ISBN 978-989-9121-55-3
64th ISTANBUL World Congress on Innovations in "Science, Engineering, Technology & Education" (WCISET-25)
Nov. 13-14, 2025 Istanbul (Türkiye)

Real-Time Traffic-Aware Routing in 5G VANETs Using Regression and Classification Models

Iclal Cetin Tas[1]

[1]Department of Computer Engineering, Başkent University, Ankara, Türkiye

Abstract**:** *Smart transportation systems have become a important research area for safe, efficient, and sustainable traffic management. 5G-enabled Vehicular Ad-hoc Networks (VANETs) and Multi-Access Edge Computing (MEC) are emerging as the leading technologies to address the need for real-time and ultra-low latency communication in these systems. 5G-enabled VANETs and MEC offer new opportunities to support ultra-low latency applications in intelligent transportation systems. In this context, efficient route selection and accurate end-to-end (E2E) delay estimation are critical for real-time and reliable vehicular communication. In this study, a machine learning-based approach is developed using the 5G-VANET Real-Time Routing Dataset. Regression and classifications methods are applied to prediction of E2E delay, and to identify the most efficient routes, respectively. Experimental findings reveal that ensemble learning methods provide the highest performance in While low error rates were achieved in the regression phase, optimal route selection was achieved with accuracy rates of up to 99% in the classification phase. The obtained results are shown that regression and classification techniques with ensemble approaches offers strong potential for intelligent route optimization in 5G-supported VANET-MEC environments.*

Keywords: *VANET, Real-Time Routing, Smart Transportation, Ensemble Learning,*

1. Introduction

In smart cities, communication and low-latency data transmission between vehicles are made possible by 5G technology and Vehicular Ad-hoc Networks (VANETs). Delay-free delivery of real-time traffic information, safety messages, and services is crucial [1]. The constant movement of vehicles, their high speeds, and varying traffic densities lead to frequent changes in network topology, leading to performance issues in applications requiring reliable communication, low latency, and high data throughput [2], [3]. This makes it difficult to estimate end-to-end (E2E) delay and make optimal routing decisions. It is important to optimize routing protocols based on criteria such as delay reduction, link availability, and reliability.

Various machine learning-based approaches have been developed in the literature to address these issues. Kandali et al. [2] proposed a new clustering algorithm based on density vertex clustering (DPC) and particle swarm optimization (PSO) to ensure network stability. In their study, PSO was used to find optimal solutions for cluster head selection, followed by DPC for clustering based on link reliability. Evaluated through simulations, the method achieved a 74% reduction in cluster handover rates, a 34% increase in intra-cluster data transfers, and a 47% increase in inter-cluster transfers, while also reducing average latency by 16%.

Tang et al. [3] focused on latency reduction in heterogeneous VANETs using machine learning-based mobility prediction. In this study, a centralized architecture based on software-defined networking (SDN) was proposed, and advanced artificial neural networks (ANNs) were used to predict mobility based on vehicle location and speed. These predictions were used to determine optimal routes via roadside units (RSUs) and base stations (BSs). Simulation results revealed that the proposed method offers lower transmission delays and more stable performance, even at different vehicle speeds.

Hota et al. [4], on the other hand, comprehensively examined the use of machine learning algorithms for optimization and intelligence not only in VANETs but also in various network types such as WSNs, MANETs,

VANETs, and USNs. The study examines how supervised learning, reinforcement learning (RL), and ensemble techniques are used in areas such as routing, resource allocation, error detection, attack prevention, and QoS enhancement. This highlights the importance of developing machine learning-based solutions tailored to the constraints of different network types.

While existing work has addressed important problems in VANETs, such as delay reduction, clustering, and mobility prediction, holistic approaches that address both E2E delay prediction and route optimization are limited in the literature[5], [6], [7], [8]. A comprehensive comparison of the performance of different machine learning algorithms on these two problems on the same dataset has also not been sufficiently investigated.

In this study, both E2E delay estimation (regression) and the most efficient route selection (classification) steps were performed using the 5G VANET Real-Time Routing Dataset. In the regression step, E2E delay estimation was performed using methods such as decision trees, stepwise linear regression, and Gaussian Process Regression. In the classification step, optimal routes were determined using Decision Tree, Random Forest (Bagging), RUSBoosted, Neural Network, Naive Bayes, and SVM algorithms. Additionally, the performance of different models was compared, and it was observed that ensemble approaches yielded superior results in both regression and classification steps.

The article follows the following structure: Section 2 details the dataset, and the methodology section explains the regression and classification steps in detail. Section 3 presents the experimental procedure, results, analyzing model performance and discussing them. Finally, Section 4 contains the conclusions section.

2. Materials and Methods

2.1. Materials

In this study, we used the open-access dataset called 5G-VANET Real-Time Routing Dataset [9]. The dataset was generated for real-time vehicular communication based on VANETs and MEC scenarios. It includes variables such as vehicle mobility (speed, location, direction), network conditions (delay, bandwidth, packet loss), and edge computing parameters (processing capacity, queue length). These parameters were generated under different traffic densities and network scenarios to simulate applications requiring ultra-low latency. The dataset contains approximately N samples, with each record representing a communication scenario. E2E delay estimation is regression step target and efficient routes is classification step target.

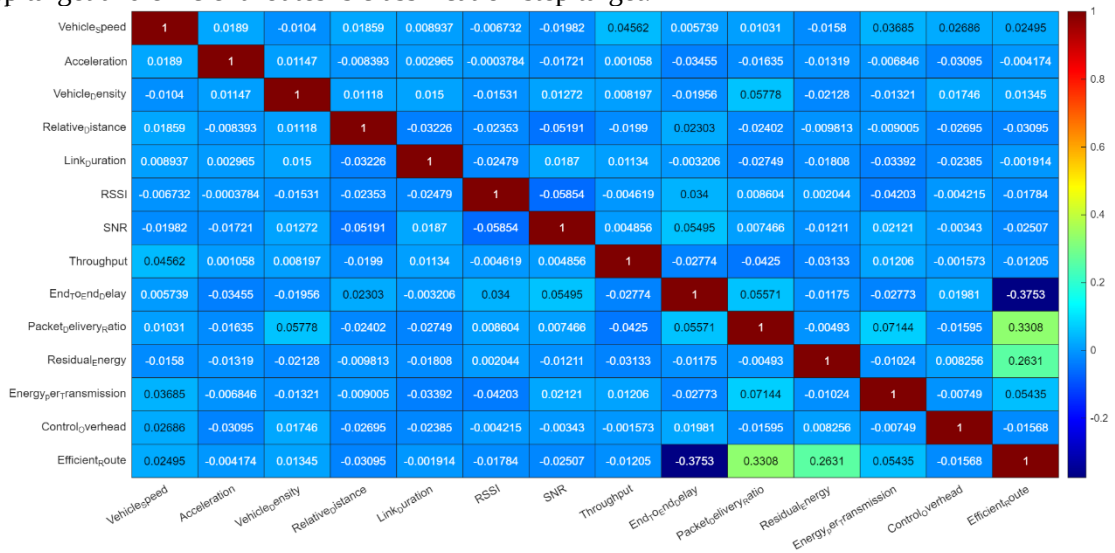

Fig. 1: Correlation matrix of dataset

2.2. Methods

In this study, we perform E2E delay estimation and efficient route selection using the 5G-VANET Real-Time Routing Dataset. The study consists of two main steps: a regression step and a classification step. While the regression step performs E2E delay estimation, the classification step uses delay and resource utilization information to determine the most efficient routes.

2.2.1. Regression Step

Bagging (Bootstrap Aggregating) is an ensemble learning approach that works in parallel. Multiple sub-datasets are created from the original dataset using bootstrap sampling, and independent predictors are trained on each sub-sample. The number of observations in each sub-sample is equal to the number of observations in the original dataset, and some observations may appear multiple times, while others may not appear at all. In regression problems, predictions are combined by averaging the sub-models, while in classification problems, the final decision is made using the voting (majority voting) method (in Equation 1).

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} h_m(X) \tag{1}$$

The Random Forest algorithm combines independent decision trees using the bagging approach. In regression problems, the final result is obtained by averaging the predictions of each tree, while in classification problems, the class with the most votes is selected. This structure reduces the overfitting tendency of a single decision tree and increases generalizability. Random Forest provides high accuracy and consistent performance, particularly in classification and regression problems.

Boosting algorithms, unlike ensemble learning methods, focus on each weak model correcting the errors of the previous model. A series of weak models are trained sequentially, and each model is weighted to compensate for the shortcomings of previous models (in Equation 2). This optimizes the final prediction based on the contributions of all weak models. The boosting approach relies on the collaboration of weak models within the ensemble to increase predictive power, rather than obtaining a single strong model. This method reduces errors in classification and regression tasks, resulting in higher accuracy and enhanced model generalizability.

$$\hat{y} = \sum_{m=1}^{M} a_m h_m(X) \tag{2}$$

Stepwise linear regression is an extended version of the classical linear regression model (in Equation 3). In the stepwise approach, statistical criteria are used to determine which variables to include or exclude from the model. This method ensures that the model includes only variables that contribute significantly, reducing the risk of overfitting and improving interpretability. It is an effective method for controlling model complexity, especially in high-dimensional datasets. $\beta_0, \beta_\mathrm{j}$ are constant and the regression coefficient respectively, x_j represents independent variables.

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{3}$$

Gaussian Process Regression (GPR) is a nonparametric regression method that offers a flexible and probabilistic approach to modeling nonlinear relationships. GPR assumes each observation is derived from a common Gaussian process and makes predictions by modeling the dependencies between data points via covariance functions (kernel functions (in Equation 4)).

$$\hat{y}_* = \mathrm{K}_*^T(\mathrm{K} + \sigma_n^2 I)^{-1} y \tag{4}$$

Here K is the covariance matrix for the training data, K_* is the covariance vector between the test data and the training data, σ_n^2 and I represent the observation noise, the identity matrix respectively.

2.2.2. Classification Step

Decision trees are frequently used in solving classification problems because they can create a hierarchical structure by repeatedly dividing the dataset according to features. At each node, purity metrics such as information gain or Gini index are calculated to select the most appropriate split point, and leaf nodes represent specific classes. While this method offers high interpretability, it is often supported by pruning strategies because it is prone to overfitting. Ensembles, a powerful extension of decision trees, aim to produce more stable and highly accurate results by combining multiple weak classifiers. In this context, RUSBoost (Random Under-Sampling Boosting) was developed to learn the minority class more effectively in imbalanced datasets. The algorithm performs

balancing by randomly extracting samples from the majority class and updates the weights of the sequentially trained classifiers using AdaBoost logic (in Equation 5).

$$a_m = ln\frac{1-e_m}{e_m} \tag{5}$$

Artificial neural networks are multilayered structures inspired by biological neurons and can provide high accuracy in classification problems. The features in the input layer are transferred to the hidden layers via weights and activation functions. The output of each neuron is obtained by applying activation functions such as sigmoid, ReLU, or tanh to the linear combination ($z = \sum w_i x_i + b$).The training process is based on minimizing the prediction error using a backpropagation algorithm and gradient descent. It offers an effective solution, especially for complex and nonlinear classification problems.

Naive Bayes classifiers are probability-based methods based on Bayes' theorem. In this approach, independence is assumed between the features, and conditional probabilities are calculated for each class. Bayes Theorem in Equation 6:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{6}$$

is implemented using the formula.The posterior probability for the class C_k is calculated, and the class with the highest probability is predicted. This computationally efficient method is widely used due to its fast results, especially on high-dimensional data. However, the assumption of independence between features can limit accuracy performance on some datasets.

Support vector machines are powerful supervised learning algorithms used in classification problems. The main objective is to find the hyperplane that maximizes the separation between classes. The decision boundary for binary classification is defined and SVM solves the following optimization problem to maximize the margin between classes as follows (in Equation 7):

$$f(x) = w^t x + b$$
$$min\frac{1}{2}\|w\|^2 \quad y_i(w^t x_i + b) \geq 1 \tag{7}$$

where w represents the weight vector and b represents the bias term. For non-linearly separable datasets, the data is transformed into a higher-dimensional space using kernel functions (e.g., polynomial, RBF) to make it separable.

3. Results and Discussions

In this section, the performance of the proposed approach is evaluated in terms of both regression-based E2E delay estimation and classification-based optimal route selection. The experiments were conducted on the 5G VANET Real-Time Routing Dataset, and the results obtained for the validation and test data are presented separately. RMSE, MSE and MAE, which are most frequently used in the literature, are used in performance evaluation (in Equation 9-10).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y_i}^2} \tag{8}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{10}$$

Tables I and II show the results obtained on the validation and test datasets in the regression step. Among the methods, the boosting-based ensemble model yielded the lowest error values in both stages. It achieved the best performance on the validation set with RMSE=22.78, MSE=518.96, and MAE=18.69, while it outperformed the other methods on the test set with RMSE=23.88, MSE=570.18, and MAE=20.54. These findings demonstrate that the boosting method can better capture the variability in dynamic VANET conditions by sequentially improving the errors of weak learners. Furthermore, the closeness of both the validation and test results demonstrate that the model has a low tendency towards overfitting and strong generalization. While the differences between the error

values are not significant, the fact that it yields the lowest error, especially on the test set, indicates that the boosting-based approach can provide more reliable predictions over the long term and under varying traffic densities. These results demonstrate that ensemble learning methods have stronger generalization capabilities compared to individual learners (such as decision trees, linear regression, and GPR). While the absolute error differences are not very large, boosting methods appear to provide more reliable results in dynamic VANET conditions.

TABLE I: Obtained Results for Regression Step (Validation)

Methods	RMSE (Validation)	MSE (Validation)	MAE (Validation)
Tree (Coarse)	23.67	560.43	19.30
Ensemble (Boosted)	22.78	518.96	18.69
Ensemble (Bagging)	23.05	531.45	18.80
Stepwise Linear Regression	23.01	529.63	18.93
Gaussian Process Regression	23.06	531.74	18.96

TABLE II: Obtained Results for Regression Step (Test)

Methods	RMSE (Test)	MSE (Test)	MAE (Test)
Tree (Coarse)	24.53	601.72	20.19
Ensemble (Boosted)	23.88	570.18	20.54
Ensemble (Bagging)	24.42	596.42	20.73
Stepwise Linear Regression	24.09	580.23	20.93
Gaussian Process Regression	24.28	589.28	20.93

Table III presents the results of the classification methods used in optimal route selection. The decision tree and RUSBoosted ensemble methods demonstrated nearly perfect classification performance. The decision tree achieved 99.88% accuracy in validation and 99.44% in testing, while the RUSBoosted method achieved the highest accuracy with 99.68% accuracy in validation and 99.98% in testing. Artificial neural networks also demonstrated competitive performance with 96.98% accuracy in validation and 99.04% accuracy in testing. However, Naive Bayes (93.88% test accuracy) and SVM (96.11% test accuracy) were seen to lag behind. The confusion matrix and ROC curves for the best result are shown in Figure 2. Experimental findings reveal that ensemble methods outperform individual learners in both regression and classification steps. The boosting approach produced the lowest error values in E2E delay estimation, while the RUSBoosted method achieved the highest accuracy in route selection classification.

TABLE IIII: Obtained Results for Classification Step

Methods	Accuracy (%)(Validation)	Accuracy (%)(Test)
Tree	99.88	99.44
Ensemble (RUSBoosted)	99.68	99.98
Neural Network	96.98	99.04
Naive Bayes	95.86	93.88
SVM	94.51	96.11

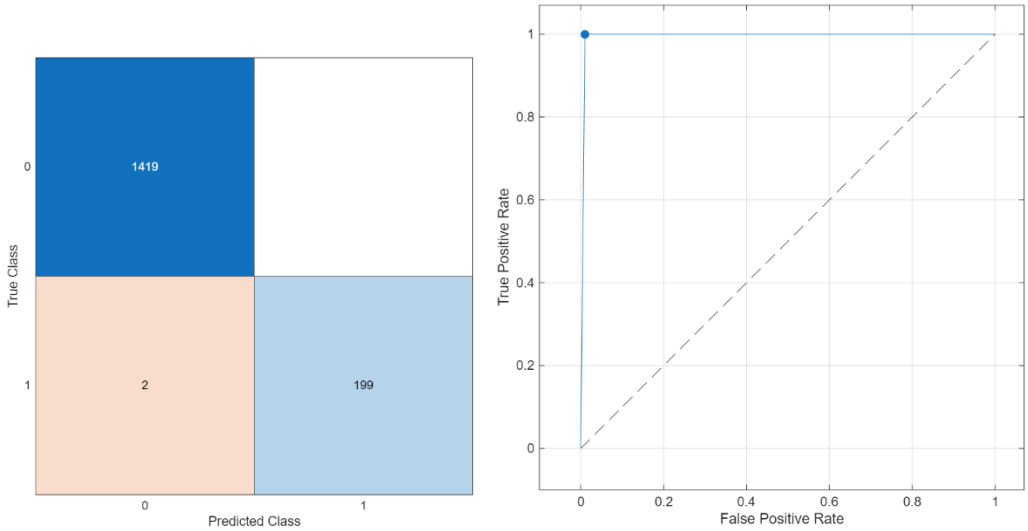

Fig. 2: Confusion matrix and ROC curve for obtained result

The results show that the proposed approach demonstrates high performance in both E2E delay estimation and efficient route selection. In the regression phase, while all methods produced results with similar error levels, ensemble methods were observed to provide lower MAE and RMSE values compared to the other models. This supports the ensemble approach's ability to produce more balanced and generalizable results by combining different model predictions. Test results are also consistent with validation results, demonstrating that the model is not prone to overfitting.

In the classification phase, accuracy rates are quite high. Decision tree-based models and ensemble methods achieved success rates above 99%. The RUSBoost algorithm's 99.98% accuracy in the testing phase demonstrates its high performance even with unbalanced data distributions. While Neural Network and Naive Bayes classifiers also exhibit satisfactory performance, the superiority of ensemble-based methods is clearly evident. These findings show that ensemble learning methods provide a reliable solution to the problem of delay estimation and route optimization.

4. Conclusion

In this study, a machine learning-based framework for E2E delay estimation and efficient route selection was developed using the 5G-VANET Real-Time Routing Dataset. Regression analyses reveal that the ensemble learning methods, including boosted and bagging, outperform other models; for example, the boosted ensemble method achieved RMSE = 22.78 on the validation set, and RMSE = 23.88 on the test set, providing reliable predictions with a low error rate.

Accuracy rates exceeding 99% in the classification step confirm that ensemble-based algorithms exhibit superior performance in optimal route selection, particularly despite unbalanced data distributions. The findings from this study demonstrate that ensemble learning approaches provide solutions for both delay estimation and route optimization in 5G-supported VANET-MEC environments. The results demonstrate the capability of this approach for developing real-time data processing and decision support mechanisms in 5G-supported VANET-MEC environments. By verifying the superiority of ensemble learning methods in delay estimation and optimal route selection, they can contribute to the design of scalable, reliable, and low-latency communication strategies in intelligent transportation systems. Future work may include modeling more complex traffic scenarios, applying different machine learning approaches, and testing the methods in real-time simulation environments.

5. References

[1] H. Bangui, M. Ge, and B. Buhnova, “A hybrid machine learning model for intrusion detection in VANET,” *Computing*, vol. 104, no. 3, pp. 503–531, Mar. 2022, doi: 10.1007/S00607-021-01001-0/FIGURES/14. https://doi.org/10.1007/s00607-021-01001-0

[2] K. Kandali, L. Bennis, O. El Bannay, and H. Bennis, “An Intelligent Machine Learning Based Routing Scheme for VANET,” *IEEE Access*, vol. 10, pp. 74318–74333, 2022, doi: 10.1109/ACCESS.2022.3190964. https://doi.org/10.1109/ACCESS.2022.3190964

[3] Y. Tang, N. Cheng, W. Wu, M. Wang, Y. Dai, and X. Shen, "Delay-Minimization Routing for Heterogeneous VANETs with Machine Learning Based Mobility Prediction," *IEEE Trans Veh Technol*, vol. 68, no. 4, pp. 3967–3979, Apr. 2019, doi: 10.1109/TVT.2019.2899627.
https://doi.org/10.1109/TVT.2019.2899627

[4] L. Hota, B. P. Nayak, and A. Kumar, "Machine learning algorithms for optimization and intelligence in wireless networks WSNs, MANETs, VANETs, and USNs," *5G and Beyond Wireless Communications: Fundamentals, Applications, and Challenges*, pp. 306–332, Sep. 2024, doi: 10.1201/9781032625034-16/MACHINE-LEARNING-ALGORITHMS-OPTIMIZATION-INTELLIGENCE-WIRELESS-NETWORKS-LOPAMUNDRA-HOTA-BIRAJA-PRASAD-NAYAK-ARUN-KUMAR.
https://doi.org/10.1201/9781032625034-16

[5] J. Grover, N. K. Prajapati, V. Laxmi, and M. S. Gaur, "Machine Learning Approach for Multiple Misbehavior Detection in VANET," *Communications in Computer and Information Science*, vol. 192 CCIS, no. PART 3, pp. 644–653, 2011, doi: 10.1007/978-3-642-22720-2_68.
https://doi.org/10.1007/978-3-642-22720-2_68

[6] A. Sharma and A. Jaekel, "Machine Learning Approach for Detecting Location Spoofing in VANET," *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, vol. 2021-July, Jul. 2021, doi: 10.1109/ICCCN52240.2021.9522170.
https://doi.org/10.1109/ICCCN52240.2021.9522170

[7] B. Karthiga, D. Durairaj, N. Nawaz, T. K. Venkatasamy, G. Ramasamy, and A. Hariharasudan, "Intelligent Intrusion Detection System for VANET Using Machine Learning and Deep Learning Approaches," *Wirel Commun Mob Comput*, vol. 2022, no. 1, p. 5069104, Jan. 2022, doi: 10.1155/2022/5069104.
https://doi.org/10.1155/2022/5069104

[8] S. Khatri *et al.*, "Machine learning models and techniques for VANET based traffic management: Implementation issues and challenges," *Peer Peer Netw Appl*, vol. 14, no. 3, pp. 1778–1805, May 2021, doi: 10.1007/S12083-020-00993-4/FIGURES/8.
https://doi.org/10.1007/s12083-020-00993-4

[9] "5G-VANET Real-Time Routing Dataset." Accessed: Sep. 08, 2025. [Online]. Available: https://www.kaggle.com/datasets/programmer3/5g-vanet-real-time-routing-dataset