Predictive QoS Management in Vehicular Networks Using Advanced Machine Learning Techniques

Aashna Kunkolienker
Department of Computer Engineering
New York University
Email: ank8919@nyu.edu

***Abstract*—This report addresses the challenges of predicting Quality of Service (QoS) in vehicular networks, where dynamic environments, high mobility, and variable communication conditions complicate reliable performance. Traditional methods, such as signal propagation models and network simulations, are often effective in static scenarios but might lack adaptability in real-time, dynamic vehicular contexts. To tackle these limitations, we leverage advanced machine learning models like tree-based algorithms, to predict key QoS metrics such as uplink throughput using features like Signal-to-Interference-plus-Noise Ratio (SINR), and Reference Signal Received Power (RSRP). Using a dataset collected in a real-world vehicular environment, our implementation showcases significant results. Practical use cases, such as efficient route planning, coordinated fleet vehicle operations, and optimized communication strategies, highlight the potential applications of this work. Finally, we propose future directions, including the use of transfer learning to adapt these models to diverse global environments, with the ultimate goal of creating scalable and robust QoS prediction systems.**

I. Introduction

The rapid evolution of wireless communication networks, particularly with the advent of 5G, has introduced new challenges in maintaining and enhancing Quality of Service (QoS) for end-users. Efficient network management requires predicting network conditions to dynamically optimize resource allocation and maintain consistent service levels. Among emerging applications, vehicular communication systems, especially Vehicle-to-Everything (V2X) communications, stand to benefit significantly from these advancements due to their stringent QoS demands. V2X involves communication between vehicles and infrastructure, other vehicles, and networks, requiring reliable and high-throughput connectivity to support efficient coordination and data exchange.

This report explores the challenge of predicting QoS in a vehicular communication scenario within the context of 5G networks. Using machine learning (ML) techniques, we focus on predicting downlink throughput, a key QoS metric, to enable efficient route planning and coordination among vehicles in a fleet. Our study leverages the Berlin V2X dataset, which includes LTE signal metrics, GPS data, and additional contextual information collected from primary and secondary cells of two LTE network operators in an urban environment. We aim to demonstrate how AI-driven models can provide accurate and adaptive solutions for QoS prediction, offering practical advantages over traditional methods.

II. Traditional vs AI-Based Methods for QoS Prediction

Traditional methods for QoS prediction, including signal propagation models like Okumura-Hata and FSPL, network simulations using tools such as NS-3 and OMNeT++, and time-series forecasting techniques like Kalman filters, rely heavily on predefined parameters and assumptions. These approaches are effective in controlled environments but often lack adaptability to the dynamic and non-linear nature of vehicular networks.

In contrast, machine learning (ML) offers data-driven solutions that can learn complex patterns and adapt to changing conditions without relying on rigid assumptions. Supervised learning models, such as Random Forests and XGBoost, handle tabular data effectively while providing feature importance insights. Neural networks, including LSTMs, excel at capturing temporal dependencies in time-series data. Furthermore, techniques like transfer learning and reinforcement learning enable models to generalize across diverse scenarios and optimize real-time decision-making. The versatility and scalability of ML make it a preferred approach for robust QoS prediction in dynamic vehicular networks.

III. Data Collection and Preparation

Data Collection and Preparation: For training AI models for QoS prediction, high-resolution datasets such as the Berlin V2X dataset are indispensable. This dataset provides extensive information, including primary and secondary cell metrics, GPS data, environmental factors, and operator information, making it well-suited for tasks like downlink throughput prediction. Additional data from controlled measurement campaigns, such as those conducted in test networks or motorway environments, can further enhance model robustness and generalizability.

Dataset Segregation: To facilitate analysis, we separated the dataset into two distinct subsets:

- **P-Cell Only Dataset:** This subset contains entries where secondary cell (SCell) data is absent. These scenarios often occur in areas with limited network infrastructure or where carrier aggregation is not active.
- **P-Cell + S-Cell Dataset:** This subset includes entries with both primary (PCell) and secondary (SCell) data. The presence of SCell metrics indicates scenarios with

active carrier aggregation, typically observed in high-demand environments or regions with strong network coverage.

Reasoning: This segregation was performed to evaluate the distinct impact of carrier aggregation on QoS prediction. The P-Cell Only dataset allows us to analyze the network's baseline performance, while the P-Cell + S-Cell dataset highlights the enhancements achieved through carrier aggregation. Such separation is crucial for understanding the varying contributions of network configurations and ensuring that the models are accurately tailored to different operational conditions.

Data Preprocessing: Preprocessing steps were essential to prepare the datasets for effective model training. These steps included:

- **Feature Selection:** Selecting relevant features, such as signal quality metrics (RSRP, SINR), environmental parameters (traffic conditions, weather), and location data (GPS coordinates), to ensure model inputs are meaningful and concise.
- **Normalization:** Scaling features to ensure uniformity and prevent dominance of features with larger numerical ranges during model training.
- **Handling Missing Values:** Employing strategies like imputation or removal to address missing data points, ensuring data quality and consistency.

By carefully organizing and preprocessing the data, we ensured that the models could effectively learn from and generalize across the diverse conditions represented in the dataset.

A. Why AI Solutions Address the Issues in Traditional Methods

While traditional approaches like signal propagation models, network simulations, and empirical measurements provide foundational insights into QoS prediction, their limitations in dynamic vehicular environments highlight the need for AI-driven solutions. In this subsection, we discuss how AI methods overcome these challenges and address the trade-offs associated with their use.

1. Adaptability to Dynamic Environments: Traditional methods, such as signal propagation models and network simulations, rely on predefined rules or scenarios, making them ill-suited for handling the highly dynamic and unpredictable nature of vehicular networks. In contrast, AI models excel at:

- Learning from diverse datasets that include real-world variations (e.g., urban, rural, high-speed).
- Adapting to unseen scenarios by capturing complex patterns and relationships in the data.

Trade-off: AI models require a sufficiently large and representative dataset to generalize well. However, once trained, these models dynamically adapt to changes in the environment with far less recalibration than traditional approaches.

2. Handling Non-Linear Interactions: High-mobility vehicular networks involve non-linear interactions between variables like signal strength, interference, and vehicle velocity. While traditional methods (e.g., Kalman filters or ARIMA) fail to capture these complexities, AI-based approaches, particularly deep learning, are designed to:

- Model non-linear relationships effectively using neural network architectures.
- Incorporate sequential patterns through models like LSTMs and Transformers, which are ideal for time-series data.

Trade-off: Deep learning solutions are computationally intensive during training, but their ability to reduce prediction error significantly justifies the investment in scenarios where accuracy is critical.

3. Resource Efficiency in Real-Time Prediction: Although traditional network simulations are resource-intensive, they primarily provide insights for fixed scenarios and cannot generalize to real-time environments. AI solutions address this by:

- Shifting computational cost to the training phase, allowing for fast and efficient real-time inference.
- Leveraging lightweight deployment options like TensorFlow Lite for edge devices, enabling predictions directly within vehicles or edge nodes.

Trade-off: While training AI models can be resource-heavy, they deliver near-instantaneous predictions post-training, making them more practical for real-time vehicular applications compared to recalibrating traditional simulations repeatedly.

4. Scalability Across Diverse Scenarios: Traditional methods require significant recalibration to account for new environments (e.g., urban vs. highway scenarios). AI models overcome this by:

- Using transfer learning to adapt pre-trained models to new datasets with minimal effort.
- Generalizing well across diverse conditions, provided the training dataset is representative.

Trade-off: AI models depend heavily on the quality and diversity of training data. Poorly curated datasets may lead to overfitting or biased predictions, but this can be mitigated with proper data preprocessing and augmentation techniques.

5. Improved Accuracy and Predictive Power: The primary advantage of AI solutions lies in their predictive accuracy. Compared to empirical models or time-series methods, AI approaches can:

- Achieve significantly lower error rates by leveraging advanced architectures like XGBoost, LSTMs, and Transformers.
- Incorporate a broader range of features, such as SINR, RSRP, and trajectory data, to improve QoS predictions.

Trade-off: While the accuracy gain is substantial, it may not always justify the computational overhead for less critical applications. However, in safety-critical scenarios like autonomous driving, the accuracy gain outweighs the cost.

6. Flexibility for Continuous Learning: AI solutions allow for incremental updates through continuous learning techniques. Unlike traditional methods that require manual recalibration, AI models can:

- Retrain periodically with new data to account for changing network conditions.
- Employ online learning for real-time adaptation in dynamic environments.

Trade-off: Continuous learning increases operational complexity, requiring periodic data collection and monitoring pipelines. However, this ensures the models remain robust and up-to-date.

Conclusion: While AI solutions come with their own resource requirements, the trade-off is clear: they provide adaptability, scalability, and significantly improved accuracy in dynamic vehicular environments. For applications like autonomous driving, where QoS predictions are safety-critical, the advantages of AI outweigh the costs. By leveraging AI models, we achieve not only better predictions but also the flexibility to deploy these solutions in real-world vehicular networks effectively.

AI-based solutions offer a more flexible and adaptive approach to QoS prediction in vehicular networks, with the potential to make real-time adjustments and proactive decisions. These models are better suited for the dynamic, high-mobility environments encountered in vehicular communications, offering improvements in accuracy and reliability over traditional methods.

IV. Network Description

In this study, we assume a network environment based on LTE cellular infrastructure operating in a metropolitan area. The dataset contains measurements from primary and secondary cells of two commercial LTE operators in Berlin. Each vehicle in the dataset serves as a user equipment (UE) connected to the network, collecting information about downlink quality metrics such as signal strength, signal-to-noise ratio (SNR), and received signal strength indicator (RSSI), among others. These metrics are recorded along with GPS coordinates, environmental factors, and traffic data, allowing us to analyze network quality under varying conditions.

A. Primary and Secondary Cells in LTE Networks

In Long-Term Evolution (LTE) networks, carrier aggregation is a key feature designed to enhance network throughput and reliability by combining multiple frequency bands. This process involves the use of **Primary Cells (PCell)** and **Secondary Cells (SCell)**, which together form a more robust communication channel.

Primary Cell (PCell):

- The PCell is the main connection point between the user equipment and the LTE network.
- It operates on the primary carrier frequency and handles critical control plane functions, such as authentication, mobility management, and signaling.
- The PCell is always active and is responsible for establishing and maintaining the initial connection when a device enters the network or moves between cells.
- This connection ensures the baseline level of communication required for reliable data exchange, even in the absence of secondary cells.

Secondary Cell (SCell):

- The SCell is an auxiliary connection that operates on additional carrier frequencies. It is dynamically added or removed based on network conditions, user demand, and device capabilities.
- SCells are primarily used to enhance data throughput by offloading traffic from the PCell and providing additional bandwidth.
- Unlike the PCell, the SCell may not be continuously active and does not typically handle control plane signaling. Its role is focused on the user plane, contributing directly to data transmission. This feature of Scell is the reason we have so much missing data in our Berlin V2X dataset.

Interaction Between PCell and SCell:

- Together, the PCell and SCell enable carrier aggregation, which is crucial for meeting the high data rate demands of modern applications.
- The PCell serves as the anchor, managing critical network tasks, while the SCell supplements it with additional resources, thereby boosting the overall network capacity and reliability.
- Carrier aggregation is especially beneficial in scenarios with high data traffic, such as streaming video, real-time gaming, or handling multiple simultaneous connections.

B. Network Environment and Impact on Performance

The environment in which the PCell and SCell operate significantly impacts their performance. LTE networks are characterized by:

- **High Mobility:** Vehicles traveling at varying speeds lead to frequent transitions between cells, known as "handovers". Effective coordination between PCell and SCell is crucial to maintaining seamless connectivity during these transitions.
- **Frequent Cell Handovers:** As the user equipment moves across geographical regions, the network dynamically assigns new PCells and SCells to ensure optimal performance.
- **Diverse Geographical Locations:**
  - *Urban Areas:* High building density causes signal reflections and obstructions, impacting both PCell and SCell connections.
  - *Highways:* High mobility introduces challenges in maintaining consistent connections, but fewer obstructions allow for better line-of-sight communication.
  - *Tunnels:* Signal degradation and limited coverage in enclosed spaces make carrier aggregation critical for maintaining communication.

Conclusion: The combined functionality of PCell and SCell ensures both reliability and high throughput in LTE networks,

particularly in challenging environments with diverse geographical and mobility conditions. By leveraging carrier aggregation, LTE networks can provide seamless communication, even in scenarios with high mobility and complex physical conditions.

V. Literature Review

A. Berlin V2X: A Machine Learning Dataset from Multiple Vehicles and Radio Access Technologies

One of the key references for this project is the paper titled "Berlin V2X: A Machine Learning Dataset from Multiple Vehicles and Radio Access Technologies," which introduces the Berlin V2X dataset. This dataset provides high-resolution measurements of QoS metrics collected across various environments in Berlin, including avenues, parks, highways, residential streets, and tunnels. The dataset encompasses features from LTE networks across two operators, capturing signal strength, SNR, RSRP, RSRQ, and downlink throughput for both primary and secondary cells.

The paper highlights the challenges and opportunities associated with applying machine learning for QoS prediction in a vehicular setting, where high mobility and diverse conditions impact network performance. The authors also present preliminary analyses, including correlation patterns across different geographical areas and network conditions, demonstrating the dataset's suitability for QoS prediction tasks. This dataset, therefore, provides a foundational resource for developing predictive models that generalize across different LTE operators and environments.

B. QoS Prediction in Radio Vehicular Environments via Prior User Information

Another relevant paper is titled "QoS Prediction in Radio Vehicular Environments via Prior User Information," which explores the concept of predictive QoS (pQoS) in vehicular networks. This study emphasizes the importance of reliable QoS for emerging vehicular use cases such as connected autonomous driving, platooning, and teleoperated driving, which rely on uninterrupted connectivity. To address the challenge of fluctuating QoS in high-mobility environments, the authors evaluate machine learning tree-ensemble methods for QoS prediction, using data from the AI4Mobile measurement campaign on the A9 Motorway in Germany.

The study proposes using information from prior vehicles to enhance the QoS prediction accuracy for target vehicles. By leveraging the correlations between the radio environment characteristics experienced by preceding vehicles, the model improves its predictive performance for vehicles following the same route. The authors demonstrate that including features from leading vehicles' physical (PHY) layer measurements, such as SNR, RSRP, and RSSI, significantly reduces prediction error, especially over longer look-ahead times.

C. Predictive QoS (PQoS): The Next Frontier for Fully Autonomous Systems

The paper titled *"Predictive QoS (PQoS): The Next Frontier for Fully Autonomous Systems"* discusses the role of predictive QoS (PQoS) in ensuring reliable and efficient communication in autonomous and vehicular systems. PQoS enables proactive adaptation to network changes by predicting QoS metrics like throughput and latency, thereby ensuring continuous operation for critical use cases such as teleoperated driving, autonomous platooning, and high-definition map sharing.

The authors highlight how machine learning, particularly deep neural networks (DNNs), can significantly improve prediction accuracy compared to traditional methods like linear regression. A case study demonstrates that SINR (Signal-to-Interference-plus-Noise Ratio) is a dominant feature for throughput prediction, achieving error rates as low as 4%. The paper also explores the use of transfer learning to adapt pre-trained models to new environments, emphasizing the potential for scalable, location-agnostic implementations.

These insights can be applied to this project by leveraging SINR as a key feature in machine learning models and exploring transfer learning for adapting predictions to diverse geographical regions and vehicular scenarios. This proactive approach can mitigate connectivity disruptions and improve QoS reliability in dynamic network environments.

VI. My Approach

To tackle the problem of predicting QoS (Quality of Service) in vehicular networks, I have developed a structured approach involving data understanding, feature engineering, model selection, training, and evaluation. This section details the methodology and justifies the chosen techniques, models, and dataset.

A. Understanding the Dataset

The dataset utilized in this study is derived from real-world vehicular environments, capturing a wide range of network and environmental metrics essential for Quality of Service (QoS) prediction. The key features include:

- **Downlink Throughput:** The target metric representing the quality of communication, indicative of the data rate achievable at the receiver.
- **Signal-to-Interference-plus-Noise Ratio (SINR):** A measure of signal quality relative to background noise and interference, crucial for assessing network reliability and performance.
- **Reference Signal Received Power (RSRP):** Represents the average received power of the reference signals, providing insights into signal strength at the receiver.
- **Reference Signal Received Quality (RSRQ):** Captures the quality of the received reference signal, complementing RSRP to provide a holistic view of signal conditions.
- **Uplink Resource Block Usage (RB Usage):** Indicates network resource utilization, reflecting the load on the uplink channel.
- **Environmental Parameters:** Features such as traffic congestion levels, weather conditions (e.g., cloud cover, visibility), and time-of-day to contextualize variations in network performance.

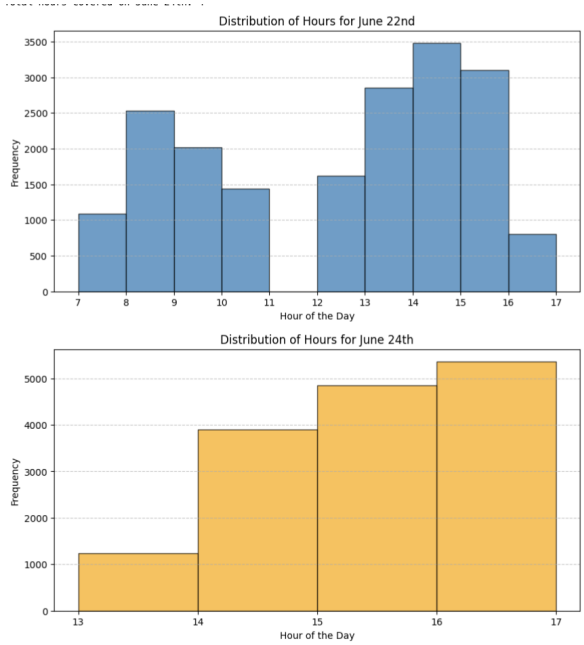

Fig. 1. Data Distribution Across Hours for June 22nd and June 24th

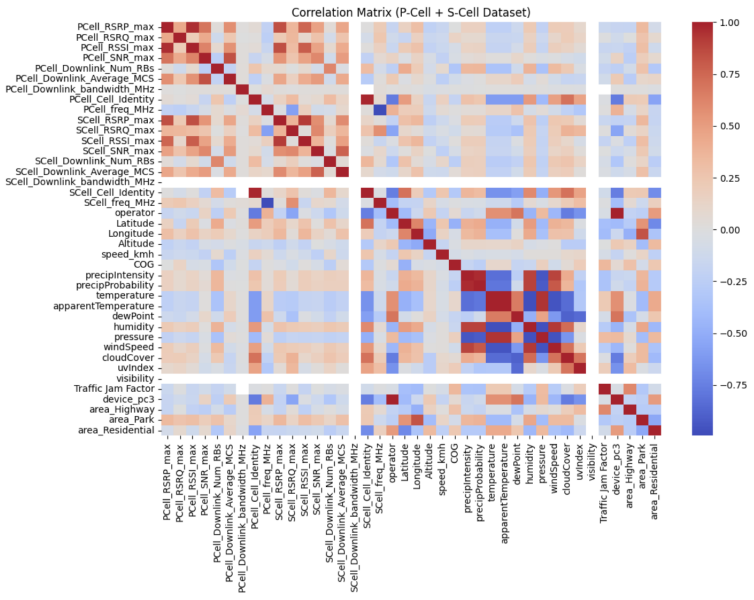

Fig. 2. Correlation Heatmap for Feature Selection.

- **Location Data:** Comprises longitude, latitude, and altitude, enabling spatial mapping of network variations and signal coverage.
- **Vehicle Velocity:** Represents the speed of the vehicle, capturing the dynamic mobility aspects that influence network behavior and QoS.
- **Primary and Secondary Cell Metrics:** Metrics from both primary (PCell) and secondary (SCell) cells, including carrier aggregation states, to study their individual and combined impact on network performance.

B. Understanding Timestamps and Data Collection

The dataset spans only two days, specifically June 22nd and June 24th. This limited timeframe necessitated careful handling of the data to ensure meaningful training and testing splits.

Figure 1 illustrates the distribution of data samples collected over the two days, grouped by hour. The data collection pattern shows that certain hours have significantly higher coverage, which allowed for informed decisions about splitting training and testing datasets while maintaining the temporal integrity of the data.

C. Feature Engineering

Feature engineering is a critical step to enhance the predictive power of machine learning models by refining and selecting the most relevant features. The following steps were undertaken:

- **Correlation Analysis and Feature Selection:** A correlation heatmap was generated to analyze the relationships between features and the target variable (downlink throughput). Features with very low correlation to the target were removed to simplify the model and reduce noise. The features removed included:
  - *Humidity:* Correlation = 0.230256
  - *Timestamp:* Correlation = 0.200607
  - *Dew Point:* Correlation = -0.244297
  - *Temperature:* Correlation = -0.248663
  - *Visibility:* Correlation = NaN

  Features such as SINR, RSRP, downlink resource block usage, and location coordinates were retained due to their higher correlation and relevance to QoS prediction.
- **Categorical Encoding:** To enable machine learning algorithms to process categorical data effectively:
  - **Device:** Device IDs were one-hot encoded.
  - **Operator:** The operator identifier (1 or 2) was one-hot encoded to represent the network operator.
  - **Area Type:** Categories such as residential, park, and work were also one-hot encoded to indicate the type of environment.
- **Timestamp Processing:** Timestamps were converted into `datetime` format to facilitate temporal analysis and ensure compatibility with downstream processing steps. This also allowed better separation of training and testing data based on time.
- **Handling Missing or Noisy Data:** Missing values in the dataset were handled using appropriate imputation techniques to maintain data consistency. Outliers were identified and removed to ensure model stability and avoid skewing predictions.
- **Derived Features:** Additional features were derived to enhance the model's predictive capabilities:
  - *Normalized Features:* SINR and RSRP values were normalized to ensure comparability across different scales.

.

D. Model Selection and Justification

Given the dataset's structure and problem complexity, I will use a combination of traditional tree-based models and neural network-based approaches:

E. Modeling Approaches

In this study, we explored two key modeling approaches for Quality of Service (QoS) prediction:

- **ARIMA (AutoRegressive Integrated Moving Average):**
  - **Description:** ARIMA is a traditional time-series forecasting method that models temporal dependencies by combining auto-regression, differencing, and moving averages to capture linear patterns in sequential data.
  - **Performance:** Despite extensive tuning, ARIMA performed poorly in this study, with a Mean Absolute Percentage Error (MAPE) of 54.43% and an R^2 score of -0.4944. These metrics indicate significant inaccuracies, likely due to the limited range of data available (only two days), which constrained the model's ability to detect meaningful trends. Given these results, ARIMA was deemed unsuitable for this analysis, emphasizing the need for alternative approaches.

.

- **XGBoost (Extreme Gradient Boosting):**
  - **Description:** XGBoost is a highly efficient tree-based ensemble learning algorithm specifically designed for tabular data. It uses gradient boosting frameworks to optimize predictions by iteratively improving weak learners (decision trees).
  - **Advantages:**
    - Handles missing values natively.
    - Strong regularization techniques prevent overfitting.
    - Fast training times and interpretable results make it ideal for medium-sized datasets.
  - **Performance:** XGBoost outperformed ARIMA in this study, demonstrating its suitability for the tabular and multi-dimensional nature of the dataset. Its ability to handle non-linear relationships and interactions between features made it a robust choice for predicting downlink throughput.
  - **Final Decision:** Based on its superior performance and compatibility with the dataset, XGBoost was selected as the primary modeling approach for this project.

A hybrid approach may also be explored by combining the strengths of both techniques to achieve better performance.

F. Model Training and Evaluation

The training and evaluation process in this study was meticulously designed to ensure robust model performance while addressing the unique challenges of predicting Quality of Service (QoS) in a vehicular communication scenario. Below, we detail the methodologies employed, the rationale behind key decisions, and the insights derived from the evaluation process.

G. Data Splits and Evaluation Metrics

To evaluate model performance under varying conditions, we experimented with different training and testing periods using timestamps spanning two days of data collection. Specifically, data from the first day (June 22nd) provided a 9-hour training window, while the second day (June 24th) offered a 4-hour testing window.

- **Testing Data:** Data from June 24th, after the cutoff time (1:00 PM to 4:00 PM), was reserved for testing. This temporal split simulated a real-world scenario where the model predicts future QoS conditions based on past data.
- **Tradeoff in Testing Duration:** Initially, multiple configurations were tested with varying durations for training and testing:
  - Longer testing windows (e.g., 6 hours) provided more data but reduced accuracy due to greater variability in network conditions.
  - Shorter testing windows (e.g., 1 hour) improved accuracy but offered less insight into model generalizability.
- **Final Configuration:** A 3-hour testing window was selected as a balance between accuracy and representativeness. This duration allowed us to evaluate model performance across meaningful variations in QoS conditions while maintaining high prediction accuracy.

	Training Hours	Testing Hours	P-Cell R^2	P+S-Cell R^2
0	9 (Day 1)	0 (Day 2)	0.708	0.7583
1	9 (Day 1) + 1 (Day 2)	3 (Day 2)	0.708	0.7583
2	9 (Day 1) + 2 (Day 2)	2 (Day 2)	0.7932	0.7292
3	9 (Day 1) + 3 (Day 2)	1 (Day 2)	0.8203	0.8431

Fig. 3. Training and Testing Splits with Evaluation Metrics

H. Sliding Window Analysis

To assess the consistency of the model across different timeframes, we implemented a sliding window analysis:

- **Rationale:** Sliding windows were employed to examine how model performance varied when trained and tested on different temporal segments of the dataset. This approach helped us understand whether the model's predictions were stable or heavily dependent on specific time periods.
- **Methodology:**

  - Training and testing windows were shifted incrementally, covering different hours on June 22nd and 24th.
  - For each window, the model was trained on a selected set of hours and tested on subsequent hours.
  - Metrics such as Mean Squared Error (MSE) were recorded for each window.
- **Results:** The sliding window analysis revealed notable variability in model performance:
  - Windows trained on early morning hours tended to perform worse on testing data from the afternoon, likely due to changing traffic patterns and environmental conditions.
  - Conversely, models trained on afternoon hours showed better performance on testing windows closer in time, reflecting the temporal dependence of network dynamics.
- **Graphical Representation:** A graph of sliding windows versus MSE is included in Figure 4, illustrating the variation in error across different windows.
- **Interpretation:** The variability in performance was expected due to the limited dataset and the inherent differences in network behavior during different hours. In larger datasets spanning multiple days or weeks, sliding window analysis could be leveraged to train models that adapt to specific temporal patterns, improving overall performance.

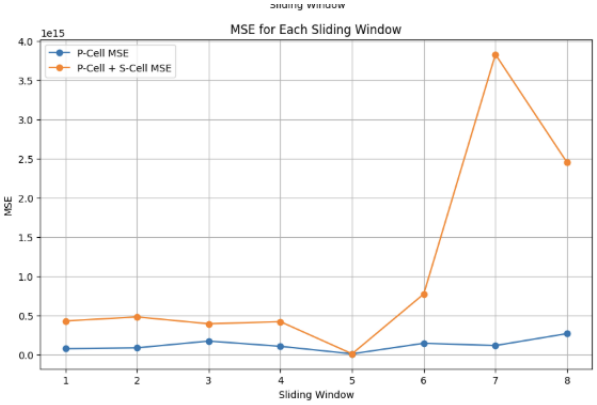

Fig. 4. Mean Squared Error (MSE) across Sliding Windows

I. Feature Importance

Understanding feature importance is crucial for interpreting the model's decision-making process and identifying the most influential variables for predicting downlink throughput. The feature importance analysis for both models, **P-Cell-Only** and **P-Cell + S-Cell**, provides valuable insights into the key drivers of network performance.

1) Top Features for Each Model: The top 5 most important features for each model are summarized below:

- **P-Cell-Only Model:**
  - **PCell_SNR_max:** Signal-to-Noise Ratio of the primary cell.

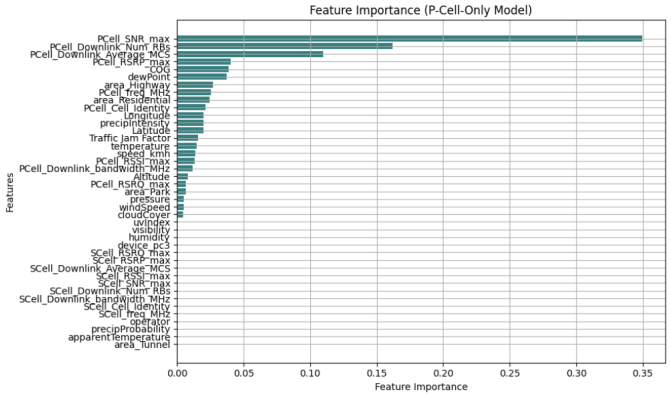

Fig. 5. Feature Importance for P-Cell-Only Model.

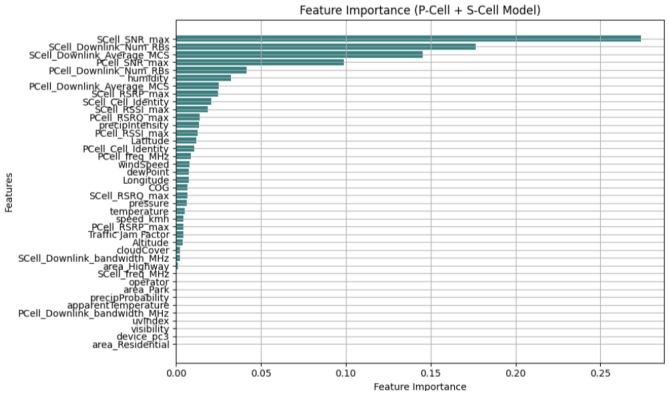

Fig. 6. Feature Importance for P-Cell + S-Cell Model.

  - **PCell_Downlink_Num_RBs:** Number of resource blocks used for the downlink in the primary cell.
  - **PCell_Downlink_Average_MCS:** Average modulation and coding scheme used for the primary cell's downlink.
  - **COG:** Course over ground, which captures vehicle direction.
  - **PCell_RSRP_max:** Maximum received signal power for the primary cell.
- **P-Cell + S-Cell Model:**
  - **SCell_SNR_max:** Signal-to-Noise Ratio of the secondary cell.
  - **SCell_Downlink_Num_RBs:** Number of resource blocks used for the downlink in the secondary cell.
  - **SCell_Downlink_Average_MCS:** Average modulation and coding scheme for the secondary cell's downlink.
  - **PCell_SNR_max:** Signal-to-Noise Ratio of the primary cell.
  - **PCell_Downlink_Num_RBs:** Number of resource blocks used for the primary cell's downlink.

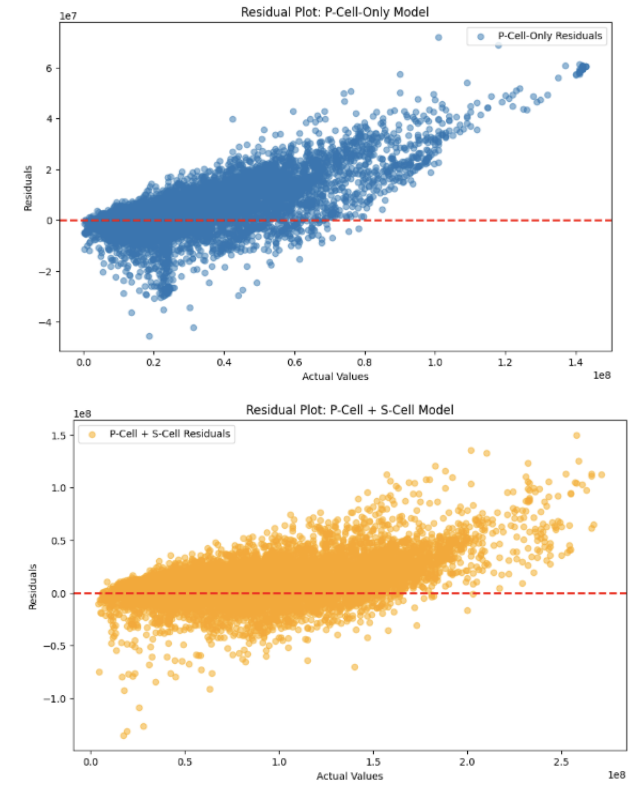

Fig. 7. Residuals Plot for (a) P-Cell-Only Model (Top) and (b) P-Cell + S-Cell Model (Bottom).

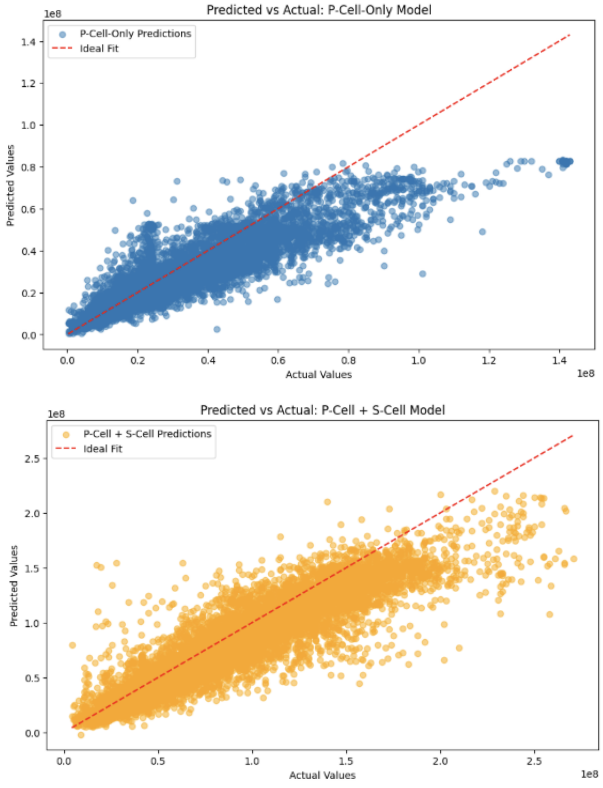

Fig. 8. Predicted vs. Actual Plot for (a) P-Cell-Only Model (Top) and (b) P-Cell + S-Cell Model (Bottom).

2) Evaluation Metrics: Model performance was evaluated using a combination of metrics and visualizations to capture various aspects of prediction accuracy:

- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values. Lower MSE indicates better accuracy.
- **Root Mean Squared Error (RMSE):** Provides a more interpretable measure of error in the same units as the target variable. It penalizes larger errors more heavily.
- **Mean Absolute Percentage Error (MAPE):** Quantifies the prediction error as a percentage, offering a normalized measure of accuracy.
- **Coefficient of Determination (R^2):** Indicates the proportion of variance in the target variable explained by the model. Higher R^2 values signify better model fit.

To complement the quantitative evaluation metrics, the following visualizations are included:

- Residuals plot for the P-Cell-Only and P-Cell + S-Cell models.
- Predicted vs. Actual values plot for both models.

VII. Conclusion

In this report, we explore the challenge of predicting Quality of Service (QoS) in vehicular networks using advanced machine learning techniques. The Berlin V2X dataset, which includes LTE metrics, GPS data, environmental factors, and network characteristics, provided a comprehensive resource for understanding network performance in dynamic vehicular environments.

Our experiments demonstrated that by leveraging key features such as **PCell SNR**, **PCell Downlink RBs**, **PCell Downlink Average MCS**, **COG**, and **area type**, we could achieve a predictive accuracy of approximately **71% for P-Cell-Only data** and **76% for P-Cell + S-Cell data** when using XGBoost. This was achieved using only **10 hours of training data** to predict network throughput for the next **2 hours**. The relatively high performance within this limited data range suggests that the model is well suited for learning temporal and spatial patterns inherent to vehicular communication environments.

This performance is significant, considering the small size of the data set and the limited temporal coverage. It indicates that models trained on larger datasets that span multiple days or diverse geographic regions could potentially exhibit even greater precision by capturing a wider variety of patterns in vehicular communication.

VIII. Future Scope

The application of AI-based QoS prediction in vehicular networks holds immense promise, particularly for use cases such as:

- **Fleet Coordination:** Reliable QoS predictions can enable smoother communication between vehicles in a fleet, ensuring consistent connectivity for shared data.

- **Proactive Route Planning:** Predicting network performance along different routes allows vehicles to adjust paths to maintain uninterrupted connectivity during critical tasks, such as package delivery or logistics operations.
- **Teleoperated Driving and Remote Assistance:** Vehicles requiring remote intervention can benefit from robust QoS predictions to ensure seamless real-time communication.

A. Opportunities for Larger Datasets

Expanding the dataset to cover longer durations and diverse regions could significantly enhance the model's robustness. A data set with weeks or months of data would allow the model to:

- **Identify Long-Term Trends:** Capture seasonal and day-to-day variations in network conditions.
- **Optimize for Temporal Patterns:** Learn repetitive patterns in traffic and network behavior, improving predictions during rush hours or special events.

B. Global Optimization through Transfer Learning

Transfer learning offers an opportunity to adapt pre-trained models to new regions, reducing the computational burden of training from scratch. Key advantages include:

- **Localized Fine-Tuning:** Adjusting models for region-specific features like traffic density, environmental factors, and road layouts.
- **Cross-Regional Collaboration:** Sharing datasets and insights across regions to create more versatile models.
- **Synthetic Data Augmentation:** Using AI-generated data to supplement training for underrepresented regions.

However, challenges such as variability in traffic patterns and environmental conditions across regions must be addressed to achieve robust, globally deployable models.

IX. Limitations

Despite the promising results, certain limitations must be acknowledged:

- **Limited Dataset Scope:** The dataset only spans two days, making it challenging to generalize the findings on larger temporal scales.
- **Regional Variability:** Different regions exhibit distinct traffic patterns, environmental conditions, and network infrastructure, making it difficult to generalize the model to new areas without additional data.
- **Real-Time Constraints:** The computational requirements of real-time QoS prediction may increase significantly with larger data sets or more complex models.

X. Final Thoughts

This study highlights the potential of machine learning for QoS prediction in vehicular networks, even with a limited dataset. The ability to predict network throughput with 71% and 76% accuracy for P-Cell-Only and P-Cell + S-Cell data, respectively, using just 10 hours of training data, is a promising outcome.

The feature importance analysis provided valuable insights into the key drivers of network performance. Features such as **SNR (Signal-to-Noise Ratio)**, **Downlink Resource Blocks**, and **Average Modulation and Coding Scheme (MCS)** for both primary and secondary cells were identified as significant contributors to accurate throughput predictions. These insights can guide network optimization strategies by focusing on metrics that have the greatest impact on QoS.

With larger datasets, global optimization techniques, and region-specific fine-tuning, these models could play a pivotal role in enabling reliable and scalable QoS prediction systems for future vehicular networks.

References

[1] Hernangomez, R., Geuer, P., Palaios, A., Schaufele, D., et al. "Berlin V2X: A Machine Learning Dataset from Multiple Vehicles and Radio Access Technologies." IEEE Vehicular Technology Conference (VTC2023-Spring), 2023.

[2] Ain, N. U., Hernangomez, R., Palaios, A., Kasparick, M., Stanczak, S. "QoS Prediction in Radio Vehicular Environments via Prior User Information." arXiv:2402.17689v1 [cs.LG], 2024.

[3] M. Althoff, L. He, and P. Koopman, "Safety Verification of Autonomous Vehicles for Realistic Traffic Scenarios," *arXiv preprint arXiv:2109.09376*, 2021. [Online]. Available: https://arxiv.org/abs/2109.09376.