ML-Based Downlink Throughput Prediction and Feature Importance Analysis in a Realistic LTE Testbed Using COTS Smartphone

Md Mahfuzur Rahman *Senior Member, IEEE*, Nishith Tripathi *Senior Member, IEEE*,
Cindy Yang Yi *Senior Member, IEEE*, Jeffrey H. Reed *Fellow, IEEE*, and Lingjia Liu *Fellow, IEEE*

***Abstract*—We propose a machine learning (ML)-based framework for downlink throughput prediction in LTE networks using real-time measurements from commercial off-the-shelf (COTS) user equipment (UE). The experimental platform integrates the srsRAN software stack deployed on a Dell desktop functioning as a Long Term Evolution (LTE) eNodeB, operating at 2.4 GHz. A Google Pixel 7a smartphone is used as the UE to collect physical layer features, including signal-to-noise ratio (SNR), channel quality indicator (CQI), modulation and coding scheme (MCS), and power headroom report (PHR). These features serve as predictors in model training. We evaluate the performance of three regression models — Linear Regression, Decision Tree Regression, and Random Forest Regression — using a supervised learning approach. The Random Forest model demonstrates superior performance, with a Root Mean Squared Error (RMSE) of 1.33 Mbps (without bandwidth), across a throughput range of 1–18 Mbps. Feature importance analysis reveals that SNR and PHR exhibit strong predictive relevance, underscoring their sensitivity to channel conditions and interference dynamics.**

***Index Terms*—Linear Regression, Decision Tree Regression, Random Forest Regression, channel quality indicator (CQI), bandwidth, modulation and coding scheme (MCS), power headroom report (PHR), signal-to-noise ratio (SNR), Google Pixel 7a, commercial off-the-shelf (COTS) UE.**

I. Background and Related Research

A. Introduction

The evolution of Long Term Evolution (LTE) networks has led to increased demands for intelligent resource management, particularly as mobile networks become more dynamic, data-intensive, and user-centric. With the growing emphasis on network automation and intelligence, there is an urgent need for experimental platforms that support data-driven optimization of LTE systems. This need has been addressed, in part, through the adoption of software-based radio access network stacks such as srsRAN and OpenAirInterface (OAI), which provide full-stack, open-source implementations of the LTE protocol stack compatible with standard 3GPP specifications [1], [2]. srsRAN has become particularly popular in academic and industrial research due to its ease of deployment, modular architecture, and seamless compatibility with commercial off-the-shelf (COTS) user equipment (UE) and software-defined radios (SDRs). These features enable realistic and cost-effective experimentation and provide researchers with the flexibility to configure LTE systems and collect reproducible performance metrics under various deployment scenarios [3], [4]. Unlike large-scale operator networks, srsRAN-based configurations allow fine-grained control over the physical and protocol layers, making them suitable for prototyping, diagnostics, and performance prediction.

As mobile data services continue to grow, accurately predicting LTE throughput based on real-time radio metrics has become crucial for achieving responsive scheduling, efficient link adaptation, and adaptive load balancing. LTE throughput is impacted by several interrelated physical-layer parameters, including signal-to-noise ratio (SNR), reference signal received power (RSRP), and modulation and coding scheme (MCS) levels [5], [6]. Capturing complex and often non-linear interactions among these variables requires sophisticated modeling approaches. To this end, machine learning (ML) methods have been applied to build predictive models capable of inferring downlink or uplink throughput using radio measurements collected from COTS smartphones. These models include Linear Regression, Decision Tree Regression, and Random Forest Regression, which are capable of learning from field measurements in real time [7], [8]. Random Forest Regression, in particular, has demonstrated strong performance in various LTE prediction tasks, including path loss estimation [9], signal strength mapping, and throughput inference.

Recent studies have confirmed the effectiveness of these models for LTE networks. For example, Rehmani et al. [8] applied ensemble learning techniques to real LTE datasets and reported high prediction accuracy, while Dias et al. [10] built LTE-specific models for predicting signal strength and planning coverage. Similar efforts by Koenig et al. [11] showed the applicability of such approaches across multiple RATs, including LTE and 5G.

This work builds on the aforementioned contributions by designing a lightweight, non-intrusive LTE measurement and prediction pipeline based entirely on open-source tools and COTS hardware. Our experimental platform uses srsRAN running on a high-performance desktop as the LTE eNodeB, interfaced with smartphones that operate as a UE. We collect physical-layer metrics such as channel quality indicator (CQI),

M. M. Rahman, N. Tripathi, C. Y. Yi, J. H. Reed, and L. Liu are with Wireless@Virginia Tech, Bradley Department of Electrical and Computer Engineering, Virginia Tech, VA, USA. This research is sponsored in part by the NSF CCRI project (award number 2235139). This research is also supported in part by the NSF RINGS project (award number: 2148212). Corresponding author's e-mail: mrahman2@vt.edu.

signal-to-noise ratio (SNR), modulation and coding scheme (MCS), and power headroom report (PHR) from YouTube traffic traces. These features are used to train predictive models that estimate downlink throughput without requiring active probing or external traffic generators. Our goal is not to modify the LTE air interface but to supplement it with predictive intelligence that can guide resource scheduling and admission control. The predictive models provide inference in real time and are particularly useful in Open RAN-based systems where decisions must be made rapidly based on fluctuating channel conditions. The proposed framework highlights the feasibility of integrating ML-based throughput prediction into LTE deployments using low-cost, reproducible setups. It also underscores the broader role of open-source platforms like srsRAN in fostering transparent, replicable wireless research. Future work may explore the integration of these models into RAN Intelligent Controllers (RICs) or xApps for closed-loop optimization in Open RAN environments.

B. Related Research

Several recent efforts have explored the application of ML techniques for estimating throughput in cellular networks, with a focus on leveraging radio metrics and environmental features. In one such study, Minovski et al. [12] utilized signal strength indicators, including RSRP, RSRQ, and SNR, to build predictive models for LTE and 5G throughput. Their measurements, collected in heterogeneous real-world contexts including subways and rural zones, were facilitated through commercial tools such as TEMS Pocket and a TWAMP-based custom probe. The study demonstrated the feasibility of prediction without dedicated testbed infrastructure. Raca et al. [13] presented a throughput prediction framework incorporating both traditional machine learning and deep learning techniques using metrics obtained via Android-based logging. Their evaluation—spanning Random Forest, support vector machine (SVM), and Long Short-Term Memory (LSTM) models—highlighted the robustness of LSTM models in dynamic environments, especially for adaptive video streaming applications. Another effort by Al-Thaedan et al. [14] collected LTE performance traces from three mobile operators and used statistical learning models including Support Vector Regression (SVR), K-th Nearest Neighbors (KNN), and Decision Tree Regression to correlate signal metrics and GPS data with downlink performance. They reported strong prediction fidelity from tree-based models, though their experiments were constrained to offline analysis of operator-generated data. In a large-scale mobile measurement campaign, Basit et al. [15] traversed 1000 kilometers using Android devices and Accuver tools, evaluating Multi-Layer Perceptron (MLP) and Gradient-Boosted Decision Tree (GBDT) models. Their findings revealed significant degradation in prediction accuracy for application level throughput, attributed to high-frequency variability and conflicting samples. Similarly, Eyceyurt et al. [16] conducted drive tests in multiple cities to develop uplink prediction models, noting that model generalization was highly sensitive to urban density and feature relevance, with KNN and Decision Tree outperforming others in certain regions. A report by Abdiel [17] and technical documentation from Clemson University [18] further examined the suitability of deep networks for modeling throughput dynamics, emphasizing their ability to learn nonlinear patterns but also pointing to the increased complexity and cost of real-time deployment.

These studies underscore the potential of ML for data rate prediction in mobile networks, but they overwhelmingly rely on datasets gathered under operator-controlled conditions or via uncontrolled field measurements. This restricts experimental flexibility and repeatability. Our work departs from this trend by employing a self-contained LTE platform powered by the open-source srsRAN stack and standard Google Pixel smartphones. This standalone setup allows precise control over the radio environment, reproducible testing, and tunable configuration of network parameters. Moreover, we emphasize model simplicity and efficiency, targeting interpretable algorithms such as linear and tree-based regressors that can be deployed in real time with minimal overhead. Unlike prior work that leans heavily on deep architectures, our lightweight design supports integration within constrained devices while maintaining predictive robustness. By embedding the learning pipeline within a software-defined LTE system, we establish a practical, controlled, and replicable framework for studying throughput prediction and optimization under modifiable conditions.

II. Motivation and Contribution

A. Motivation

The shift toward modular, intelligent, and software-defined mobile networks has positioned Open Radio Access Network (Open RAN) as a foundational architectural model for the future of cellular systems. By promoting disaggregation, open interfaces, and AI-native design, Open RAN aims to accelerate innovation and reduce vendor lock-in [19]. However, translating these principles into operational systems introduces a number of unresolved challenges—particularly the need for high-fidelity training data, non-disruptive testing environments, and realistic platforms for evaluating ML models. At the same time, 4G LTE networks continue to play a pivotal role in global connectivity, particularly in areas where 5G infrastructure is nascent or economically unfeasible. In these settings, accurately forecasting LTE throughput remains crucial for effective radio resource management and service optimization. Metrics such as CQI, MCS, PHR, and SNR are deeply intertwined with performance outcomes, but their interactions are inherently nonlinear and context-dependent, making them poorly suited to rule-based or analytical models.

Machine learning provides a compelling mechanism for learning these complex relationships from data. Yet, deploying and validating ML models in live commercial networks is fraught with limitations—including restricted access to internal parameters, lack of control over operating conditions, and risk of service disruption. These barriers underscore the need for configurable, low-cost LTE research platforms that enable controlled experimentation. Open-source projects such

as srsRAN, when coupled with COTS smartphones and SDRs, offer such a platform. They provide full-stack control and monitoring capabilities, facilitating experimentation across a wide range of radio conditions. Despite the availability of these tools, systematic investigations into ML-based throughput prediction using real-time COTS UE data in srsRAN-driven environments remain sparse. Our work addresses this gap by illustrating how open LTE testbeds can be leveraged for reproducible, data-centric experimentation, thereby contributing to both the Open RAN vision and the practical advancement of LTE network intelligence.

B. Contribution

To highlight the novelty and practical relevance of our study, we outline the main contributions as follows:

- **Low-Cost LTE Platform Using Open-Source Tools:** We present an LTE experimentation framework built using the srsRAN open-source stack and a Google Pixel 7a smartphone.
- **Lightweight Regression Models for Throughput Estimation:** A data-driven pipeline is developed for predicting LTE downlink throughput using real-time physical-layer metrics. The framework benchmarks three interpretable regressors — Linear, Decision Tree, and Random Forest — and highlights their trade-offs in terms of complexity and performance.
- **Insights from Data-Driven Feature Relevance:** We conduct a rigorous feature importance study to uncover which UE-side metrics most influence throughput prediction.
- **Prediction Accuracy Validated with Real Measurements:** Experimental results indicate that the Random Forest model offers the most robust performance, achieving an RMSE of 1.33 Mbps (without bandwidth) accross a test range spanning from 1 to 18 Mbps.

III. Regression Model Formulations

We frame LTE downlink throughput estimation as a supervised regression task using physical-layer features collected from COTS UEs. Each input vector is defined as:

$$\mathbf{x} = [\text{CQI}, \text{SNR}, \text{PHR}, \text{MCS}, \text{BW}], \quad y = \text{throughput}$$

The goal is to learn a mapping $\hat{y} = f(\mathbf{x})$ that minimizes the prediction error. We evaluated three models, Linear Regression, Decision Tree Regression, and Random Forest Regression, selected for their trade-offs between complexity, interoperability, and accuracy.

A. Regression Models

Linear Regression assumes a linear dependency between input and output, optimized via least squares:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \tag{1}$$

Decision Tree Regression partitions the input space based on feature thresholds, fitting a constant output in each region to minimize intra-node variance:

$$\min_{j,t} \left(\sum_{x_i \in R_1(j,t)} (y_i - \bar{y}_{R_1})^2 + \sum_{x_i \in R_2(j,t)} (y_i - \bar{y}_{R_2})^2 \right) \tag{2}$$

Random Forest Regression constructs an ensemble of B randomized decision trees and averages their predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}) \tag{3}$$

Random Forest Regression helps reduce overfitting and models nonlinearities, while offering feature importance measures.

B. Data Normalization and Evaluation

All features are standardized using Z-score normalization to zero mean and unit variance, implemented with `StandardScaler` from `scikit-learn`, to improve learning stability. We assess models using three standard metrics:

- **Mean Squared Error (MSE):**

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

- **Root Mean Squared Error (RMSE):**

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

- **Coefficient of Determination (R^2):**

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

These metrics evaluate accuracy (RMSE), sensitivity to large errors (MSE), and model explanatory power (R^2). The importance of the Random Forest feature further reveals which physical parameters influence the most throughput, enhancing the interpretability of the results.

IV. System Model and Data Collection

We developed a custom 4G LTE testbed using the srsRAN open-source stack to enable controlled and repeatable experimentation in an Open RAN-compliant environment. The eNB and EPC functions were deployed on a Dell desktop (Intel i7, 32 GB RAM) running Ubuntu 22.04 LTS with a real-time kernel to support low-latency, over-the-air operations. A USRP B210 [20] software-defined radio was used as the RF front-end, configured at 2.4 GHz for indoor propagation. A Google Pixel 7a smartphone, operating on Android 14 and configured using OpenCells guidelines [21], acted as the UE. Ubuntu terminal captured key physical-layer metrics—including SNR, CQI, MCS, PHR, and throughput—which served as inputs to

the ML models. YouTube data was collected across multiple locations within the lab to induce varying link conditions. All logs were cleaned to discard invalid entries, and input features were normalized using Z-score scaling before training and evaluation.

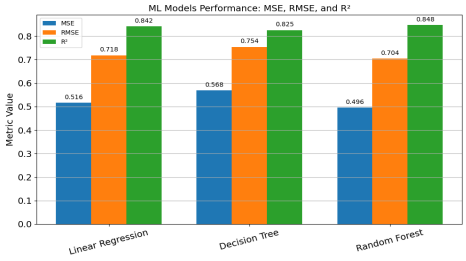
Fig. 1: Realistic LTE testbed with a smartphone

V. Result and Analysis

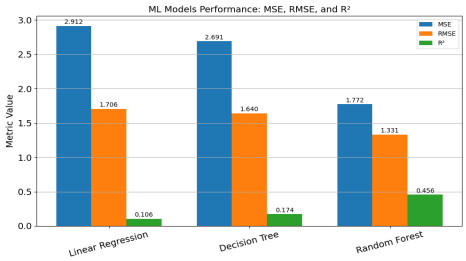
The performance of the three machine learning models — Linear Regression, Decision Tree Regression, and Random Forest Regression — was evaluated using the test dataset. The models were compared using MSE, RMSE, and R^2 score. The results are presented in Table I, and a grouped bar chart visualization is shown in Fig. 2 for clarity. Fig. 2 compares

TABLE I: ML Models Comparison

Model	MSE ($Mbps^2$)	RMSE (Mbps)	R^2 Score
Linear Regression	0.516 (a),2.912 (b)	0.718 (a), 1.706 (b)	0.842 (a), 0.106 (b)
Decision Tree	0.568 (a), 2.691 (b)	0.754 (a), 1.640 (b)	0.825 (a), 0.174 (b)
Random Forest	0.496 (a), 1.772 (b)	0.704 (a), 1.331 (b)	0.848 (a), 0.456 (b)

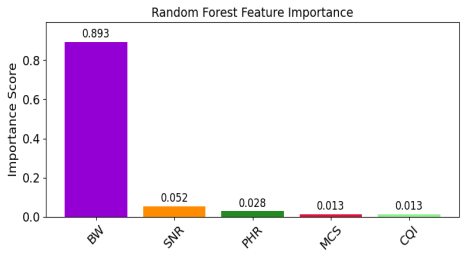


(a) With bandwidth feature

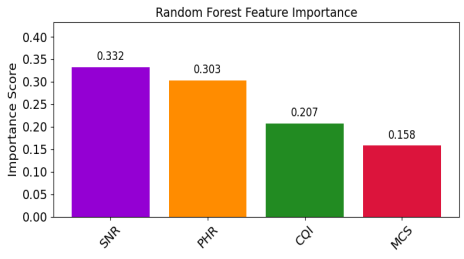


(b) Without bandwidth feature

Fig. 2: Comparison of model performance



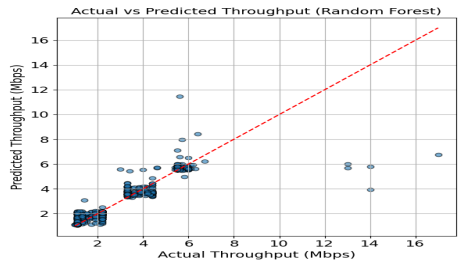
(a) With bandwidth



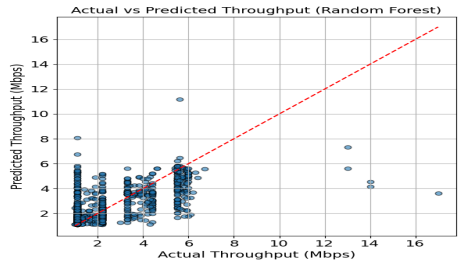
(b) Without bandwidth

Fig. 3: Feature importance from Random Forest

error. Feature ranking (Fig. 3) confirms bandwidth and SNR as the top predictors. Their strong contribution validates the importance of resource allocation in LTE throughput and motivates bandwidth-aware adaptation strategies. The prediction



(a) With bandwidth

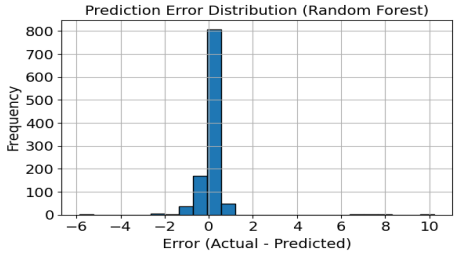


(b) Without bandwidth

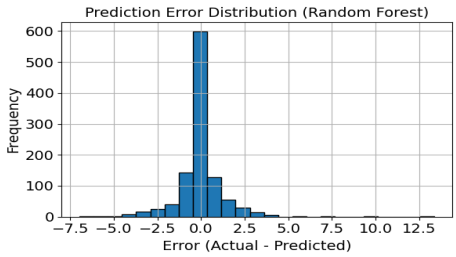
Fig. 4: Actual vs. predicted throughput

the performance of the regression model with and without bandwidth as one of the input features. Including bandwidth (2a) significantly improves all metrics—particularly for Random Forest, which achieves RMSE ≈ 0.70 and $R^2 \approx 0.85$. Excluding it (2b) degrades performance, with Linear and Decision Tree models showing $R^2 < 0.25$ and increased

scatterplots in Fig. 4 show better alignment with the ground truth when bandwidth is used, with tighter clustering around the ideal diagonal. Without bandwidth, predictions deviate more—especially at higher throughputs—highlighting limited generalization. Fig. 5 illustrates the effect of bandwidth on error distribution. With bandwidth, errors are tightly centered



(a) With bandwidth



(b) Without bandwidth

Fig. 5: Prediction error distribution

around zero, indicating high confidence. Without it, errors are more dispersed and skewed, with deviations exceeding ± 5 Mbps—suggesting reduced reliability. Overall, adding bandwidth as a feature leads to 20%–30% RMSE improvement and boosts R^2 from 0.46 to 0.85, demonstrating its critical role in throughput prediction. These findings support the design of lightweight, feature-aware schedulers that prioritize bandwidth, SNR, and PHR for QoS optimization.

VI. Conclusion

This work introduced a machine learning framework for downlink throughput prediction in LTE networks using measurements from a Google Pixel smartphone connected to an srsRAN-based testbed. The dataset captured key radio metrics (CQI, SNR, MCS, PHR, bandwidth) under varied link conditions, including mobility and non-line-of-sight scenarios. We compared three regression models—Linear, Decision Tree, and Random Forest—using standard error metrics. Random Forest achieved the best performance in both scenarios, effectively capturing nonlinear feature interactions. Feature ranking highlighted SNR and bandwidth as dominant predictors, reinforcing their role in dynamic link adaptation. The proposed approach relies solely on accessible UE-side features, enabling practical, real-time quality of service estimation. Future directions include joint modeling of throughput and error estimation, cross-environment validation, and adaptation to 5G systems.

References

[1] Software Radio Systems (SRS), *srsRAN 4G Documentation*, 2024, [Online; accessed May 6, 2025]. [Online]. Available: https://docs.srsran.com/projects/4g/en/latest

[2] OpenAirInterface Software Alliance, "OpenAirInterface 5G Wireless Implementation," https://www.openairinterface.org/, 2024, accessed: 2025-05-07.

[3] R. P. Alves, J. G. A. da Silva Alves, M. R. Camelo, W. O. de Feitosa, V. F. Monteiro, and F. R. P. Cavalcanti, "Experimental Comparison of 5G SDR Platforms: srsRAN x OpenAirInterface," *arXiv preprint arXiv:2406.01485*, 2024. [Online]. Available: https://arxiv.org/abs/2406.01485

[4] X. Ye *et al.*, "Design and Performance Study of 4G Communication System Based on srsRAN," in *Proc. SPIE 12474, Sixth International Conference on Photonics and Optical Engineering*, 2022, p. 124742B. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12474/124742B/Design-and-performance-study-of-4G-communication-system-based-on/10.1117/12.2653437.full

[5] A. Teixeira and J. Timóteo, "A Predictive Resource Allocation for Wireless Communications Systems," *SN Computer Science*, vol. 2, no. 2, pp. 1–14, 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00854-8

[6] H. Perveen, M. Zafar, S. Abbas, S. Rehman, I. Ahmad, and M. Rathore, "Dynamic Traffic Forecasting and Fuzzy-Based Optimized Admission Control in Federated 5G-Open RAN Networks," *Neural Computing and Applications*, vol. 34, pp. 2925–2940, 2022. [Online]. Available: https://doi.org/10.1007/s00521-021-06206-0

[7] H. Elsherbiny and M. Abbas, "4G LTE Network Throughput Modelling and Prediction," *Queen's Telecommunications Research Lab*, 2020.

[8] A. Rehmani *et al.*, "Machine Learning for Wireless Network Throughput Prediction," *ScholarWorks @ UTRGV*, 2023.

[9] S. Chang and A. Baliga, "Development of machine learning-based radio propagation models and benchmarking for mobile networks," *J. Stud. Res*, vol. 10, pp. 1–12, 2021.

[10] T. Dias, A. Oliveira, L. Gonçalves, and J. Martins-Filho, "RSRP Prediction on LTE Network Testbed Using a Software Defined Radio Platform," in *XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, 2022. [Online]. Available: https://biblioteca.sbrt.org.br/articlefile/3582.pdf

[11] N. Koenig *et al.*, "Throughput Prediction Using Machine Learning in LTE and 5G Networks," https://people.computing.clemson.edu/~jmarty/projects/lowLatencyNetworking/papers/AI-ML/MLAppliedToNetworks/Throughput_Prediction_using_Machine_Learning_in_LTE_and_5G_Networks.pdf, 2021.

[12] D. Minovski, N. Ögren, K. Mitra, and C. Åhlund, "Throughput Prediction Using Machine Learning in LTE and 5G Networks," *IEEE Transactions on Mobile Computing*, vol. 22, no. 3, pp. 1825–1840, 2023.

[13] D. Raca, A. H. Zahran, C. J. Sreenan, R. K. Sinha, E. Halepovic, R. Jana, and V. Gopalakrishnan, "On Leveraging Machine and Deep Learning for Throughput Prediction in Cellular Networks: Design, Performance, and Challenges," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 11–17, 2020.

[14] A. Al-Thaedan, Z. Shakir, A. Y. Mjhool, R. Alsabah, A. Al-Sabbagh, M. Salah, and J. Zec, "Downlink throughput prediction using machine learning models on 4G-LTE networks," *International Journal of Information Technology*, vol. 15, no. 6, pp. 2987–2993, 2023.

[15] O. Basit, P. Dinh, I. Khan, Z. J. Kong, Y. C. Hu, D. Koutsonikolas, M. Lee, and C. Liu, "On the Predictability of Fine-Grained Cellular Network Throughput Using Machine Learning Models," in *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, 2024, pp. 47–56.

[16] E. Eyceyurt, Y. Egi, and J. Zec, "Machine-learning-based uplink throughput prediction from physical layer measurements," *Electronics*, vol. 11, no. 8, p. 1227, 2022.

[17] E. Abdiel, "Forecasting LTE Network Throughput for Optimizing Operational and Business Aspects," https://medium.com/@earlyanabdiel/forecasting-lte-network-throughput-for-optimizing-operational-and-business-aspect-a9599a565d6a, 2022.

[18] Clemson University, "Machine Learning and Deep Learning for Throughput Prediction," 2022. [Online]. Available: https://people.computing.clemson.edu/~jmarty/projects/lowLatencyNetworking/papers/AI-ML/MLAppliedToNetworks/Machine_Learning_and_Deep_Learning_for_Throughput_Prediction.pdf

[19] N. D. Tripathi and V. K. Shah, *Fundamentals of O-RAN*. John Wiley & Sons, 2025.

[20] Ettus Research, "Universal Software Radio Peripheral (USRP)," https://www.ettus.com/, accessed: 2025-05-07.

[21] Open Cells Project, "Open Cells Project: 4G and 5G Open-Source Testbed Platform," https://open-cells.com/, accessed: 2025-05-07.