upon these insights by proposing a model that maintains competitive accuracy while reducing complexity, enabling faster adaptation to network changes without the prohibitive data and processing requirements characteristic of many deep learning approaches.

Our study addresses a notable gap by focusing on throughput prediction solely based on the user-base station distance, a readily available yet often underutilized feature. We conduct a thorough performance and runtime comparison of multiple ML models, emphasizing practical considerations for deployment in scenarios with limited data and computational resources. Unlike many prior works that rely on rich and high-dimensional feature sets, our approach highlights the potential of minimalistic inputs for effective prediction, which is particularly relevant in privacy-sensitive applications or initial network deployment phases. Furthermore, we extend the analysis by considering the implications of model choice on energy optimization, contributing valuable insights to both the academic community and network practitioners.

The significance of this study is in its capacity to greatly improve the optimization of cellular network infrastructure by offering precise forecasts of downlink and uplink data rates depending on the distance to the nearest serving cell eNB. Given the increasing consumption of cellular network data with the widespread adoption of 5G and future technologies, it is crucial for both service providers and customers to guarantee optimal network performance. Conventional approaches to network design generally depend on empirical models or theoretical assumptions, which may not comprehensively encompass the intricacy of real-world settings. By utilizing ML models such as KNN, Random Forest, and ensemble learning methods, this work presents a data-driven strategy to forecast network performance with more precision. Enhanced resource allocation and infrastructure planning not only contribute to improved resource management but also optimize user experience by guaranteeing high-quality data transmission in both indoor and outdoor settings. Furthermore, the results obtained from this study have the potential to enhance the advancement of more sophisticated and flexible networks that can effectively fulfill the requirements of contemporary communication systems.

To the best of our knowledge, comprehensive distance-only throughput prediction together with a systematic accuracy–runtime trade-off across RFR, GPR, KNN, and SVR has not been documented in prior work. Our results quantify the extent to which a single, readily available feature can support reliable prediction, and identify when richer radio features may be required.

3. Dataset and Preprocessing

In this section, the dataset employed in our study is described. A publicly available benchmark dataset is utilized to ensure reproducibility and comparability with existing research [20].

3-1. Data Acquisition

The dataset is based on a 4G Long Term Evolution (LTE) network and comprises 135 individual traces, each averaging approximately fifteen minutes in length. Throughput measurements are recorded at one-second intervals, exhibiting values that span from zero up to 173 Mbps, capturing fine-grained temporal variations in network performance. The dataset contains different mobility patterns, and we focus on both of them, namely static trials and pedestrian trials. The dataset contains measurements such as signal strength, throughput, latency, and other performance indicators across different network conditions and locations, which are suitable for training and evaluating learning-based approaches. The dataset improves the accuracy of the forecasts by including a wide variety of network conditions.

3-2. Data Preprocessing

During the preparation stage, we eliminated duplicate entries and any entries with missing data (including NaN). In addition, we have identified and removed unsuccessful measurements in order to enhance the dependability of the dataset.

3-3. Data split

We divided the dataset into three distinct subsets: training, validation, and test. We use an 80/10/10% train/validation/test split, stratified across traces to avoid leakage between related samples. Only the distance feature is used as the predictor, and all other KPIs are omitted to isolate distance-only performance.

- **Training set:** Contains the largest share of samples and is used to fit each model.
- **Validation set:** Serves exclusively for hyper-parameter tuning and model selection.
- **Test set:** Remains untouched until the end and provides an unbiased estimate of the final model's performance (accuracy, R^2, MSE, and MAE).

3-4. Libraries

All experiments were conducted in Python. We relied on scikit-learn for the learning algorithms, data normalization, and metric calculations; pandas for data handling; and Matplotlib for visualizations.

4. ML Models

The configuration and architecture of the ML models used for simulation and performance evaluation are presented in this section. These models are selected based on their proven effectiveness in similar tasks within the domain. Fig. 1 summarizes the proposed pipeline from data ingestion to evaluation.

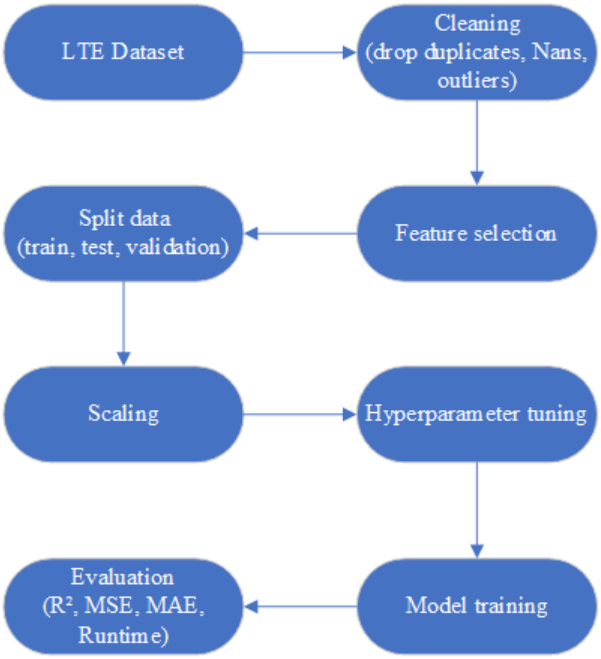


Fig. 1. Pipeline for throughput prediction using distance-only feature: data ingestion, preprocessing, model training/tuning, and evaluation.

4-1. RFR

RFR is a resilient supervised ML technique that has proven effective in modeling and predicting performance metrics in cellular networks. Utilizing ensemble learning in the construction of multiple decision trees on randomly selected subsets of the input features and data samples, the model captures complex, nonlinear dependencies between network parameters and throughput outcomes. The final prediction is obtained by averaging the outputs of individual trees, which helps mitigate overfitting and improve generalization. This makes RFR particularly suitable for handling heterogeneous and dynamic cellular environments where multiple interacting factors influence network behavior [21].

4-2. GPR

GPR is a Bayesian, non-parametric regression technique that models the distribution over possible functions that fit the observed data. Unlike deterministic models, GPR not only provides predictions but also quantifies uncertainty in those predictions, making it especially valuable in scenarios where confidence estimation is important. This approach enables the model to not only provide point estimates but also quantify the uncertainty associated with each prediction, which is especially valuable in dynamic and noisy wireless environments. By defining a covariance function (kernel) that captures the similarity between input features such as geographic location, time, and network load, GPR effectively models nonlinear and spatially correlated behaviors, making it a powerful tool for performance forecasting in cellular systems [9].

4-3. KNN

KNN is a data-driven, non-parametric approach that has been widely applied in cellular networks for predicting KPIs such as signal strength, throughput, and latency. This algorithm estimates the value of a target network parameter by averaging the values of the k most similar historical observations, identified based on a distance metric in the feature space. Its simplicity and ability to model nonlinear relationships without prior assumptions make it suitable for dynamic wireless environments. However, KNN's effectiveness can be influenced by the selection of k, the quality and density of training data, and the dimensionality of input features [22,23].

4-4. SVR

SVR is a supervised learning algorithm based on Support Vector Machines (SVM), designed to perform regression tasks. The algorithm is based on finding a function that deviates from the actual target values by no more than a predefined margin, while maintaining model simplicity. In the context of cellular networks, SVR is effective in predicting performance metrics such as throughput, latency, and signal quality, particularly when the underlying relationships are nonlinear and complex. Its robustness to overfitting and strong generalization ability make SVR a suitable choice for modeling in dynamic and heterogeneous wireless networks [24].

4-5. Hyperparameters and Tuning Protocol

Four regressors (RFR, GPR, KNN, and SVR) are tuned using 3-fold cross-validation with grid

search, selecting the configuration with the lowest mean squared error (MSE). The chosen grids reflect commonly used values in the literature and scikit-learn defaults, balancing parameter-space exploration with computational feasibility. For RFR, we varied the ensemble size n_estimators ∈{50,100,200}, tree depth ∈{3,5,10,"None"}, and minimum split size ∈{2,5,10}. These parameters probe the bias–variance tradeoff inherent to tree ensembles.

In GPR, four kernel initializations were considered (RBF with length-scales 1.0, 0.5, and 0.1 (each with a White Kernel component), as well as DotProduct+White) while the optimizer refined kernel hyperparameters during training. The alpha values {10^(-5),10^(-2),1} provide numerically robust noise regularization. The White Kernel noise term models independent Gaussian observation noise and stabilizes inference when measurement noise or ill-conditioned data are present.

In the KNN, we explored neighborhood sizes {3,5,7,11,15}, weighting schemes (uniform vs. distance), and Minkowski exponents p∈{1,2}, capturing the locality and metric sensitivity of the model. Finally, For SVR, both RBF and linear kernels were tested with C∈{0.1,1,10,100} and ϵ∈{0.01,0.1,0.5,1}. These parameters represent the margin–smoothness tradeoff and the loss-insensitive region.

Because distance is the sole input feature and carries meaningful physical scale, we did not apply feature standardization; tree-based models are scale-invariant, and for distance-based GPR/KNN, monotonic scaling does not improve performance. The literature supports the negligible benefit of normalization in single-feature, physically interpretable inputs.

Table 1 summarizes the full hyperparameter grids and explains the conceptual impact of each parameter.

5. Numerical Results

In this section, we present a comparative evaluation of the four machine learning models (RFR, GPR, KNN, and SVR) applied in this study. The assessment is based on four key performance indicators: the coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), and runtime. Together, these metrics offer a balanced perspective on each model's predictive accuracy, error magnitude, and computational efficiency, enabling a comprehensive comparison of their effectiveness for cellular network performance forecasting.

Figs. 2 and 3 summarize the predictive performance of four regression models for downlink and uplink throughput estimation based on user–base station distance. For the downlink predictions (Fig. 2), GPR achieved the highest coefficient of determination (R^2=0.881), followed closely by RFR (R^2=0.879), while KNN recorded a slightly lower but still competitive score (R^2=0.838). The uplink results (Fig. 3) exhibit a similar ranking, with GPR, RFR, and KNN attaining R^2 scores of 0.794, 0.791, and 0.767, respectively. These values indicate that, despite using a single spatial feature, the models (particularly RFR and GPR) were able to capture a substantial portion of the nonlinear patterns between distance and achievable data rates.

Table 1. Model-wise summary of tuned hyperparameters and their effects on performance.

Model	Hyperparameters tuned (range or grid)	Impact on the Model
RFR	n_estimators= {50,100,200};	Controls ensemble size; governs bias–variance tradeoff.
	max_depth= {3,5,10, None};	Limits tree growth; prevents overfitting; increases interpretability.
	min_samples_split= {2,5,10}	Avoids overly small, noisy leaf nodes; stabilizes variance.
GPR	kernel ∈ {RBF+White (ℓ=1.0), RBF+White ℓ=0.5), RBF+White (ℓ=0.1) DotProduct+White};	RBF models smooth spatial trends; White Kernel accounts for measurement noise.
	alpha={10^{-5},10^{-2},1}	Adds ridge-like noise for numerical stability and robust inference.
KNN	neighbors= {3,5,7,11,15};	Controls locality of averaging and the bias–variance balance.
	weights= {uniform, distance};	Distance weighting reduces bias when samples are unevenly distributed.
	p= {1,2}	Chooses Manhattan vs. Euclidean metric; impacts similarity geometry.
SVR	kernel∈{RBF, linear} (RBF with gamma='scale')	RBF captures nonlinear structure; linear supports simpler relations.
	C= {0.1,1,10,100};	Penalizes margin violations; influences model flexibility.
	epsilon= {0.01,0.1,0.5,1};	Defines the insensitivity zone around the target function.

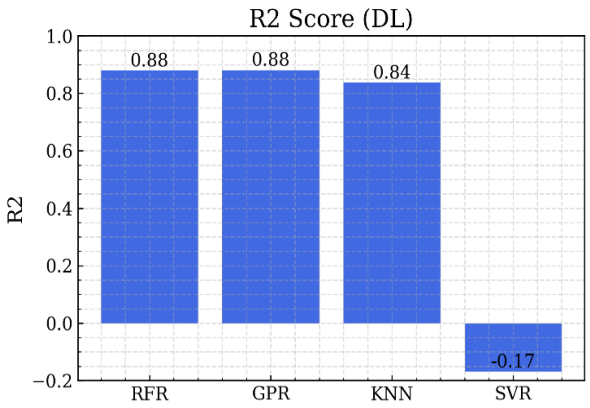


Fig. 2. R^2 scores for downlink throughput prediction using RFR, GPR, KNN, and SVR algorithms.

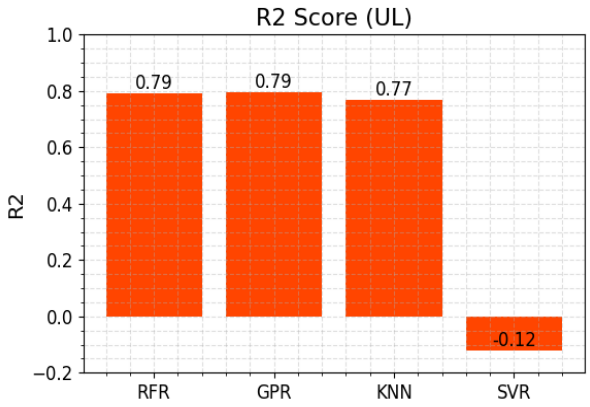


Fig. 3. R^2 scores for uplink throughput prediction using RFR, GPR, KNN, and SVR algorithms.

The near parity in performance between GPR and RFR across both tasks highlights their robustness in handling complex, nonlinear behaviors in network performance. RFR's ensemble approach of averaging over multiple decision trees helps reduce variance and overfitting, whereas GPR's kernel-based probabilistic framework excels in modeling smooth nonlinearities and providing uncertainty quantification. KNN's reasonable accuracy suggests that local instance-based predictions can still be effective when the training set adequately covers the feature space. On the other hand, SVR's performance was markedly poor, yielding negative R^2 values in both downlink (−0.168) and uplink (−0.121), indicating results worse than a mean predictor. This underperformance is likely due to the high sensitivity of SVR to kernel choice and hyperparameter settings, combined with the limited single-feature input space, which constrains its ability to find an optimal separating function in the regression context.

Beyond algorithmic considerations, the inherent stochastic nature of cellular networks imposes a ceiling on achievable prediction accuracy. Real-world data rates are influenced by numerous dynamic factors, such as multipath fading, shadowing, interference, user mobility, and base station scheduling policies, which are not explicitly represented in the input features. As a result, even high-capacity models like GPR and RFR cannot fully capture instantaneous variations solely from distance-based information, leading to accuracy plateaus below 0.89 for downlink and 0.80 for uplink. The results suggest that incorporating additional features such as received signal strength, channel quality indicators, or network load metrics would likely improve model accuracy and narrow the performance gap between more feature-sensitive algorithms like SVR and the stronger ensemble and kernel-based methods observed in this study.

Figs. 4 and 5 present the Mean Squared Error (MSE) values for the downlink and uplink throughput prediction tasks. In both cases, the ensemble-based RFR and the probabilistic GPR models achieve the lowest error levels, closely followed by KNN, while SVR produces substantially larger errors. This trend is consistent with the R^2 results, reinforcing the observation that models capable of capturing nonlinear dependencies and handling noise in the data, such as RFR and GPR, are better suited to the inherent variability of wireless communication channels. KNN's moderate error levels reflect its ability to perform reasonably well with sufficient neighborhood representation, although it lacks the global generalization capacity of the top-performing models. The significantly higher MSE values for SVR indicate its inability to model the underlying throughput patterns effectively when limited to a single spatial feature.

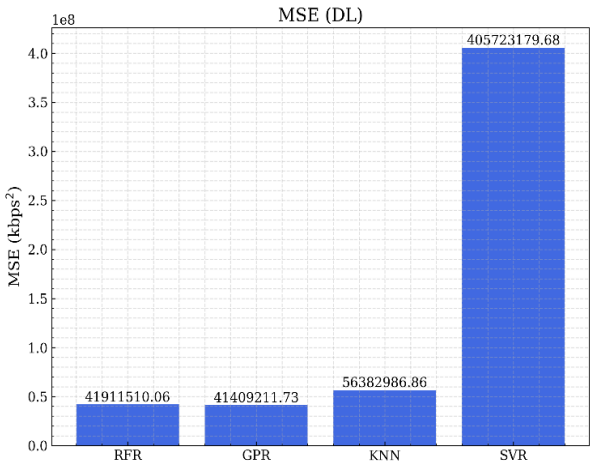


Fig. 4. Mean Squared Error (MSE) for downlink throughput prediction.

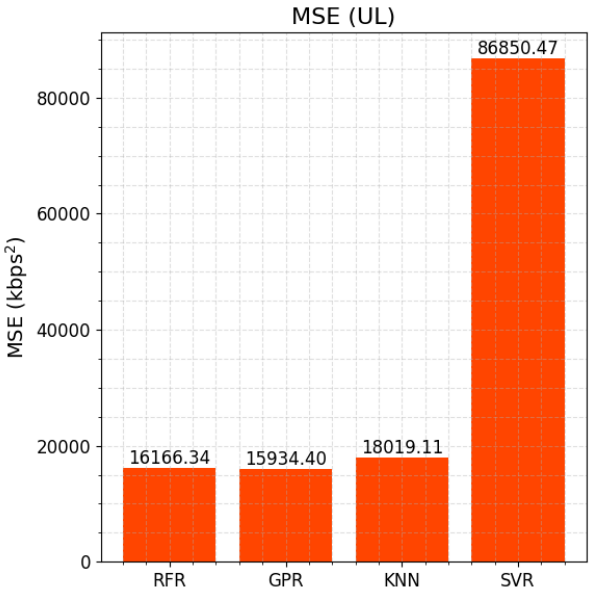


Fig. 5. Mean Squared Error (MSE) for uplink throughput prediction.

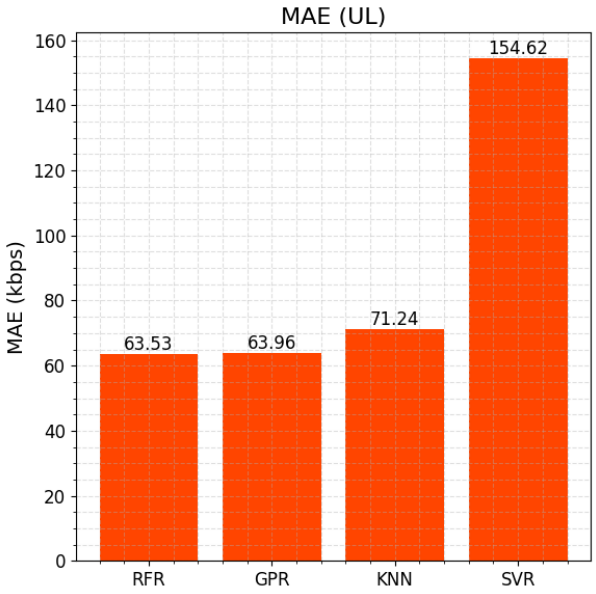


Fig. 7. Mean Absolute Error (MAE) for uplink throughput prediction.

Figs. 6 and 7 show the Mean Absolute Error (MAE) for the same prediction tasks. Similar to the MSE results, RFR and GPR outperform the other models, producing the smallest average deviations between predicted and actual throughput values. This consistent performance advantage across both MSE and MAE metrics highlights their robustness in modeling the complex, noisy, and nonlinear relationships that characterize real-world cellular network data. KNN again yields competitive but slightly higher errors, reflecting its dependence on local sample density. In contrast, SVR's much larger MAE values confirm its instability and suboptimal fit under the given feature constraints and parameter settings.

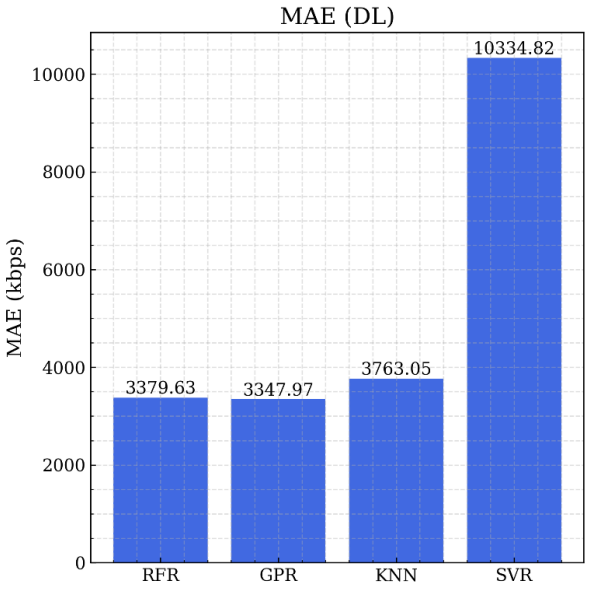


Fig. 6. Mean Absolute Error (MAE) for downlink throughput prediction.

Table 2 presents the training times (in seconds) for four different ML algorithms, measured for both downlink and uplink models. The experiments were conducted on Google Colab with a high-resource runtime, as indicated by 334.6 GB of available RAM and 225.3 GB of disk space. The results show a clear trade-off between predictive performance and computational efficiency. KNN achieves the fastest runtime in both tasks, owing to its non-parametric nature and the absence of a costly training phase. However, this efficiency comes at the expense of slightly reduced prediction accuracy compared to the top-performing models. RFR demonstrates a good balance, delivering high prediction accuracy with moderate computation times.

In contrast, GPR and SVR are significantly more computationally expensive. GPR, while achieving accuracy comparable to RFR, requires considerable processing time due to the complexity of kernel matrix computations, especially when dealing with large datasets. SVR not only exhibits the highest runtimes but also delivers the poorest predictive performance, suggesting that its high computational cost is not justified under the given feature and data constraints. These findings highlight the importance of considering both accuracy and runtime when selecting models for real-time or resource-constrained wireless network prediction scenarios.

The comprehensive evaluation of four ML models-Random Forest Regressor (RFR), Gaussian Process Regressor (GPR), k-Nearest Neighbors (KNN), and Support Vector Regressor (SVR)-for predicting downlink and uplink

throughput reveals clear differences in accuracy, error metrics, and computational efficiency. Both RFR and GPR consistently outperform other models, achieving the highest R2R^2R2 scores (above 0.87 for downlink and near 0.79 for uplink) and the lowest error values across MSE and MAE metrics. KNN shows competitive but slightly lower accuracy and higher error rates, benefiting from its simplicity and local approximation capabilities.

Table 2. Training runtimes (in seconds) for the four prediction models in both downlink and uplink tasks.

Model	DL train time (s)	UL train time (s)
RFR	7.551	6.939
GPR	31.191	38.504
KNN	0.407	0.395
SVR	54.486	57.554

SVR, on the other hand, demonstrates the weakest predictive performance, with negative R2scores and significantly higher error metrics, suggesting poor generalization from the limited single-feature input. From a computational perspective, KNN offers the fastest runtime by avoiding an explicit training phase, while RFR provides a balanced trade-off between accuracy and execution time. GPR and SVR require substantially longer runtimes, with SVR being the slowest despite its low accuracy, limiting their practicality in real-time or resource-constrained environments.

These findings emphasize the importance of selecting models that not only provide strong nonlinear modeling capabilities but also maintain computational efficiency, particularly for wireless network applications where rapid and accurate throughput estimation is critical.

5. Limitations and Future Works

To further improve the prediction results, expanding the feature set beyond just the distance to the base station could significantly enhance model accuracy; for example, incorporating signal quality indicators (RSSI, RSRP, RSRQ) or environmental factors. Increasing the size and diversity of the dataset, including measurements from different locations, times, and network conditions, would help the models generalize better to unseen scenarios. Finally, testing additional ML algorithms such as gradient boosting methods, neural networks, or other ensemble approaches could reveal models that outperform the current RFR, KNN, GPR, and SVR configurations.

6. Conclusion

This research systematically evaluated the effectiveness of four widely used ML models (RFR, GPR, KNN, and SVR) in predicting downlink and uplink throughput in cellular networks based solely on the spatial feature of user-base station distance. Both RFR and GPR consistently demonstrated superior predictive accuracy and robustness, achieving the highest coefficients of determination and the lowest error metrics across multiple evaluation criteria. KNN exhibited moderate predictive capabilities, leveraging local instance-based learning to approximate nonlinear relationships with reasonable accuracy. In contrast, SVR underperformed significantly, highlighting its limitations in handling regression tasks with minimal and low-dimensional feature sets. Additionally, an analysis of computational runtimes revealed that while KNN offers minimal processing delay, RFR provides a balanced compromise between computational efficiency and prediction quality. The notably higher runtimes of GPR and SVR further constrain their practical deployment in real-time or resource-limited settings. These findings underscore the importance of selecting appropriate ML frameworks that balance accuracy, robustness, and computational demands for wireless network throughput estimation.

References

[1] “A Survey on Energy Optimization Techniques in UAV-Based Cellular Networks: From Conventional to Machine Learning Approaches.” Accessed: Aug. 12, 2025. [Online]. Available: https://www.mdpi.com/2504-446X/7/3/214

[2] Khaled, H., & Alkhazraji, E. (2024). AI optimization-based heterogeneous approach for green next-generation communication systems. *Sensors*, *24*(15), 4956.

[3] Das, B. R., Hasan, S. R., Sabuj, S. R., Hossain, M. A., & Ray, S. K. (2025). A Comprehensive Survey on Emerging AI Technologies for 6G Communications: Research Direction, Trends, Challenges, and Opportunities. *International Journal of Intelligent Networks*.

[4] Andrews, J. G., Buzzi, S., Choi, W., Hanly, S. V., Lozano, A., Soong, A. C., & Zhang, J. C. (2014). What will 5G be?. *IEEE Journal on selected areas in communications*, *32*(6), 1065-1082.

[5] Islam, M. Z., Ali, R., Haider, A., & Kim, H. S. (2022). QoS provisioning: key drivers and enablers toward the tactile internet in beyond 5G Era. *IEEE Access*, *10*, 85720-85754.

[6] Coronado, E., Valero, V., Cambronero, M. E., & Orozco-Barbosa, L. (2023). Modelling, simulation and performance evaluation of the IEEE 802.11 e protocol with station mobility. *PeerJ Computer Science*, *9*, e1457.

[7] Minovski, D., Ögren, N., Mitra, K., & Åhlund, C. (2021). Throughput prediction using machine learning in LTE and 5G networks. *IEEE Transactions on Mobile Computing*, *22*(3), 1825-1840.

[8] Zhang, K., Wang, J., Zhang, W., Wang, K., Zeng, J., Fan, G., & Gui, G. (2019). Random forest algorithm-based lightweight comprehensive evaluation for wireless user perception. *IEEE Access*, *7*, 173477-173484.

[9] Xu, Y., Yin, F., Xu, W., Lin, J., & Cui, S. (2019). Wireless traffic prediction with scalable Gaussian process: Framework, algorithms, and verification. *IEEE Journal on Selected Areas in Communications*, *37*(6), 1291-1306.

[10] Costa, M., Ayanda, D., Sátiro, B., Mate, D., & Ortega, A. (2023, September). Throughput Slopes Prediction in 5G Networks with Gaussian Regression Process. In *2023 16th International Conference on Signal Processing and Communication System (ICSPCS)* (pp. 1-5). IEEE.

[11] Tao, L., Bao, D., & Cui, Y. (2024, October). Optimization Analysis of Wireless Sensor Coverage Indoor Positioning Based on KWAP-KNN Algorithm. In *2024 First International Conference on Software, Systems and Information Technology (SSITCON)* (pp. 1-5). IEEE.

[12] Shiwei, G. U. O. (2021, July). An improved KNN based decision algorithm for vertical handover in heterogeneous wireless networks. In *2021 40th Chinese Control Conference (CCC)* (pp. 3011-3016). IEEE.

[13] Singh, S., Kumar, A., Kankane, B., & Mishra, R. (2025, January). Predicting K-Coverage in Wireless Multihop Networks with Boundary Effects Using Support Vector Regression: A Feature Sensitivity Analysis for Smart City Applications. In *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)* (pp. 1516-1521). IEEE.

[14] Kim, W., Park, J., Yoo, J., Kim, H. J., & Park, C. G. (2012). Target localization using ensemble support vector regression in wireless sensor networks. *IEEE transactions on cybernetics*, *43*(4), 1189-1198.

[15] Selvamanju, E., & Shalini, V. B. (2022, October). Deep learning based mobile traffic flow prediction model in 5G cellular networks. In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1349-1353). IEEE.

[16] Wongphatcharatham, T., Phakphisut, W., Jaruvitayakovit, T., Boonkajay, A., & Huang, J. (2024, October). Deep Learning Aided Robust RSRP Prediction in Cellular Networks. In *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)* (pp. 1-5). IEEE.

[17] Alsuhli, G., Banawan, K., Attiah, K., Elezabi, A., Seddik, K. G., Gaber, A., ... & Gadallah, Y. (2021). Mobility load management in cellular networks: A deep reinforcement learning approach. *IEEE Transactions on Mobile Computing*, *22*(3), 1581-1598.

[18] Lin, Y., Tong, Y., Zhong, Q., Gao, R., Yin, B., Liu, L., ... & Chai, H. (2021, December). Dccp: Deep convolutional neural networks for cellular network positioning. In *2021 IEEE Global Communications Conference (GLOBECOM)* (pp. 1-6). IEEE.

[19] Wang, Z., & Wong, V. W. (2022, May). Cellular traffic prediction using deep convolutional neural network with attention mechanism. In *ICC 2022-IEEE International Conference on Communications* (pp. 2339-2344). IEEE.

[20] "4G LTE Speed Dataset and Bandwidth." Accessed: Aug. 12, 2025. [Online]. Available: https://www.kaggle.com/datasets/aeryss/lte-dataset

[21] Fauzi, M. F. A., Nordin, R., Abdullah, N. F., & Alobaidy, H. A. (2022). Mobile network coverage prediction based on supervised machine learning algorithms. *Ieee Access*, *10*, 55782-55793.

[22] Jeske, M., Sansò, B., Aloise, D., & Nascimento, M. C. (2024). Received Signal Strength Indicator Prediction for Mesh Networks in a Real Urban Environment Using Machine Learning. *IEEE Access*.

[23] Azoulay, R., Edery, E., Haddad, Y., & Rozenblit, O. (2023). Machine learning techniques for received signal strength indicator prediction. *Intelligent Data Analysis*, *27*(4), 1167-1184.

[24] "A machine learning approach to TCP throughput prediction | Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems," ACM Conferences. Accessed: Aug. 13, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/1254882.1254894

Biography

Pouya Deabae Shishavan is currently pursuing his M.Sc. degree in Electrical Engineering (Communication Systems) at the University of Guilan, Iran. He received his B.Sc. degree in Electrical Engineering (Communication) from Hamedan University of Technology. His research interests include machine learning, data science, and their applications in modern communication systems.

Siavash Rajabi. received his B.Sc. degree in Electrical Eng. from the University of Tehran in 2003 and the M.Sc. degrees in telecommunications from Khajenasir University of Technology, Tehran, Iran, in 2006. He received the Ph.D. degree in electrical engineering from the Shahid Beheshti University, Tehran, Iran, in 2018. He is currently with the Department of Electrical Engineering, Hamedan University of Technology, Hamedan, Iran, as an Assistant Professor. His research interests are mainly in wireless communications and networking, applications of machine learning in wireless communication networks, focusing on 5G/6G technology.

Reza Shahbazian (IEEE Senior Member) received his PhD in Computer Science from the University of Calabria, Italy. He previously worked as a postdoctoral researcher and assistant professor at the University of Calabria. He is currently an assistant professor at the University of Palermo. He is the author of more than 50 papers in prestigious international journals and conferences. His research interests include applications of machine learning, neural artificial intelligence, optimization, and signal processing.