IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Estimating Quality-of-Service in Urban Vehicular Networks through Machine Learning

DUARTE DIAS[1], MIGUEL LUÍS[1,3], PEDRO RITO[1], (Member, IEEE), and SUSANA SARGENTO[1,2], (Member, IEEE)

[1]Instituto de Telecomunicações, 3810-193 Aveiro, Portugal (e-mail: duarterochadias@ua.pt, nmal@av.it.pt, pedrorito@av.it.pt, susana@ua.pt)
[2]Departamento de Eletrónica, Telecomunicações e Informática, Universiade de Aveiro, 3810-193 Aveiro, Portugal
[3]Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal

Corresponding author: Duarte Dias (e-mail: duarterochadias@ua.pt).

This work was supported by the European Union/Next Generation EU, through Programa de Recuperação e Resiliência (PRR) [Project Nr. 29: Route 25], by the European Regional Development Fund (FEDER), through the Competitiveness and Internationalization Operational Programme (COMPETE 2020) of the Portugal 2020 framework [Project IMMINENCE with Nr. 112314 (POCI-01-0247-FEDER-112314)] and by FCT/MEC through national funds under the project MH-SDVanet (PTDC/EEI-COM/5284/2020).

ABSTRACT Machine Learning (ML) has emerged as a promising tool for addressing complex challenges in multiple domains. In the context of Vehicular Ad-Hoc Networks (VANETs), ML has gained much more attention due to its ability to solve major known problems in areas such as traffic management, road safety and communication infrastructure management. In a VANET, vehicles generate a significant amount of data, which can be explored to, for example, enhance the network management regarding the connectivity between the vehicles and the infrastructure.
This work studies the performance of ML models regarding the estimation of the Quality-of-Service of different network access technologies (ITS-G5 and 5G) in urban vehicular environments. To this end, data collection campaigns were carried out throughout the city of Aveiro, Portugal, which included vehicular and network performance data for ITS-G5 and 5G cellular technologies. After an initial characterization of the data collected, several ML algorithms were trained, considering different combinations of features (represented by the collected metrics). The results have shown that, for the same configurations, similar estimation errors were obtained by the Random Forest Regression and the Extreme Gradient Boosting algorithms, with the last one presenting a shorter estimation time. The results also show that location-independent configurations, i.e., when no geographic positions are used in the ML model, are slightly worse than GPS-based ML models, creating the possibility of being applied in different urban environments, making them quite versatile.

INDEX TERMS Intelligent Transport Systems, Machine Learning, Vehicular Networks, ITS-G5 and 5G, Vehicle Data Collection, Network Performance.

I. INTRODUCTION

Vehicular Ad-Hoc Networks (VANETs), a key part of the Intelligent Transport System (ITS) framework [1], have emerged as a promising technology to enable effective communication among vehicles and the infrastructure, leading to improved road safety, traffic management, air pollution and enhanced driving experiences. As VANETs generate a massive amount of data, the integration of machine learning techniques can play a crucial role in extracting valuable insights.

In the realm of VANETs, Quality of Service (QoS) stands as a central concern. Given the dynamic nature of this domain, fluctuations in radio signal propagation pose challenges to maintain fast and stable connections, especially for services reliant on such connectivity.

Ensuring optimal network performance requires precise evaluations of communication link quality. These evaluations enable higher-priority network services to enjoy precedence over lower-priority ones, thereby enhancing the overall network performance.

Machine Learning (ML) algorithms have emerged as vital tools to predict the signal quality. By analyzing real-world data encompassing factors like distance to Point of Attachments (PoAs), signal strength, and environmental conditions, ML models can forecast the network performance, facilitating intelligent decision-making processes.

The proposed approach uses ML techniques to create models capable of predicting network performance metrics, such as bitrate, jitter, packet loss and round-trip time (RTT), for various communication technologies. Taking advantage of the dynamism of a vehicular network, it is possible to obtain a large set of data essential for training the forenamed models. The solution considers multiple ML algorithms alongside different configurations of features to get the best possible predictions.

In order to train and evaluate the models, extensive data collections for ITS-G5 and cellular technologies were necessary. The collected data is genuine city data, collected in an urban vehicular environment with multiple vehicles.

The main contributions of this article are the following:

- Use of ML techniques to create models to predict network metrics for different communication technologies;
- Network data obtained in a city environment for both ITS-G5 and 5G cellular technologies;
- Vehicle and network performance data obtained using real vehicle communication equipment (On-Board Units (OBUs) and Road-Side Units (RSUs)) in a city environment;
- Evaluation of multiple ML algorithms and analysis of multiple configurations of model features.

The remaining of the article is organized as follows. Section II discusses the topic of ML in mobile networks with special empahsis on VANETs. Section III details the data and its collection, while the proposed approach is presented in section IV. Section V discusses the obtained results. Finally, Section VI enumerates the conclusions and introduces directions for future work.

II. MACHINE LEARNING IN VANETS

The ML concept has emerged as a life-changing technology in multiple areas, revolutionizing the way data is analysed. One area that has undergone a major transformation with the integration of ML techniques is the optimization of communication in VANETs. These techniques have the potential to address several complex challenges that define the vehicular environment. By using the power of ML algorithms, it is possible to extract valuable insights from the vast amount of generated data by vehicular networks, enabling intelligent decision-making processes. From the data obtained, it is possible to make the algorithms learn patterns and behaviours from historical data, predict future events, and optimize network operations in real-time.

The sub-area of vehicular communication networks that makes the most sense to address, linked to the scope of this article, is the area of VANET QoS metrics prediction. QoS metrics are crucial in VANETs to ensure reliable and efficient communication between vehicles and the infrastructure. By using ML techniques, ML models can effectively capture complex relationships within VANETs and predict QoS metrics under varying conditions, obtaining much more accurate and reliable predictions. These techniques [2] have been recently used for predicting link quality in wireless environments.

The work in [3] proposes an approach to online the QoS estimation based on the Adaptive Random Forest algorithm, where a trained model can be taken as a base estimator and fine-tuned with information from the user equipment and the cell itself. The work concludes that online learning alleviates the complexity of the data collection procedures (required in offline learning), while also reacting timely to performance drops, significantly reducing the overhead of offline re-validation and retraining. However, online models are especially susceptible to parameter tuning, which can cause an overreaction to drifts with non-optimal parameters.

The work in [4] proposes a supervised-ML-based prediction model that estimates the Packet Reception Rate on the road, updating communication zones to adapt to changes in traffic conditions. Performance tests showed excellent results in the vast majority of cases. However, this work only considers one QoS metric (i.e., Packet Reception Rate), with results based on a simulator, which cannot necessarily translate into identical results using genuine city data.

The work in [5] proposes a deep learning model based on an Artificial Neural Network (ANN) technique to predict wireless link quality in VANETs. The proposed algorithm uses location-based features, i.e. latitude, longitude, heading and speed of a vehicle, to estimate the wireless link (i.e. Received Signal Signal) quality. Performance results showed that the developed model significantly outperforms conventional models. Although promising, the results only consider a single metric, RSS, which may not reflect the actual network performance.

In terms of QoS predictors outside vehicular networks, certain studies draw attention. The work in [6] proposes a combination of teletraffic and Neural Network (NN)-based approaches, called IESA-NN, to estimate the blocking probability metric in a mobile cellular network. The work uses an NN to estimate a tuning parameter, which is in turn used to estimate the blocking probability via a modified IESA approach. The results demonstrate that IESA-NN significantly outperforms both direct-NN and pure teletraffic-based approaches.

The study in [7] introduces a ML approach for detecting the interference and estimating the throughput

in WiFi networks across both the 2.4 GHz and 5 GHz bands. These ML models are trained with real-world data using tree-based and deep learning-based algorithms. The evaluation of the performance indicates that all ML models proposed for throughput estimation outperform the benchmark analytical metric representing the throughput.

The study outlined in [8] proposes a ML based QoS estimator for Aerial Wireless Networks. The proposed estimator is based on convolutional neural networks, and estimates the mean aggregate QoS for a given network by considering the Unmanned Aerial Vehicles (UAVs) positions, the user positions and their offered traffic. The performance results demonstrated that the QoS estimator provides accurate estimations, with an average error lower than 5%. Despite its considerable potential, the model's reliance on GPS coordinates restricts its versatility, limiting its practicality beyond training settings.

Many of the approaches presented in the literature to estimate the QoS metrics in vehicular communication exhibit some degree of limitation. These studies ultimately focus solely on RSS as the main signal quality metric or prioritize GPS coordinates, thereby constraining their effectiveness. Another significant drawback is that the majority of performance evaluations rely on simulations, potentially failing to accurately reflect real-world outcomes in urban environments.

The work developed here presents estimations of multiple performance metrics along with different technologies, which can be beneficial for different types of services on the network. Another advantage is that performance results are based on real data obtained in a genuine city environment with multiple vehicles and communication technologies.

III. DATA COLLECTION AND PROCESSING

We propose ML models based on data extrapolated from vehicular environments. The ML models provide an estimate of a set of network-related Key Performance Indicators (KPIs) that indicate the quality of each of the RSUs signal, as well as the cellular signal, as can be seen in Figure 2. These models have been trained in advance using network KPIs collected from a vast dataset in the city of Aveiro, Portugal. Figure 1 depicts the data collection process, which is based on a set of vehicles that travel around the city performing network performance tests. These tests are carried out for both ITS-G5 and 4G/5G cellular technologies. The obtained data are thus the core of the proposed approach. The overall architecture of the solution is thus divided into two categories: data collection and ML model training.

To support the proposed approach, it was necessary to obtain a set of data, specifically network performance data. This data is divided into two types, the ITS-G5 data and the data coming from the cellular network (5G). The acquisition was made from multiple network

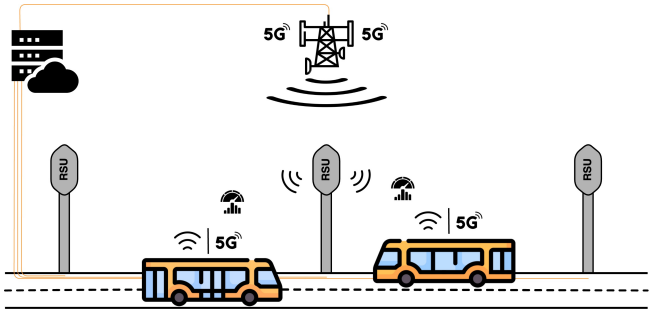


FIGURE 1. Network performance data collection.

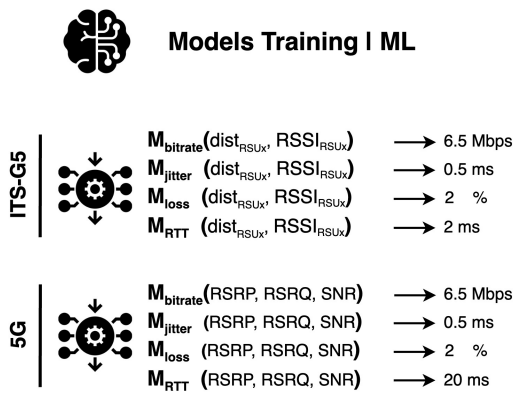


FIGURE 2. ML models training and estimation.

data collection campaigns around the city of Aveiro. For this purpose, both the Aveiro Tech City Living Lab (ATCLL) infrastructure [9] as well as a commercial cellular network were used. More information about the dataset used in this study, as well as the dataset itself is available in [10].

These data collection campaigns consist of three main components: a server, a series of ITS-G5/Cellular points, and a fleet of vehicles. Figure 3 illustrates the procedure of data collection. This process yields a collection of log files which, when interconnected, form a comprehensive dataset detailing the status of the VANET.

A. AVEIRO TECH CITY LIVING LAB (ATCLL)

The Aveiro Tech City Living Lab (ATCLL) [9], deployed in Aveiro, Portugal, is an initiative created by the city and Instituto de Telecomunicações, that combines a multi-technology advanced, large-scale communication, sensing and computation infrastructure for data management and innovative analytics.

This infrastructure, illustrated in Figure 4, comprises 44 nodes, called Road-Side Units (RSUs), strategically spread across the city, and are connected through fiber link technology (16 km long) to a data processing center. These fixed nodes contain a set of sensory devices: video cameras, radars, lidars, etc.; as well as various

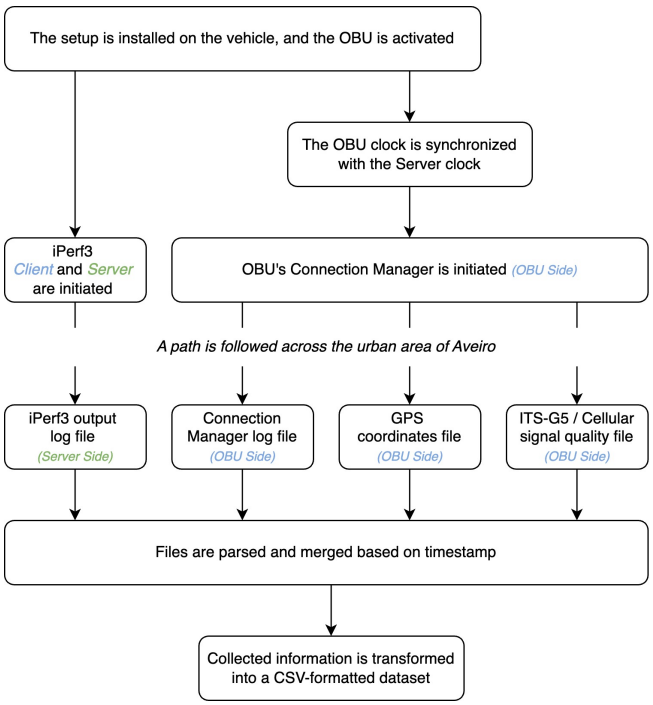


FIGURE 3. **Data Collection campaign flowchart.**

communication interfaces: 5G, ITS-G5, WiFi, mmWave and LoRa.

FIGURE 4. **ATCLL Map [9].**

The platform includes a set of more than 10 mobile nodes, called On-Board Units (OBUs), in public buses and private vehicles, that interact with the fixed infrastructure through the different communication technologies. These nodes are thus the data source used to obtain the network-related KPIs through multiple campaigns to collect this type of data around the city of Aveiro.

B. KEY PERFORMANCE INDICATORS COLLECTION CAMPAIGNS

The acquisition of data is based on multiple data collection campaigns for both ITS-G5 and 5G, that monitored both radio and network KPIs. For this purpose, the Iperf3 tool [1] is used to obtain the network performance metrics. Other types of information are also collected, such as temporal and spatial (GPS) information.

1) Information collected

Regarding the information collected in the campaigns, it is divided into multiple categories. The information acquisition frequency is variable; however, it is possible to induce lower frequency values, so all information is available at a frequency of 10Hz.

The first category is radio signal. There are four metrics obtained in this category: the Received Signal Strength Indication (RSSI) (for ITS-G5); the Reference Signal Received Power (RSRP), the Reference Signal Received Quality (RSRQ) and the Signal-to-Noise Ratio (SNR) (for 5G Cellular). The RSSI and the RSRP are measured in dBm, as the RSRQ and SNR are measured in dB. The RSSI measures the signal strength between an OBU and a given RSU. This metric is only obtained for the ITS-G5 data, as it is not possible to obtain it from the cellular side (for 5G). Regarding the RSRP, which is similar to RSSI, it measures the power of the reference signal received at the antenna of the device. The RSRQ measures the quality of the reference signal received by the device, which indicates the level of interference on the radio link. The SNR measures the ratio of the desired signal power to the sum of the power of all other interfering signals and noises. The RSRP, the RSRQ and the SNR are metrics not obtainable for the ITS-G5, therefore, exclusive to 5G cellular. The frequency of all metrics is 1Hz, but since their variability is quite low, it is possible to interpolate the remaining information to a frequency of 10Hz without compromising the veracity of the data.

The second category is the network performance. The KPIs obtained in this category are the bitrate (in Mbps), the jitter (in ms), the packet loss (in %) and the round-trip time (RTT) (in ms). The first three KPIs are obtained from the Iperf3 tool that represents a file transfer (UDP) from the the OBU to the cloud (uplink). The RTT is obtained from an ICMP generated traffic. This data has a base frequency of 10Hz and no data interpolation is required. These KPIs are of great importance for the proposed approach, since they will serve as labels for training machine learning models in the

[1] https://iperf.fr

future. These labels are obtained for both technologies (ITS-G5 and 5G); however, the traffic characteristics are different between the two, which will be explained in greater detail later on.

The third and last category is location. It includes a set of four types of information: latitude and longitude, measured in degrees; heading (direction), also measured in degrees and speed, measured in m/s. These values are obtained through a GPS module/antenna present in the OBU which captures these values. The frequency of the data is 1 Hz (standard); however, it is possible to interpolate intermediate data in order to obtain a frequency of 10 Hz without losing its veracity. There is a fifth metric in this category, the distance (OBU-RSU), measured in meters, which was derived from the GPS coordinates of the OBU and the RSU to which it was connected. This metric is also recorded at a frequency of 10Hz, achieved through data interpolation.

These three categories of data presented here represent all the data that is possible to acquire, encompassing physical, network, and movement levels. The data is acquired concurrently but through multiple methods. Nonetheless, the connection between information is established using the timestamp of each operation. The cumulative dataset comprises over 48 hours of interaction between the mobile nodes (OBUs) and each ITS-G5 RSU, as well as the 5G network.

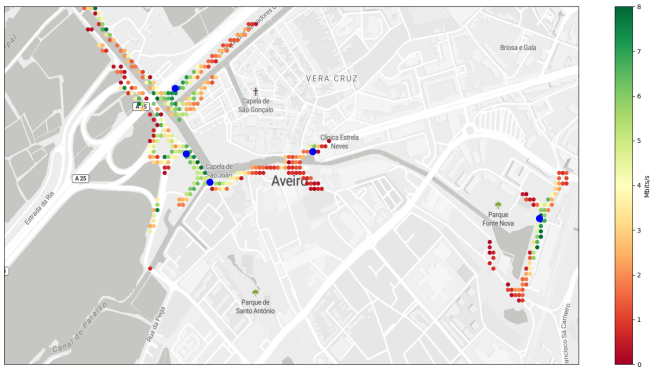
2) ITS-G5 Data

Regarding the ITS-G5 data obtained in the data collection campaigns, these are obtained from network performance tests carried out between the ITS-G5 interfaces of the mobile nodes (OBUs) and the fixed infrastructure (RSUs) of the ATCLL platform.

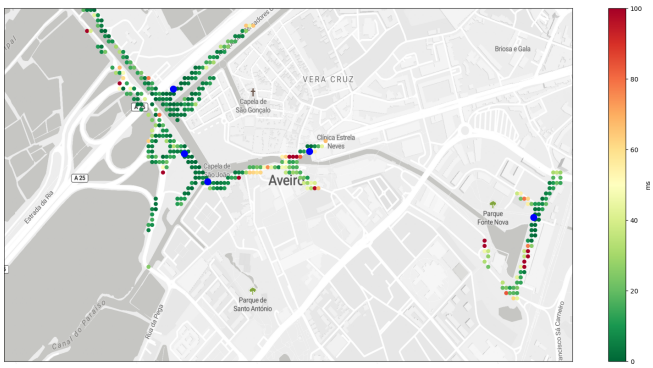
As previously mentioned, the data is obtained from the Iperf3 tool that represents a file transfer (UDP) in uplink. However, in the case of the ITS-G5, some characteristics of the traffic generated were adapted. Regarding the bitrate generated in the tests, it is limited to 10 Mbps, due to the ITS-G5 bandwidth limitations.

The results of the data collection campaigns can be seen in a map format in Figure 5, which includes the four metrics, bitrate, jitter, packet loss and RTT:

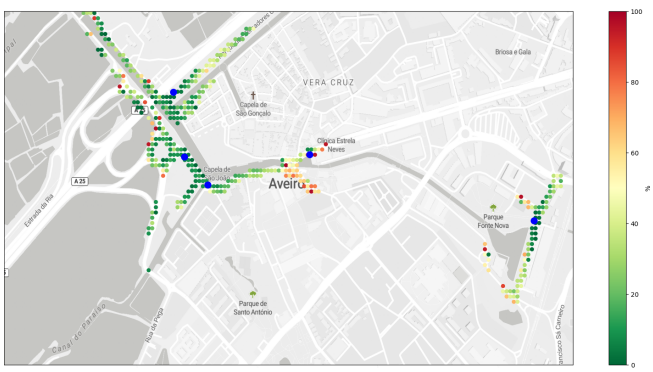
- In subfigure 5(a), relative to the bitrate, it is possible to state that, as we move away from the RSUs (blue dots), the bitrate reduces substantially. Bitrates around 5 to 7 Mbps are observed in close proximity to the RSUs, within a distance of less than 50 meters without obstructions. However, with increasing distance, bitrates quickly drop below 1 Mbps, persisting at this rate for the next hundred meters. Despite offering some coverage, ITS-G5 technology is constrained by low bitrates and vulnerable to road obstructions. Consequently, the dominent color representing ITS-G5 bitrate coverage is red.



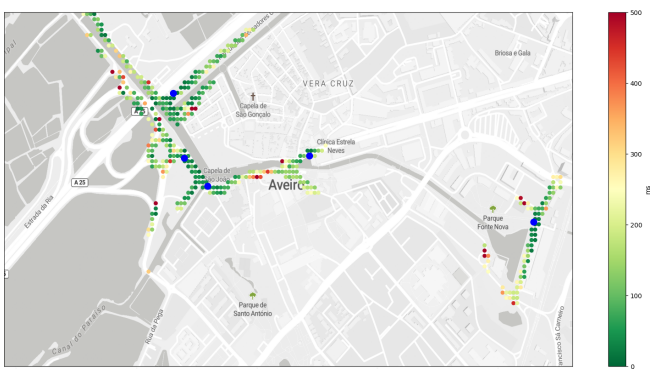
(a) Bitrate (Mbps).



(b) Jitter (ms).



(c) Packet Loss (%).



(d) RTT (ms).

FIGURE 5. ITS-G5 KPIs. Blue dots represent ATCLL RSUs.

- In subfigure 5(c), relative to packet loss, the dependence between distance and packet loss is also observable, as the latter increases as we move away from an RSU. However, in this case, the sensitivity diminishes, capable of sustaining distances of up to 100 meters while keeping packet loss below 30 %. This range results in a predominantly light green coverage. With greater distance, packet loss escalates to 50 - 70 %, consequently yielding orange zones.
- In subfigures 5(b) and 5(d), the map is predominantly dark green. It is notable that the sensitivity of these two metrics concerning distance is significantly lower compared to the others. When it comes to jitter, the majority of values fall below 50 ms, and for RTT, they remain below 150 ms. This allows the map to undergo minor fluctuations in values, leading to only a few instances of red dots in both cases. This behaviour is anticipated since the ITS-G5 is known for its reduced latencies, designed specifically for real-time applications.

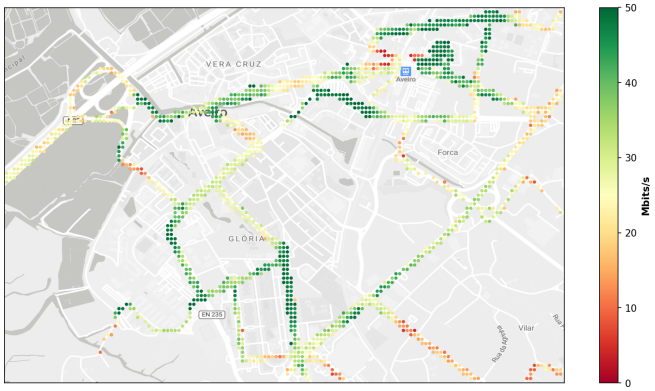
3) 5G Data

Regarding the 5G data obtained in the data collection campaigns, these are obtained from network performance tests carried out between the cellular interface of the mobile nodes (OBUs) of the ATCLL platform and the cell towers of the commercial 5G infrastructure.

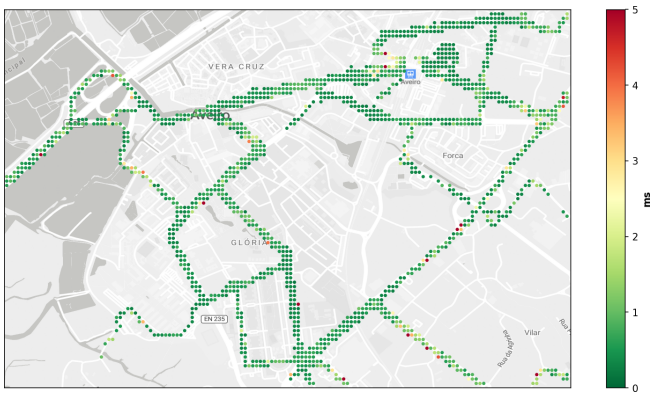
The collected 5G data is also obtained from the Iperf3 tool that represents a file transfer (UDP) in uplink. However, one characteristic of the traffic generated is different from the ITS-G5. It was determined to limit 5G to 50 Mbps (a value 5x higher than the ITS-G5 limit) to avoid saturation of the commercial 5G network. This value allows it to distance itself from the ITS-G5 without needing to exploit the maximum that the 5G network can reach.

The results of the data collection campaigns can be observed in a map format in Figure 6, which includes the four metrics: bitrate, jitter, packet loss and RTT.

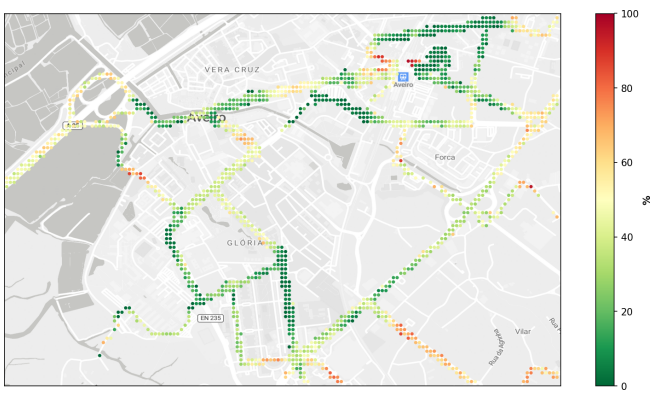
- In subfigure 6(a), concerning bitrate, it is possible to observe the extensive coverage of the cellular network and its superiority compared to ITS-G5. Here, coverage is solely affected by bus routes, which gather data, rather than the infrastructure itself. Regarding bitrate, it reaches significantly higher values but remains constrained by the imposed limit of 50 Mbps. While we lack access to the precise locations of 5G cells, areas with reduced bitrate suggest their distance or connection to cellular technologies with lower rates, such as 4G or even 3G.
- In subfigure 6(b), concerning jitter, the map displays a predominant dark green color. The sensitivity of this metric to distance is nearly insignificant, as the majority of points fall below the 1 ms



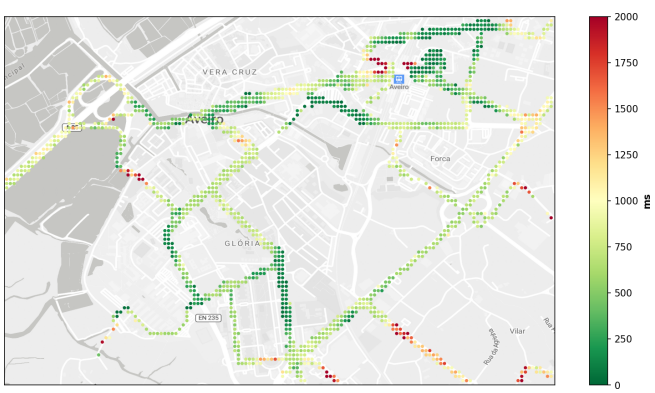
(a) Bitrate (Mbps).



(b) Jitter (ms)



(c) Packet Loss (%).



(d) RTT (ms).

FIGURE 6. Cellular KPIs.

threshold. Occasional red dots may suggest potential handovers between cells or cellular technologies. Among all the metrics, this one presents the least variability.

- In subfigures 6(c) and 6(d), concerning packet loss and RTT, we can observe a pattern similar to that of bitrate. This pattern is characterized by increased packet loss and RTT as the distance from 4G/5G cells increases. Near the cells, packet losses remain below 20 % and RTT stays under 250 ms. However, as the distance increases, packet loss rises to 40 % and RTT values increase to 750 ms.

4) ITS-G5 vs 5G

In the world of vehicular networks, the two technologies that stand out are the ITS-G5 and cellular. Each technology offers distinct advantages and disadvantages, serving diverse needs within the transportation domain. The last two sections have enabled us to draw some conclusions into the distinct attributes of each technology. In terms of coverage and bitrate, the cellular technology presents a significant advantage, saturating the entire city with notably higher bitrate values, despite capped at 50 Mbps. On the other hand, ITS-G5, designed specifically for vehicular environments, achieves substantially lower RTT values, which allows for rapid exchange of safety-critical information, making it particularly well-suited for such applications.

IV. MACHINE LEARNING MODEL TRAINING

The proposed approach uses ML to estimate the network performance metrics. It is then necessary to analyze all achievable relevant information to fulfil the proposed goal. This information is then used to train different models that return estimates of network performance values that indicate the state of the network. Since it is an approach where the desired outcomes are obtainable, multiple supervised ML algorithms are tested to try to get the best achievable estimates. In this section, the features and labels of the models are discussed in greater detail, followed by the algorithms used with their respective advantages and disadvantages. Finally, an analysis of the performance of these models is performed.

A. MODELS FEATURES & LABELS

ML algorithms build models based on sample data, known as training data, to make predictions or decisions. As this is a supervised ML problem, the training data consists of features and labels. Features are independent variables that act as inputs to the system. Labels are the expected outputs of the system. Models trained from this data generate functions that try to approximate the relationship between inputs and outputs observable in the data.

Regarding the outputs of the models, i.e. the labels, these are the same regardless of the technology. The considered labels are presented in Table 1.

TABLE 1. Labels.

Labels	Description
bitrate	Bitrate (Mbps).
jitter	Jitter (ms).
loss	Packet loss (%).
RTT	Round-Trip Time (ms).

Each of the four ML models will predict a specific network metric (bitrate, jitter, packet loss, and RTT) as their respective output labels.

As mentioned in the previous section, the data obtained for each technology (ITS-G5 and 5G) are not the same. Therefore, it is natural that the features differ between technologies.

1) ITS-G5

In the case of the ITS-G5, it is possible to extract from the training data the following features presented in Table 2. This table contains a set of location data (GPS coordinates, direction, speed, and distance between the OBU and RSUs), as well as signal data, which indicates the signal quality (RSSI) with a given RSU.

TABLE 2. ITS-G5 features.

Features	Description
Latitude	Latitude geographical coordinate.
Longitude	Longitude geographical coordinate.
Heading	Indicates the direction of the OBU (degrees).
Speed	OBU speed (m/s)
Dist	Distance between OBU and RSU (meters).
RSSI	RSU's Received Signal Strength Indicator (dBm).

The level of correlation, that is, of the interrelation between the various features/labels, is given in Figure 7. In this Figure, it is possible to observe that the features that have the highest (positive or negative) correlation with the labels are the RSSI and the distance, even though these are low correlations. Location features have very little influence on labels, with these having almost zero correlation values. The exception case is that of packet loss, where latitude and longitude also have a so-called low correlation.

The correlation matrix allows us to perform various analyses between labels and key features to highlight potential interesting behaviors.

Figure 8 shows the relationship between bitrate and RSSI. The bitrate exhibits a relatively linear relationship with signal quality (RSSI), demonstrated by the correlation value of 0.35. The RSSI values are asymmetrically distributed, with a peak density around -84 dBm.

Figure 9 illustrates the bitrate's behaviour relative to distance. The trend remains linear up to 300 meters, deviating beyond this threshold. Additionally, the density of values decreases significantly beyond this distance due

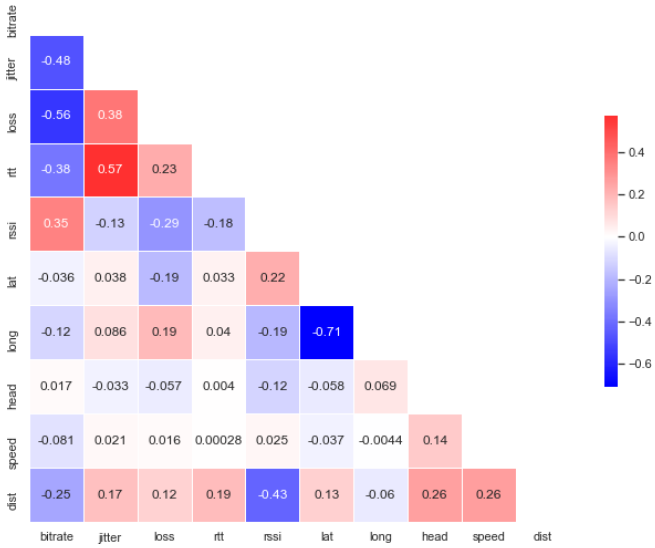


FIGURE 7. ITS-G5 labels/features correlation matrix.

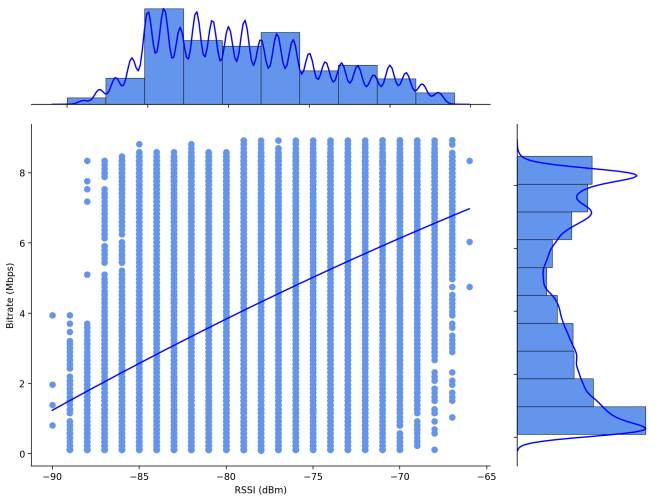


FIGURE 8. Bitrate over RSSI | Histograms/PDFs of Bitrate and RSSI (ITS-G5)

to challenges in maintaining connectivity in urban areas with obstructions at greater distances.

2) 5G Cellular

Regarding 5G, it is possible to extract from the training data the following features presented in Table 3. This also contains a smaller set of location data (GPS coordinates, direction, speed). The big difference compared to the ITS-G5 is that the signal quality data includes a higher number of metrics, including RSRP, RSRQ and SNR.

Regarding the level of correlation between the various features/labels, this is presented in Figure 10. It is possible to observe that both the RSRP and the SNR features have an average correlation level with the bitrate, packet loss and RTT. Regarding jitter, no feature is linked to its behaviour. The RSRQ has almost zero correlation,

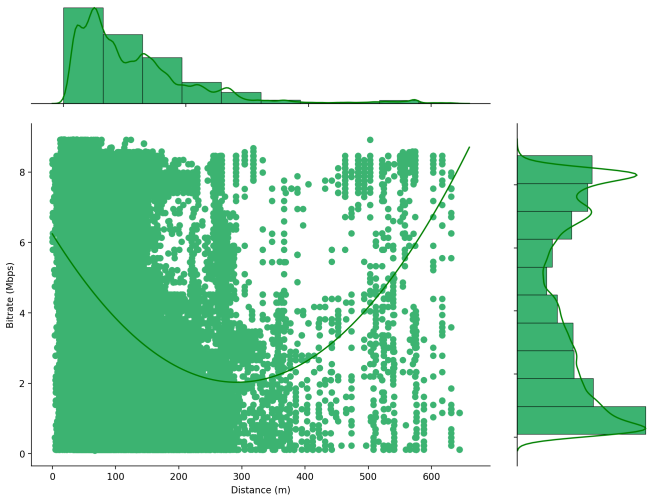


FIGURE 9. Bitrate over Distance | Histograms/PDFs of Bitrate and Distance (ITS-G5).

TABLE 3. 5G features.

Features	Description
RSRP	Reference Signal Received Power (dBm).
RSRQ	Reference Signal Received Quality (dB).
SNR	Signal to Noise Ratio (dB).

with no behavioural link between this feature and the labels.

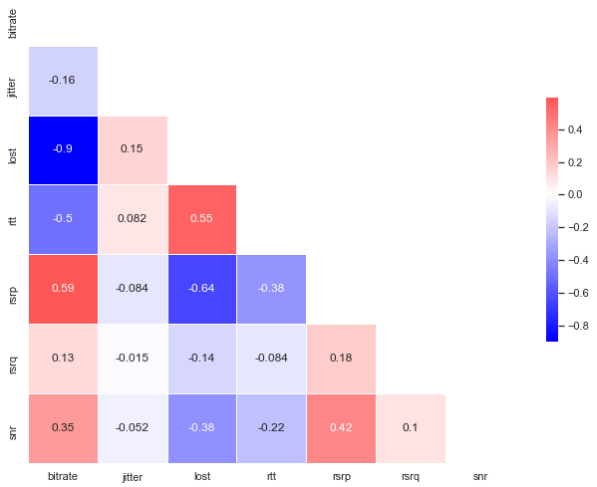


FIGURE 10. 5G labels/features correlation matrix.

As depicted in Figure 11, the bitrate demonstrates a linear variation with the signal quality, stabilizing towards the end for elevated RSRP values. The observed stabilization can be explained by the imposed 50Mbps limitation in the iPerf tests conducted for data acquisition.

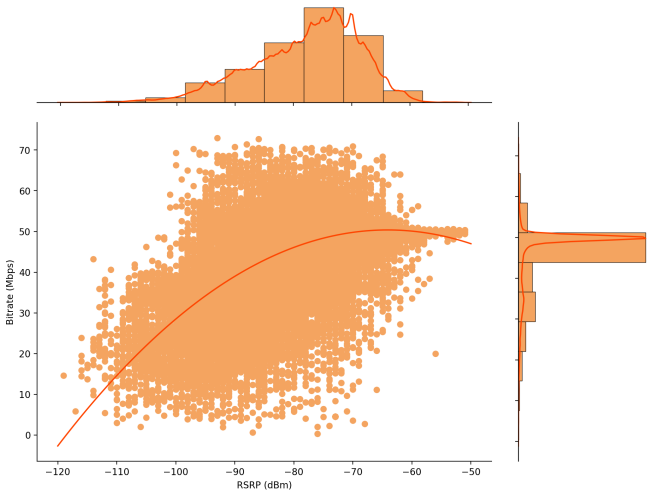


FIGURE 11. Bitrate over RSRP | Histograms/PDFs of Bitrate and RSRP (Cellular)

B. ALGORITHMS

The proposed approach aims to obtain a prediction of network metrics (bitrate, jitter, packet loss and RTT) from ML algorithms. Having a history of desirable outputs (network metrics), we can consider supervised ML algorithms. The outputs are continuous values, which make this problem a regression problem.

This work examines two ML algorithms separately, Random Forest Regression and XGBoost, to determine which one yielded the most favorable results for the dataset.

1) Random Forest Regression

Random Forest Regression is a machine learning algorithm that combines the principles of random forests and regression analysis. It is a supervised learning algorithm that uses the ensemble learning method, which is a technique that combines predictions from multiple machine learning algorithms to make a more precise prediction than a single model algorithm.

The algorithm works by building several decision trees during the training time. Each decision tree is built using a random subset of the training data and a random subset of features. In the end, it outputs the mean of the predictions of all the trees [11].

This algorithm has multiple advantages that make it ideal for the proposed problem. The qualities are as follows:

- Robustness: It can handle large and complex amounts of data, including numerous features, without over-fitting.
- Tolerance to outliers: The fact that the algorithm is based on multiple trees makes it less sensitive to outliers.

However, the algorithm also has some limitations:

- Memory usage: Although multiple trees improve prediction, they increase memory substantially.
- Computationally Demanding: The creation and training of models are computationally demanding due to the algorithm's complexity. For this reason, the prediction time is high.

2) XGBoost

XGBoost (Extreme Gradient Boosting) is a machine learning algorithm designed for gradient boosting that is widely used in data science. It is also an ensemble of decision trees, just like Random Forest Regression. The algorithm is designed to be highly efficient, flexible and portable, outperforming other machine learning algorithms in most tasks, including classification, regression and ranking.

The algorithm integrates several weak learners (decision trees) into one strong learner. One weak learner is trained on a subset of the data but has a poor performance. The second weak tree is trained to predict what the first tree was not capable of. The predictions from each weak learner are combined to form a strong learner that is much more accurate.

This algorithm has numerous advantages, such as:

- High accuracy: It is a popular choice for machine learning tasks that require high precision.
- Efficiency: It is designed to be fast and efficient, even for large datasets. It is an excellent choice for tasks that require fast predictions.

However, the algorithm also has a few limitations:

- Overfitting: Without regularization techniques, it is possible for the algorithm to overfit the training data, which can lead to inaccurate predictions of new data.

Initially, we considered using Linear Regression (LR) and Artificial Neural Network (ANN) algorithms, but these were quickly disregarded due to their specific limitations.

For LR, the weak correlation between features indicates low linearity, which compromises the effectiveness of this algorithm as it relies on linear assumptions. Additionally, its simplicity increases the risk of overfitting, a problem less pronounced in Random Forest Regression (RFR) and Extreme Gradient Boosting (XGBoost) algorithms used in our work.

As for ANN, its requirement for a substantial amount of high-quality training data to achieve good results makes it unsuitable for our case. Our datasets are relatively small and contain outliers, which significantly reduces the performance of ANN.

Due to these reasons, we did not explore LR and ANN further in our research.

C. MODELS TRAINING

To turn data into ML models capable of making reliable predictions, it is necessary to use techniques that correctly validate the performance of the models.

Traditionally, the model validation methodology is based on a "Train-Validation-Test Split" technique. This technique divides the dataset into three parts, one for training the models, another for validation used to tune the model's hyperparameters, and a last one for testing, used to evaluate the final performance of the model. However, this methodology is quite limited in cases where the dataset is reduced, with only a fraction used for the actual training.

Another technique that overcomes this problem is cross-validation, which helps to evaluate how well a trained model performs on unseen data by running the model's performance on multiple subsets of the available data.

The most common form of cross-validation is k-fold cross-validation. This technique divides the dataset into k equally sized folds. The model is trained and evaluated k times, with each fold serving as the validation set while the remaining folds are used for training. Finally, an average of all iterations is made, thus evaluating the overall model's performance. This technique offers a more robust estimation since it uses multiple train-validation splits instead of relying on a single one. It also prevents overfitting as it uses different subsets of the data. Consequently, this technique is used cooperatively with both XGBoost and Random Forest Regression algorithms to obtain the best possible models.

In addition to cross-validation, hyperparameter tuning is essential for optimizing the model's performance. One effective method for this, used in this work, is RandomizedSearchCV. This method performs a random search over specified hyperparameter values, evaluating each combination using cross-validation. By testing a fixed number of hyperparameter combinations, it efficiently explores the hyperparameter space. This method often finds good hyperparameters faster than the grid search method, which is more computationally intensive. When used with k-fold cross-validation, RandomizedSearchCV helps identify the optimal hyperparameters, therefore enhancing the performance of models.

V. PERFORMANCE RESULTS

This section delves into the multifaceted evaluation of machine learning models performance. It is divided into two parts, each focusing on a specific communication technology (ITS-G5 and 5G). The analysis aims to identify the optimal features of each technology for achieving the most favourable results.

A. ITS-G5 MODELS

Regarding the ITS-G5 data and the corresponding models, multiple feature configurations were considered: GPS + RSSI + DIST, GPS + RSSI, RSSI + DIST, RSSI, DIST. The choice of these settings is based partially on the correlation matrix between features and labels, addressed in Figure 7. This correlation matrix examines the relationships and dependencies among multiple variables, in this case, features and labels of the models studied here.

In Figure 12 below, we observe a segment of the test path, showcasing the bitrate over time alongside predictions based on various features.

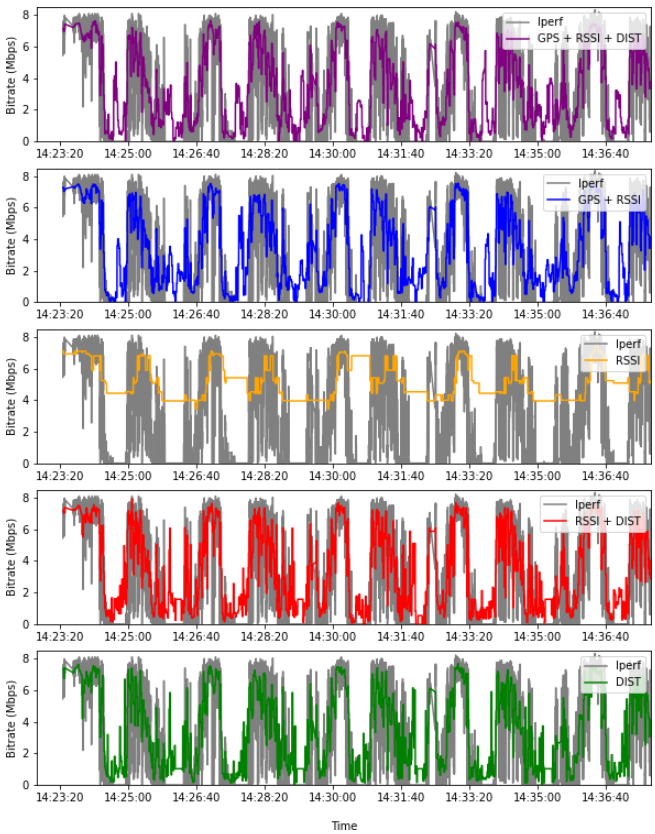


FIGURE 12. ITS-G5 Bitrate over time (XGBoost).

Each sub-figure features a grey curve representing values from iPerf tests. Above each one, a uniquely coloured curve illustrates the forecast of each ML model configuration.

Upon closer inspection, it becomes apparent that the yellow configuration (RSSI - subplot 3) exhibits notable deviations from expectations, producing predictions confined to a specific range. This simplistic model, with a sole feature and a limited RSSI range, delivers less reliable forecasts, thus qualifying as the least effective model among all evaluated. Conversely, the other feature configurations (subplots 1, 2, 4, and 5) demonstrate relatively consistent alignment, delivering comparable outcomes. These curves provide valuable insights into the model's behaviour and associated errors, but a definitive assessment of model performance needs

a closer examination of actual error values (RMSE).

The actual errors (RMSE) are calculated for both algorithms across each configuration/label, providing insights into the actual performance of the configurations, as outlined in Tables 4 and 5.

TABLE 4. ITS-G5 RMSE for each model/configuration (XGBoost).

Configuration	Bitrate	Jitter	Loss	RTT
GPS RSSI DIST	1.23 Mbps	96.44 ms	13.93 %	131.45 ms
GPS RSSI	1.23 Mbps	96.20 ms	14.05 %	134.12 ms
RSSI	2.60 Mbps	252.49 ms	23.49 %	189.33 ms
RSSI DIST	1.28 Mbps	111.41 ms	14.35 %	135.46 ms
DIST	1.50 Mbps	163.08 ms	16.33 %	141.71 ms

TABLE 5. ITS-G5 RMSE for each model/configuration (Random Forest Regression).

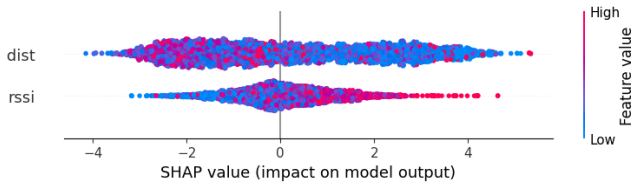
Configuration	Bitrate	Jitter	Loss	RTT
GPS RSSI DIST	1.23 Mbps	96.04 ms	13.97 %	132.92 ms
GPS RSSI	1.24 Mbps	96.03 ms	14.04 %	134.93 ms
RSSI	2.60 Mbps	252.51 ms	23.49 %	189.33 ms
RSSI DIST	1.29 Mbps	115.00 ms	14.56 %	137.00 ms
DIST	1.50 Mbps	164.24 ms	16.41 %	141.72 ms

The outcomes for each configuration/label are remarkably similar for both algorithms, deserving a joint consideration of the results.

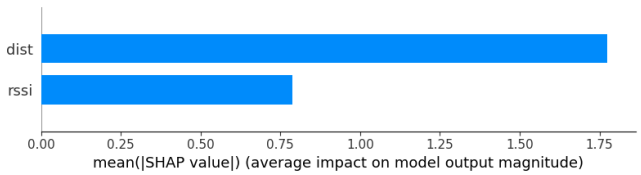
In terms of bitrate, the values align with the conclusions drawn from the preceding figure. The (RSSI) configuration exhibits the highest error at 2.60 Mbps, while configs (GPS + RSSI + DIST, GPS + RSSI, and RSSI + DIST) show closely clustered values ranging from 1.23 to 1.29 Mbps. The (DIST) configuration slightly surpasses these with an error of 1.50 Mbps. Consequently, it is evident that, across both algorithms, the configurations yielding the best result (smallest RMSE) for bitrate prediction is (GPS + RSSI + DIST), with an RMSE of 1.23 Mbps. For the remaining labels, (RSSI) configuration consistently yields the least favourable outcomes across all scenarios. GPS-based configurations consistently achieve the lowest errors, with (RSSI + DIST) configurations closely following, displaying nearly identical values in the case of packet loss and RTT at 14.6 % and 137 ms, respectively.

These features remain independent of the specific location in which data was collected, enhancing their overall generality. While GPS-based configurations yield the lowest RMSE, they tend to be less versatile across different environments. Consequently, even though the (RSSI + DIST) configuration produces slightly less promising results in the city of Aveiro, it could potentially yield the best outcomes in any other city.

By conducting a SHapley Additive exPlanations (SHAP) analysis, used to interpret the output of ML models, some of the conclusions drawn from this work can be confirmed.



(a) Beeswarm plot.



(b) Bar plot.

FIGURE 13. ITS-G5 SHAP summary for the bitrate model (XGBoost).

In Figure 13(a), which depicts a SHAP analysis for the bitrate model, it is evident that the DIST feature plays a significant role: high distance values (red dots) reduce the model's performance, whereas low distance values (blue dots) enhance it. By contrast, for the RSSI feature, higher values improve performance, while lower values diminish it. This behavior concerning distance was previously observed in Figure 9, which indicates that higher distance values tend to exhibit more random behavior due to the limited amount of data available for these higher values.

Figure 13(b), which presents the same SHAP analysis but in a bar plot format, confirms that the DIST feature contributes the most to the model. This behavior is also evident in Tables 4 and 5, where models based solely on the DIST feature achieve error values not too far apart.

B. CELLULAR MODELS

Regarding cellular configurations, the following setups are considered: (RSRP + RSRQ + SNR, RSRP + SNR, RSRP). By examining the cellular correlation matrix in Figure 10, it is evident that RSRP shows the strongest correlation with the labels. On the other hand, SNR and RSRQ exhibit a lower impact individually. However, when combined with RSRP, they contribute to improving the model's performance.

The actual errors (RMSE) were calculated for both algorithms across each configuration/label, providing insights into the realistic performance of the configurations, as outlined in Tables 6 and 7.

The outcomes for each configuration/label are remarkably similar for both algorithms, deserving a joint consideration of the results.

TABLE 6. Cellular RMSE for each model/configuration (XGBoost).

Configuration	Bitrate	Jitter	Loss	RTT
RSRP RSRQ SNR	6.17 Mbps	2.65 ms	10.34 %	268.86 ms
RSRP SNR	6.24 Mbps	2.65 ms	10.45 %	269.58 ms
RSRP	6.52 Mbps	2.65 ms	11.22 %	250.46 ms

TABLE 7. Cellular RMSE for each model/configuration (Random Forest Regression).

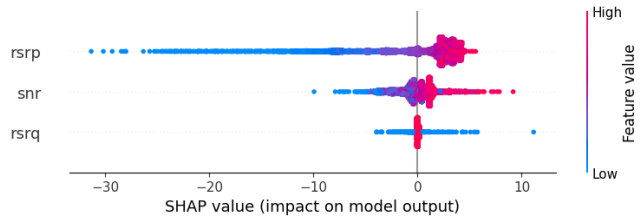
Configuration	Bitrate	Jitter	Loss	RTT
RSRP RSRQ SNR	6.15 Mbps	2.54 ms	10.31 %	275.22 ms
RSRP SNR	6.24 Mbps	2.56 ms	10.47 %	275.88 ms
RSRP	6.52 Mbps	2.65 ms	11.22 %	250.35 ms

Regarding bitrate, the configuration (RSRP + RSRQ + SNR) emerges with top results, delivering 6.17 Mbps and 6.15 Mbps of RMSE for XGBoost and RFRegression, respectively. The other configurations produce competitive results, with RMSE ranging from 6.24 to 6.52 Mbps, showcasing their proximity in performance. For the remaining labels, the (RSRP) configuration achieved the lowest RMSE value for RTT but lagged behind in comparison to the other labels. Notably, jitter remains unchanged in the XGBoost algorithm, while for the RFRegression algorithm, a variation is observed only in the (RSRP) configuration, resulting in a slightly higher error than the others.

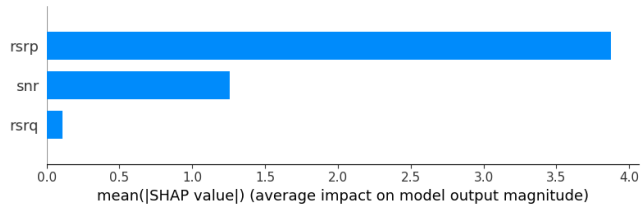
Although RSRQ and SNR metrics exhibit minimal impact on the labels individually, as shown in the correlation matrix (Figure 10), their combined effect with RSRP contributes to a slight improvement in RMSE. It clarifies why the configuration containing the highest number of features achieves the best RMSE value.

Since none of the configurations relies on GPS coordinates, there are no geographical constraints on where these models can operate, as these are universal metrics independent of location. Consequently, the RMSE values presented remain applicable even beyond the area used for training the models, in this case, the city of Aveiro.

Figure 14 presents a SHAP analysis for the cellular bitrate ML model obtained with XGBoost.



(a) Beeswarm plot.



(b) Bar plot.

FIGURE 14. Cellular SHAP summary for the bitrate model (XGBoost).

In Figure 14(a), it is clear that the RSRP feature plays a significant role. For both RSRP and RSRQ features, high values (red dots) enhance the model's performance, whereas low values (blue dots) reduce it. The SNR feature is more ambiguous due to its short range and neutral SHAP values.

Figure 14(b) confirms the conclusions drawn from Tables 6 and 7: RSRP is by far the most significant positive contributor to the model's performance, whereas RSRQ has a much lower impact, and SNR is almost negligible.

C. PERFORMANCE CONCLUSIONS

The analysis conducted in this section enables the discovery of, not only the optimal algorithm, but also the most effective combination of features, leading to the creation of models that produce satisfactory results.

1) Algorithms

The two discussed algorithms delivered remarkably similar results, with nearly identical RMSE values for the same configurations. Despite their promising predictive performance, these algorithms are distinguished by certain obstacles, namely memory usage and estimation time. Table 8 illustrates these distinct characteristics for each algorithm.

Clearly, the RFRegression algorithm models demand significantly more memory space, approximately 12600 % more than the XGBoost algorithm models, despite achieving very similar results. A similar pattern emerges in terms of prediction time, where the RFRegression algorithm requires 2200 % more time than XGBoost. These results shed light on the inherent drawbacks these obstacles pose to the scalability of the number of models used. Therefore, in this article, the XGBoost algorithm is recognized as the best choice,

TABLE 8. Algorithms Characteristics.

Algorithm	Average Size p/model	Average Estimation Time p/model *
RFRegression	533,9 MB	538,1 ms
XGBoost	4,2 MB	23,2 ms

* estimation time on Macbook Air M1 (2021). Results were not obtainable in OBU devices, namely PCEngines APU3, for the RFRegression algorithm due to lack of memory.

providing an optimal ratio with the lowest RMSE for both memory usage and prediction time.

2) Models

For the two technologies discussed here, a variety of feature configurations were explored with the goal of attaining the most satisfactory predictions possible. Across both technologies, features based on the signal quality and distance characteristics exhibited superior performance coupled with broad generalizability. While GPS-based features showcased promising results, their utility was constrained to the areas where data was acquired, thus restricting the generality of the models created.

The increased complexity of ML models enables them to capture complex relationships, thereby facilitating more precise estimation of QoS metrics compared to traditional heuristics. The latter relies on predefined rules or heuristics, constraining their efficacy. Models generated in this study bypasses these limitations with predictors based on real data.

VI. CONCLUSION

This article researched an ML-based solution that estimates network performance metrics in a city environment. These estimates pave the way for intelligent and efficient control of the vehicular network. To achieve this goal, multiple data collections were carried out in the city of Aveiro using a set of vehicles dedicated to that purpose. Data collections include vehicular and network performance data for ITS-G5 and 5G technologies. The collected data was analyzed to determine which ML algorithms were most suitable. Several feature configurations were also considered to obtain the best possible results.

Performance results showed that, for both technologies, configurations based on signal quality and distance characteristics offer both excellent performance with high generality. Although GPS-based feature configurations provide the best results, these are restricted to the area where the data was collected, thus being limited. While both algorithms, RFRegression and XGBoost, achieved comparable results, the drawbacks related to memory and prediction time led the choice away from the RFRegression algorithm in favour of XGBoost. These results conclude that it is possible to transform vehicular domain data into usable information that can be used to predict the quality of diverse available links.

Future work based on the ML models generated here may have different applications. The area that makes the most sense is network management. Estimates of network metrics can dynamically allocate bandwidth, manage power consumption, and optimize communication protocols based on the specific requirements of the vehicular network. Another potential area for future work involves assessing the models in scenarios beyond those used for the data collection campaign. Such an evaluation would enable the effective validation of the models applicability in cities other than the one used for training.

REFERENCES

[1] L. Liang, H. Ye, and G. Y. Li, "Toward intelligent vehicular networks: A machine learning framework," IEEE Internet of Things Journal, vol. 6, no. 1, pp. 124–135, 2019.

[2] G. Cerar, M. Mohorcic, T. Gale, and C. Fortuna, "Link quality estimation using machine learning," CoRR, vol. abs/1812.08856, 2018. [Online]. Available: http://arxiv.org/abs/1812.08856

[3] R. Hernangómez, A. Palaios, G. Guruvayoorappan, M. Kasparick, N. U. Ain, and S. Stańczak, "Online qos estimation for vehicular radio environments," in 2022 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 1701–1705.

[4] R. Chakroun, T. Villemur, and K. Benoit Nougnanke, "Learning-based infrastructure to vehicle link quality estimation," in 31st International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2023). Split, Croatia: University of Split, FESB and Croatian Communications and Information Society (CCIS), sep 2023.

[5] N. R. Sudesh Pahal and B. Singh, "A deep learning-based model for link quality estimation in vehicular networks," IETE Journal of Research, vol. 69, no. 8, pp. 5159–5168, 2023.

[6] Y.-C. Chan, J. Wu, E. W. Wong, and C. S. Leung, "Integrating teletraffic theory with neural networks for quality-of-service evaluation in mobile networks," Applied Soft Computing, vol. 152, p. 111208, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623012267

[7] E. D. Salık, G. Görbilek, B. Okur, and A. G. Önalan, "Non-wifi interference detection and throughput estimation at the wifi edge for 2.4 and 5 ghz bands with machine learning," in 2023 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), 2023, pp. 371–378.

[8] E. N. Almeida, K. Fernandes, F. Andrade, P. Silva, R. Campos, and M. Ricardo, "A machine learning based quality of service estimator for aerial wireless networks," in 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 2019, pp. 1–6.

[9] P. Rito, A. Almeida, A. Figueiredo, C. Gomes, P. Teixeira, R. Rosmaninho, R. Lopes, D. Dias, G. Vítor, G. Perna, M. Silva, C. Senna, D. Raposo, M. Luís, S. Sargento, A. Oliveira, and N. B. de Carvalho, "Aveiro Tech City Living Lab: A Communication, Sensing, and Computing Platform for City Environments," IEEE Internet of Things Journal, vol. 10, no. 15, pp. 13 489–13 510, 2023.

[10] D. Dias, R. Rosmaninho, A. Figueiredo, P. Almeida, M. Luís, P. Rito, D. Raposo, and S. Sargento, "A Dataset of ITS-G5 and Cellular Vehicular Connectivity in Urban Environment," Data in Brief, p. 109846, 2023.

[11] "Random forest regression," https://levelup.gitconnected.com/random-forest-regression-209c0f354c84, online; accessed March 2023.

DUARTE DIAS received the M.Sc. degree in Computers and Telematics Engineering from the University of Aveiro, Portugal, in 2021. Since 2020, he has been a Researcher with Instituto de Telecomunicações (IT), Aveiro, Portugal, exploring areas such as Software defined Vehicular Networks, Machine Learning in Vehicular Networks and Quality-of-Service. Currently, he's associated with research projects such as IMMINENCE (Celtic-NEXT program) and Route 25 (PRR Agenda). His current research interests include vehicular communications, network management, virtualization and smart cities.

MIGUEL LUÍS received the M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from the Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal, in 2009 and 2015, respectively. He is an Assistant Professor in Instituto Superior Técnico (IST) and Researcher with Instituto de Telecomunicações (IT), and has been involved in several national and European research projects targeting new communications for mobile networks. Currently, he is the coordinator of "MH-SDVanet: Multihomed Software Defined Vehicular Networks", a national funded research project, and he contributes to several other research projects such as SNOB-5G (FCT-MIT program), IMMINENCE (Celtic-NEXT program), CityCatalist and POWER (P2020 program) and Route 25 and New Space (PRR Agenda), to name a few. Miguel has published more than 90 scientific works, including 3 book chapters and 45 publications in peer-reviewed international journals. His research interests include medium access control for wireless systems, routing and dissemination mechanisms for mobile networks and management, orchestration and softwarization of future networks.

PEDRO RITO received the M.Sc. degree in Electronics and Telecommunications Engineering from University of Aveiro, Portugal and Ph.D. degree in Electrical Engineering from Technische Universität Berlin, Germany, in 2011 and 2019, respectively. In 2012 he joined IHP GmbH, Frankfurt (Oder), Germany, as a Researcher, and in 2018 he joined Cisco Optical GmbH, Nürnberg, Germany, as a Hardware Engineer, investigating and developing state-of-the-art high-efficient and high-bandwidth electro-optical interconnects for datacenters, metro and long-haul communications. From 2020, he has been with Instituto de Telecomunicações (IT) in Aveiro, Portugal, as an Assistant Researcher in the Network Architectures and Protocols group. During his research activity he has published more than 50 publications (20 of which in peer-reviewed journals) and he holds 1 US/EU patent. His current research interests include radio access networks, software-defined networking, vehicular communications, edge computing and smart cities.

SUSANA SARGENTO is a Full Professor in the University of Aveiro and a senior researcher in the Instituto de Telecomunicações, where she is leading the Network Architectures and Protocols group. She was a visiting PhD student in Rice University (2000-2001), and a Guest Faculty in Carnegie Mellon University (2008). Susana has been leading research projects with telecom operators and OEMs. She has been involved in several FP7 projects (4WARD, Euro-NF, C-Cast, WIP, Daidalos, C-Mobile), EU Coordinated Support Action 2012-316296 "FUTURE-CITIES", EU Horizon 2020 5GinFire, EU Steam City, and CMU-Portugal projects (S2MovingCity, DRIVE-IN with the Carnegie Melon University) and MIT-Portugal Snob5G project. She has organized several international conferences and workshops, such as ACM MobiCom, IEEE Globecom, and has also been a reviewer of conferences and journals, such as IEEE Networks, IEEE Communications. Susana has co-founded a vehicular networking company in 2012, Veniam (www.veniam.com), she is the winner of the 2016 EU Prize for Women Innovators, and the winner of Femina 2020 prize in Science. She was the co-coordinator of the national initiative of digital competences in the research axis INCoDe.2030, belonged to the evaluation committee of the Fundo200M (www.200m.pt), and she is one of the Scientific Directors of CMU-Portugal Programme. Her main research interests, with more than 400 scientific papers, are in the areas of self-organized networks, Intelligent Transportation Systems, 5G and beyond networks and services, and content distribution networks. She regularly acts as an Expert for European Research Programmes.