Enhancing Vehicular Communication with Predictive Quality of Service: A Real-World Proof of Concept

Maroua Drissi
Orange Research
France
maroua1.drissi@orange.com

***Abstract*—Vehicular applications, particularly time-sensitive ones like tele-operated driving and traffic management, require strict Quality of Service (QoS) to ensure reliable communication within the Vehicle-to-Everything (V2X) ecosystem. Predictive QoS (p-QoS) enhances these systems by forecasting network conditions, enabling real-time adaptation to maintain continuous safety and availability. Sudden QoS fluctuations can disrupt operations, but p-QoS facilitates proactive adjustments through early notifications. AI-driven predictive models monitor critical network metrics, anticipating variations to dynamically optimize communication protocols. This paper presents a Proof of Concept (PoC) conducted at the UTAC Linas-Montlhéry test site, evaluating p-QoS impacts on tele-operated driving, particularly for critical requirements like real-time video stream transmission. Unlike prior simulations or theoretical models, this PoC demonstrates the operational feasibility of AI-driven p-QoS in a controlled real-world tele-operation setting. Results highlight p-QoS's potential to improve V2X system resilience and adaptability, ensuring reliable performance even in dynamic environments.**

***Index Terms*—Predictive Quality of Service (p-QoS), Vehicular Communication, Real-time Video Transmission, AI-driven QoS Forecasting, Tele-operated Driving**

I. Introduction

Autonomous driving has seen rapid advancement, driven by its potential to improve traffic safety, energy efficiency, and network optimization. Vehicles are typically categorized along a spectrum from driver-assisted to fully autonomous (SAE levels 0 to 5), with current implementations—primarily at levels 2 to 3—relying heavily on onboard sensors. However, these sensors face inherent limitations in perception due to restricted visual horizons [1].

To overcome these limitations, vehicular communication technologies—especially those within the Vehicle-to-Everything (V2X) framework—have gained prominence. The deployment of advanced networking technologies such as 5G has markedly enhanced the capabilities of autonomous systems by supporting high-bandwidth, low-latency communication and fulfilling stringent Quality of Service (QoS) requirements for time-critical applications [2], [3]. Despite these improvements, challenges such as high mobility and network congestion can still introduce QoS variability, directly affecting latency-sensitive applications like tele-operated and autonomous driving.

While current 5G networks can adapt reactively to QoS fluctuations, proactive adaptation mechanisms at the application level remain underdeveloped. This gap highlights the importance of Predictive Quality of Service (p-QoS), a paradigm that provides advance notification of potential network degradations. Implementing p-QoS at the network edge enables vehicular applications to anticipate and adjust to changing conditions, ensuring uninterrupted service and improved user experience—particularly in safety-critical scenarios [3].

Machine learning techniques play a central role in enabling p-QoS through real-time monitoring and forecasting of network metrics. In vehicular networks, AI-driven predictive models can dynamically assess and forecast key QoS indicators such as latency, allowing communication protocols to adjust proactively. This is particularly vital for applications like tele-operated driving, where any latency spike could compromise operational safety.

This study presents a Spatio-Temporal supervised learning model aligned with the 3GPP standard, aimed at predicting end-to-end (E2E) latency for V2X communication. Unlike earlier multi-stage approaches, this model leverages the RANdom SAmple Consensus (RANSAC) [4] algorithm, which is particularly robust against noisy and outlier-prone data. In vehicular communication scenarios, where data may be affected by environmental disturbances or sudden network fluctuations, RANSAC offers the advantage of isolating and focusing on meaningful patterns while minimizing the influence of aberrant values. This ensures reliable latency predictions even under imperfect data conditions typical of C-V2X environments.

To enhance model performance and generalization, RANSAC is used in conjunction with a Random Forest (RF) [5] estimator. RF builds an ensemble of decision trees trained on random subsets of the data, thereby reducing overfitting and improving robustness, especially in complex and dynamic vehicular settings. This hybrid approach allows the model to effectively account for both the spatial and temporal variability of QoS parameters, such as latency, bandwidth, or connection reliability.

Moreover, the prediction model supports periodic retraining based on updated measurements, enabling it to adapt over time to shifts in traffic patterns or environmental conditions. For example, retraining can be scheduled daily using latency measurements collected during the preceding day. This continual update mechanism further enhances the accuracy and relevance of predictions in real-world vehicular networks.

To validate the proposed approach, a Proof of Concept (PoC) was conducted at the UTAC Linas-Montlhéry test track, providing a real-world setting to evaluate the system's performance. This builds on previous work in predictive latency modeling [6], extending it into a live vehicular environment. The PoC results demonstrate the feasibility and benefits of real-time p-QoS in enhancing the resilience of V2X applications by enabling anticipatory communication adjustments and ensuring continuous, safe operation in dynamic network conditions.

The primary objectives of this work are to:

- Develop and validate an AI-driven predictive model for latency forecasting for tele-operated driving applications.
- Evaluate the effectiveness of p-QoS in adapting to dynamic network conditions in real-world environments.
- Assess the impact of network latency on critical systems such as video stream transmission and vehicle control cameras.

The remainder of this paper is structured as follows: Section 2 provides an overview of the system architecture, detailing the components and workflow. Section 3 covers the implementation aspects, including data ingestion, storage, workflow automation, and model training for prediction. Section 4 presents the evaluation and results, followed by Section 5, which offers conclusions and directions for future work.

II. QoS Prediction for 5G-V2X Communications

A. QoS in V2X Scenarios

V2X applications, particularly those involving safety-critical services like tele-operated driving (ToD), require stringent Quality of Service (QoS) guarantees. For example, ToD demands ultra-low latency (20 ms) and high reliability for real-time video, sensor data, and control command exchanges. Any degradation in QoS can compromise safety, potentially requiring route changes, speed adjustments, or controlled stops.

Table I highlights various V2X use cases and their associated QoS KPIs, along with potential application responses to predicted QoS variations [3].

In addition to reaction strategies, Table II summarizes the performance targets defined for key V2X scenarios, illustrating their dependency on precise QoS metrics [7].

B. 5G Architecture for Predictive QoS

To support proactive QoS assurance, the 3GPP introduces a QoS prediction framework using the NWDAF (Network Data Analytics Function), as illustrated in Figure 1 [8]. NWDAF delivers predictive insights to various network functions (NFs) via NEF (Network Exposure Function), enabling applications to adapt before QoS degradation occurs.

TABLE I
QoS-Aware V2X Use Cases and Reactions [3]

Use Case	Predicted KPIs	Possible Application Responses
Tele-operated driving	Latency, reliability, data rate	Change route or mode, park, transfer control, adjust sensors
Platooning	Latency, reliability	Adjust inter-vehicle distance or speed, terminate platoon
Hazard warning	Reliability	Notify driver, reduce speed, reroute
Lane merge	Latency, reliability	Delay or abort merge maneuver
Software update	Data rate	Reschedule or pause download
Infotainment	Data rate	Reduce media quality

TABLE II
V2X QoS Requirements [7]

Use Case	Latency (ms)	Reliability (%)	Data Rate (Mbps)
Vehicle Platooning	10	99.99	50–65
Advanced Driving	3–10	99.999	30–53
Extended Sensors	3–50	99–99.999	10–1000
Remote Driving	5	99.999	UL: 25 / DL: 1

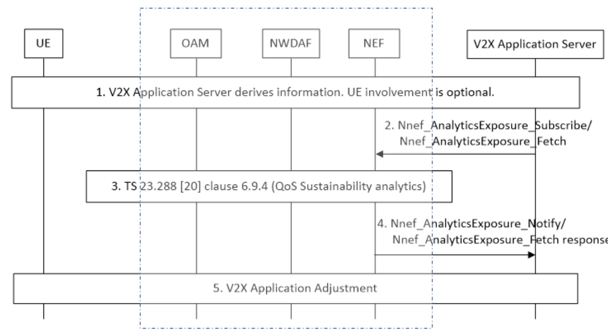

Fig. 1. QoS Change Notification to V2X Application Server [8]

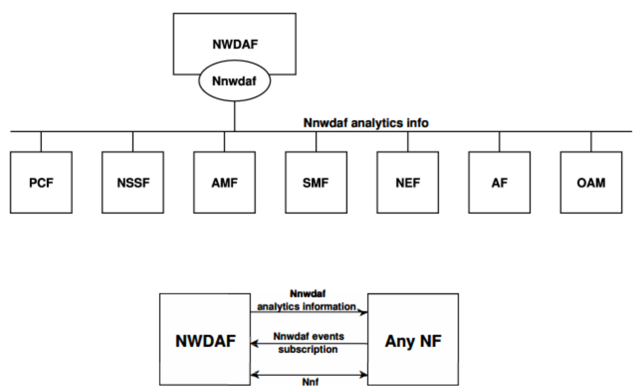

Fig. 2. NWDAF-based Analytics Framework [10]

As shown in Figure 2, NWDAF acts as an analytics hub within the 5G core, providing real-time QoS evaluations and forecasts to NFs such as PCF, AMF, SMF, and OAM. Its ML-driven capabilities enable functions like policy enforcement, mobility management, and network slicing to operate with predictive awareness. NWDAF collects data from OAM and other NFs, evaluates QoS trends, and issues early alerts to V2X applications through NEF. These alerts can prompt operational changes such as mode switching or safe service termination, ensuring continuity and safety in dynamic vehicular environments.

III. Robust Latency Prediction via RANSAC–RF Model

To handle the noisy and variable conditions in C-V2X networks, we use a robust spatio-temporal supervised learning framework combining RANSAC with a Random Forest (RF) estimator. This hybrid model predicts end-to-end latency with high resilience under fluctuating QoS.

A. Model Robustness

RANSAC (RANdom SAmple Consensus) resists outliers common in V2X data by fitting regression models to random subsets and focusing on consistent patterns. RF complements RANSAC by aggregating predictions from multiple decision trees trained on bootstrapped samples, capturing complex nonlinearities and reducing overfitting. Together, they address the spatial-temporal variability and reliability needs of C-V2X latency prediction.

B. Input Processing and Feature Augmentation

Input data is augmented to improve temporal context:

- **Historical QoS Injection:** Each sample includes the last 100 latency measurements per quadkey as additional features, preserving raw temporal patterns.
- **Temporal Features:** Timestamps are split into year, month, and day columns to provide periodic context.

C. Training and Model Updates

The model trains on historical data with augmented features and is retrained bi-weekly using recent measurements to adapt to changing network conditions. The computational overhead for inference is negligible in practice, while retraining is performed offline on a dedicated server, thus not impacting real-time operation.

D. Inference Pipeline

Deployed on a dedicated server, the model predicts latency for queries defined by spatial location and forecast horizon, using the latest QoS data. This enables proactive V2X adaptation based on predicted network performance.

This approach integrates robust algorithms, temporal feature engineering, and continuous learning to deliver accurate, real-time QoS predictions vital for safe C-V2X operation.

IV. System Architecture

Vehicular applications, especially within the Vehicle-to-Everything (V2X) ecosystem, require stringent Quality of Service (QoS) standards to guarantee reliable communication. Applications such as tele-operated driving and traffic management systems are critically dependent on latency and other real-time performance metrics. To address these needs, Predictive Quality of Service (p-QoS) is introduced as a mechanism that forecasts network conditions, enabling systems to dynamically adjust and ensure continuous safety and availability.

This section presents an overview of the system architecture, including how p-QoS integrates with vehicular communication frameworks and how it enables real-time adaptation for various V2X applications.

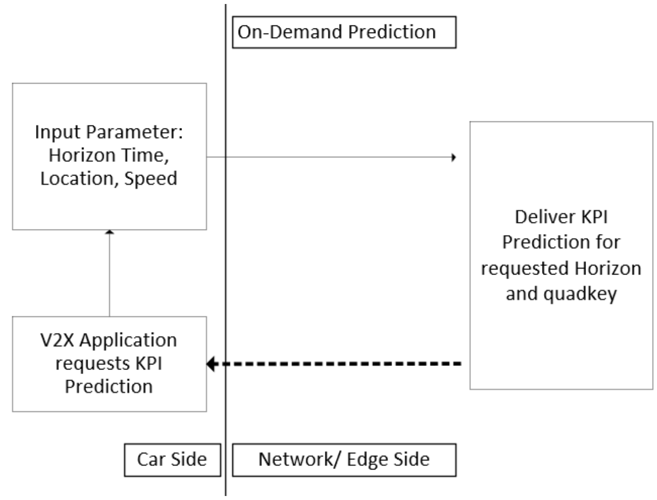

Fig. 3. System Architecture Overview: High-Level p-QoS Integration in V2X Systems

A. System Components and Operational Workflow

The proposed p-QoS architecture comprises a modular ensemble of components that collectively enable end-to-end quality of service monitoring, predictive analytics, and real-time adaptation in vehicular communication systems. The key components include: (i) a data ingestion and storage pipeline, (ii) a stream processing and orchestration layer, (iii) machine learning-based predictive models, and (iv) adaptive control mechanisms integrated with application-layer services.

The operational workflow initiates with continuous monitoring of critical network and application metrics such as latency, throughput, jitter, and signal quality—captured from in-vehicle and roadside units. This raw telemetry is ingested in real time, preprocessed, and stored for both immediate inference and offline model training. Predictive algorithms process the ingested data to forecast short-term QoS conditions. Based on these forecasts, proactive system-level adaptations are triggered to mitigate potential degradation in performance, such as reconfiguration of streaming parameters or prioritization of mission-critical data.

The architectural design aims to maintain service continuity and performance integrity under varying network conditions, including high congestion, mobility-induced fluctuations, and load imbalances. This architecture was deployed in the UTAC testbed to assess end-to-end latency forecasting and its impact on camera switching for tele-operation.

The subsections below provide a detailed examination of each architectural module, highlighting their respective roles in supporting predictive QoS capabilities.

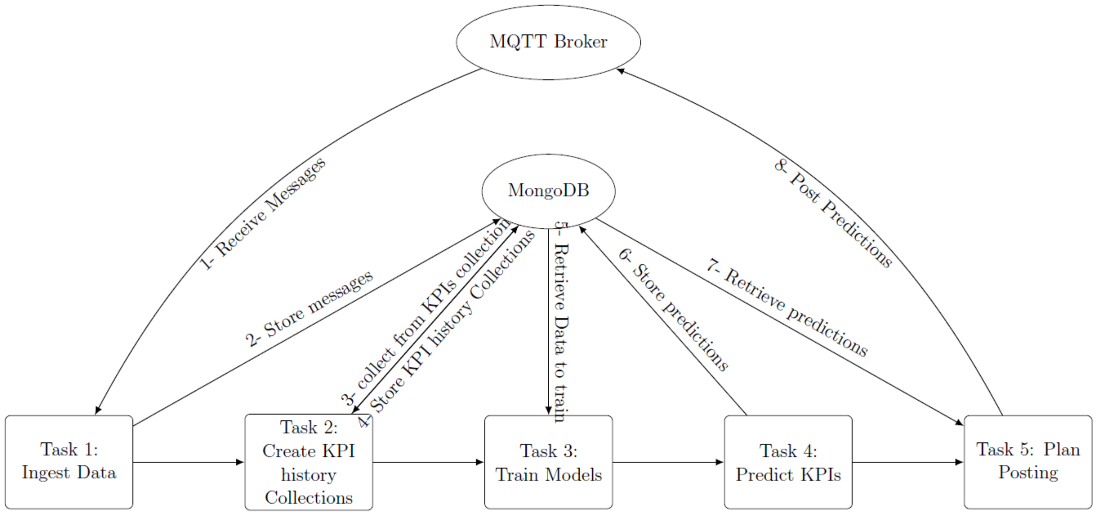

Fig. 4. High-Level System Architecture: Modular Workflow for Predictive QoS in Real-Time Vehicular Communication

1) Real-Time Data Ingestion and Storage: To enable predictive QoS, the system implements a high-frequency telemetry pipeline that captures fine-grained network statistics directly from the vehicular edge. Parameters such as round-trip time (RTT), packet loss rate, and signal-to-noise ratio (SNR) are logged at sub-second intervals. The collected data is structured and stored in a centralized time-series database to support real-time access and retrospective model training. The storage system also facilitates efficient querying, ensuring scalability across multiple vehicle nodes and road segments.

2) Predictive Modeling and QoS Forecasting: The machine learning component forms the analytical core of the p-QoS system. Supervised learning algorithms are employed to model the temporal dynamics of network latency and bandwidth under diverse vehicular and environmental conditions. Historical data is used to train models such as Random Forests, Gradient Boosted Trees, or LSTM-based neural networks. These models are continuously retrained using recent data to capture evolving network behaviors. Predictions are generated in real time and are used to inform QoS-sensitive decision logic, such as adaptive bitrate control for video streams or latency-aware camera switching in autonomous vehicles.

3) Feedback Loop and Control Adaptation: Forecast outputs feed into a real-time control layer responsible for adapting communication protocols or application behavior. For instance, when a latency spike is predicted, non-critical data transmission can be deferred, or video resolution may be downgraded to preserve responsiveness. This closed-loop system maintains service quality by dynamically reconfiguring system parameters based on predicted network states, ensuring optimal operation even in degraded or volatile environments.

V. Evaluation & Results

This section presents the results of the Proof of Concept (PoC) at the UTAC Linas-Montlhéry test site, evaluating p-QoS impact on tele-operated driving, particularly in relation to critical requirements such as camera flux transmission.

Results demonstrate the predictive model's ability to forecast latency variations and enable adaptive communication. Performance is assessed by comparing predictions with measured values using Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Mean Squared Error (MSE), highlighting latency's effect on system performance and p-QoS effectiveness under dynamic conditions.

A. Tests Overview

To validate the robustness and accuracy of the latency prediction model under various network and mobility conditions, a series of tests were designed and executed. Table III summarizes the test scenarios.

B. Test 1: Static Conditions

Objective: Evaluate prediction accuracy in a controlled, static environment.

Results:

- Actual latency: 17.5 ms (avg)
- Predicted latency: 13.1 ms (avg)
- MAE: 4.03 ms
- MSE: 20.17 ms^2
- MAPE: 24.83%

Fig. 5. Latency predictions vs actual values during static test

Discussion: Predictions are reasonably accurate, with slightly higher error margins likely due to network variability.

TABLE III
OVERVIEW OF THE CONDUCTED TESTS

Test No.	Objective	Description	Expected Outcome
0	Hardware Verification	Check equipment readiness, run iperf, configure Dashboard.	Functional setup
1	Static Evaluation	Compare actual and predicted latencies at rest.	MAPE $< 25\%$
2	Real Conditions Performance	Evaluate prediction during vehicle motion.	Stable predictions
3	Load Test (DL)	Apply download traffic with iperf.	Stable prediction
3bis	Load Test (DL + UL)	Apply simultaneous DL+UL traffic.	Stable prediction
4	Latency Impact on Cameras	Observe camera behavior under latency thresholds.	Camera control by latency

Fig. 6. MAE and MAPE values during static test

C. Test 2: Dynamic Environment

Objective: Measure prediction under vehicular movement.
Results:

- Actual latency: 22.0 ms
- Predicted latency: 22.03 ms
- MAE: 5.59 ms
- MSE: 57.07 ms^2
- MAPE: 20.42%

Fig. 7. Prediction performance under real driving conditions

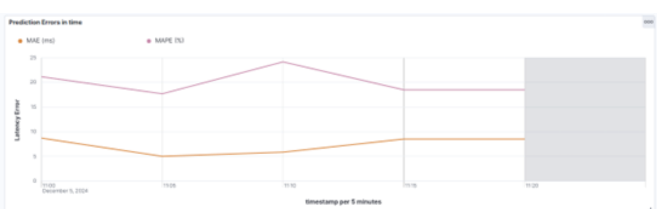

Fig. 8. MAE and MAPE values under real driving conditions

Discussion: Despite added variables such as movement and variable signal quality, the model maintained stable predictions.

D. Test 3 and 3bis: Network Load Conditions

Objective: Assess robustness under download (Test 3) and combined DL+UL (Test 3bis) loads.

Results Summary:

- MAE: 4.91 ms
- MSE: 43.92 ms^2
- MAPE: 17.71%

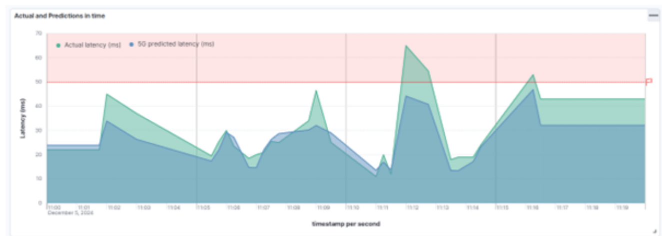

Fig. 9. Latency prediction under network load (DL and DL+UL)

Fig. 10. MAE and MAPE values under network load (DL and DL+UL)

Discussion: The model demonstrated consistency even under significant load, affirming its resilience.

E. Test 4: Impact on Camera Systems

Objective: Verify latency thresholds triggering camera deactivation.

Results:

- Actual latency: 20 ms
- Predicted latency: 17.93 ms
- MAE: 4.82 ms
- MSE: 38.07 ms^2
- MAPE: 19.50%

Camera Observations:

- Cameras deactivate when latency > 30 ms (e.g., at 16:15:17)
- Reactivation when latency falls below 25 ms (e.g., at 16:17:30)
- Central camera remains unaffected; lateral ones respond to thresholds.

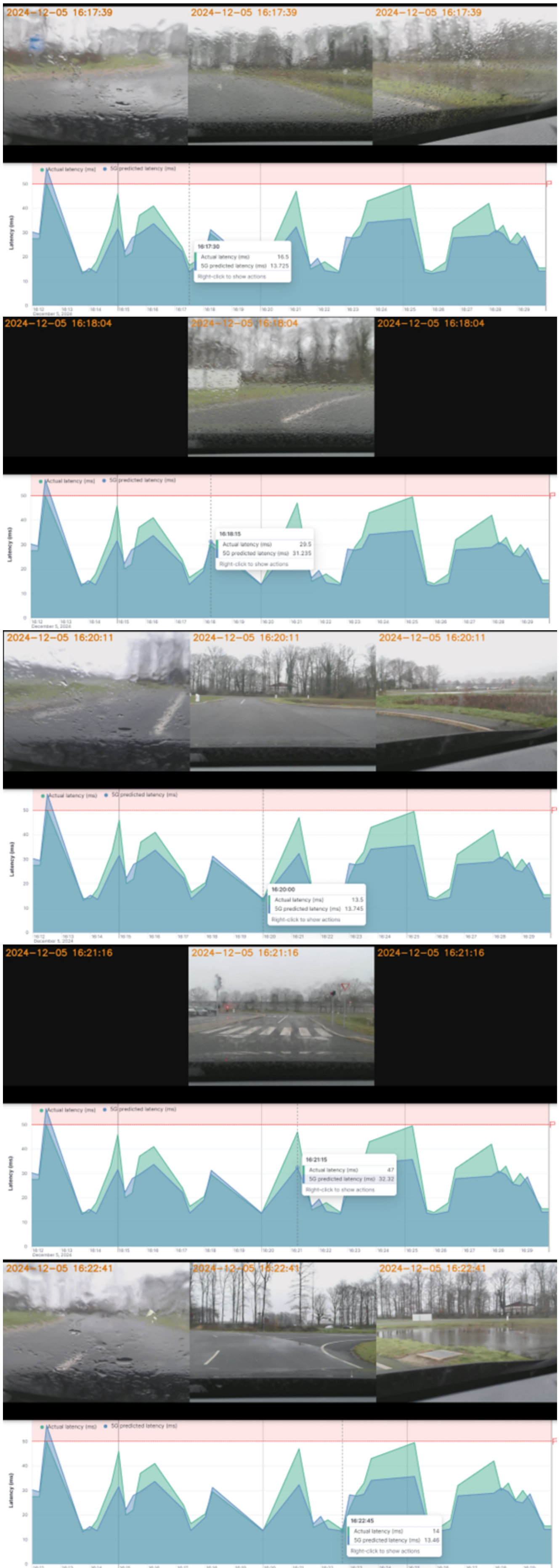

Fig. 11. Camera behavior in response to latency spikes

Discussion: The latency-aware deactivation logic performed effectively, with the prediction model showing satisfactory accuracy (MAPE $< 25\%$) across static, mobile, and loaded conditions. Predictions were robust for real-time, safety-critical applications such as camera activation, and dashboard visualizations supported interpretability. Future work will focus on further reducing errors in dynamic and congested scenarios.

VI. Conclusion

This paper demonstrated the potential of p-QoS to enhance resilience and adaptability in V2X systems, with a focus on safety-critical tele-operated driving. A real-world Proof of Concept at the UTAC Linas-Montlhéry test track showed that a hybrid RANSAC–RF latency prediction model can anticipate network fluctuations and enable proactive adaptation. The architecture, combining data ingestion, predictive analytics, and real-time feedback, maintained service reliability under dynamic conditions, bridging the gap between theoretical models and operational deployment.

VII. Limitations and Future Work

While promising, the study has limitations. Experiments were restricted to a single site and operator, limiting generalization to diverse environments (urban, highway, adverse weather). Prediction errors (MAPE $< 25\%$) remain non-negligible in highly dynamic scenarios, though sufficient to trigger proactive mechanisms such as latency-aware camera deactivation. Inference overhead was negligible, but offline retraining requires further optimization. Finally, the system complements the 3GPP NWDAF framework; future studies will investigate long-term stability, prediction horizon trade-offs, and comparisons with alternative models such as LSTM or gradient-boosted trees.

References

[1] Naranjo, José, et al. "Cross-Border interoperability for Cooperative, Connected and Autonomous Driving." IEEE Intelligent Transportation Systems Magazine (2021).

[2] Wang, Jiadai, Jiajia Liu, and Nei Kato. "Networking and communications in autonomous driving: A survey." IEEE Communications Surveys & Tutorials 21.2 (2018): 1243-1274.

[3] Making 5G Proactive and Predictive for the Automotive Industry, 5GAA, White paper, December 2019

[4] O. Chum and J. Matas, "Optimal Randomized RANSAC," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, pp. 1472-1482, Aug. 2008, doi: 10.1109/TPAMI.2007.70787

[5] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[6] M. Drissi and S. Allio, "End-to-End Spatio-temporal Latency Prediction for Vehicular Applications," 2024 International Wireless Communications and Mobile Computing (IWCMC), Ayia Napa, Cyprus, 2024, pp. 126-131, doi: 10.1109/IWCMC61514.2024.10592518.

[7] 3GPP, "Service requirements for enhanced V2X scenarios (release 15)," 3rd Generation Partnership Project, Sophia Antipolis, France, Tech. Rep. 3GPP TR 22.186 V15.0.0, Mar. 2017.

[8] 3GPP TS 23.288, "Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (v16.2.0, Release 16)," Dec. 2020.

[9] Pateromichelakis, Emmanouil, et al. "End-to-end data analytics framework for 5G architecture." IEEE Access 7 (2019): 40295-40312.

[10] 3GPP, "5G System; Network data analytics services; Stage 3," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.520), April 2021, (Release 16).