BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 73(4), 2025, Article number: e154287
DOI: 10.24425/bpasts.2025.154287

Estimation of CO_2 emissions from vehicles using machine learning and multi-model investigation

Mükerrem Sinem MUNGAN[1] and **Orhan ARPA**[2]*

[1] Mardin Artuklu University Vocational School Machinery and Metal Technologies Department Machinery Program, Mardin, Türkiye
[2] Dicle University, Faculty of Engineering, Department of Mechanical Engineering, Diyarbakır, Türkiye

Abstract. This study presents a comprehensive analysis of the prediction of carbon dioxide emissions from vehicles using machine learning-based regression models. Linear regression, lasso regression, k-nearest neighbor regression, random forest, and CatBoostRegressor algorithms are systematically evaluated using a dataset of vehicle specifications and emissions data. Hyper-parameter optimization was performed using a grid search method and the performance of the models was measured using mean squared error, root mean squared error, mean absolute error, and R-squared metrics. CatBoostRegressor stood out for its high predictive accuracy, while random forest and k-nearest neighbor models also produced notable results, while linear models failed to model complex data relationships. Correlation analysis showed that engine displacement, number of cylinders, and fuel consumption were strongly correlated (0.92–0.99) with carbon dioxide emissions. The comparison with the literature showed that the study was characterized by its multi-model approach, rigorous data pre-processing, and systematic optimization. However, the geographical limitation of the dataset and the lack of dynamic variables such as driving conditions restrict its generalizability. In the future, explainable artificial intelligence methods and larger datasets may overcome these limitations. By highlighting the applicability of CatBoostRegressor, this study strengthens the contribution of machine learning to environmental sustainability policy and provides methodological innovation in the literature.

Keywords: CO_2 emissions; machine learning; CatBoostRegressor; regression analysis; environmental sustainability.

1. INTRODUCTION

Global climate change is one of the most critical environmental issues of our time, and the accumulation of carbon dioxide (CO_2) emissions in the atmosphere is one of the main causes of this problem [1]. The transport sector in particular is responsible for about a quarter of global greenhouse gas emissions due to the widespread use of fossil fuel vehicles [2]. This situation threatens human health through increased air pollution and causes irreversible damage to ecosystems [3]. Accurate estimation of CO_2 emissions is critical for achieving environmental sustainability goals, developing emission reduction strategies, and formulating policies [4]. Machine learning (ML) has the potential to make these estimation processes more efficient by extracting meaningful patterns from complex data sets [5]. In this context, the prediction of vehicle CO_2 emissions using ML-based models has emerged as a key area of research from both academic and practical perspectives [6].

This study presents a comprehensive regression analysis for the estimation of CO_2 emissions from vehicles. Using the dataset "CO_2 Emission by Vehicles" on the Kaggle platform, the performance of different machine learning algorithms is compared. In the study, linear regression, lasso regression, k-nearest neighbor (KNN) regression, random forest, and CatBoostRegressor algorithms are configured as base learners. To optimize the model performance, hyperparameters were adjusted using the grid search method and the accuracy of the models was evaluated using metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE) and R-squared (R^2). In the data pre-processing stage, steps such as elimination of missing data, coding of categorical data and scaling of numerical data were carefully applied. Correlation analysis was also carried out to understand the relationships between the variables in the dataset and it was found that engine displacement, number of cylinders and fuel consumption were highly correlated with CO_2 emissions. As a result of the study, the CatBoostRegressor model was found to have the best prediction performance compared to other algorithms.

- High performance model: CatBoostRegressor provided the best prediction accuracy with MSE = 3.8707, RMSE = 1.9674, and R^2 = 0.9956, and effectively modeled complex data relationships.
- Comprehensive comparison: A systematic comparison of five different regression algorithms was conducted to demonstrate the superiority of ensemble learning and boosting-based models over linear models.
- Optimization and data analysis: Hyperparameter optimization with grid search and detailed correlation analysis improved the reliability and generalizability of the model.

This study presents a machine learning-based multi-model approach for predicting vehicle CO_2 emissions. The main advantage of this research is that the best performance values are obtained by systematically comparing many different regression algorithms on the same dataset. Secondly, model performance

*e-mail: orhana@dicle.edu.tr

Manuscript submitted 2025-04-14, revised 2025-04-30, initially accepted for publication 2025-05-08, published in July 2025.

is maximized by grid search-based hyperparameter tuning and a repeatable methodology is proposed. Thirdly, the effects of variables such as engine displacement, number of cylinders, and fuel consumption on CO_2 emissions are investigated in detail using correlation analysis, providing valuable insights for both academic and industrial applications. In addition, the open-access Kaggle dataset used in the study enhances scientific transparency by allowing other researchers to replicate similar analyses [7]. As an innovative contribution, the success of CatBoostRegressor in predicting CO_2 emissions with high accuracy and low computational cost supports the wider application of this algorithm in environmental sustainability studies. However, the main limitations of the study include the geographical limitation of the dataset to a single region and the lack of additional variables representing dynamic driving conditions. These shortcomings may limit the generalizability of the resulting models to different geographical regions or real-time driving scenarios.

In the literature, studies on vehicle CO_2 emission estimation usually focus on a single algorithm or consider a limited number of variables. In contrast, this study systematically compares five different regression algorithms and discusses in detail the advantages, disadvantages, and application scenarios of each model. Furthermore, while most studies focus only on emission estimation, this research combines data pre-processing, correlation analysis, and hyper-parameter optimization in a holistic framework. In particular, the ability of CatBoostRegressor to work with categorical data and its low computational cost stand out when compared to other models in the literature. The study offers a different perspective to the literature by focusing not only on prediction accuracy but also on the practical applicability of the models and their potential impact on environmental policy.

2. LITERATURE REVIEW

This section reviews recent work in the literature on CO_2 emission estimation in vehicles. Zhou *et al.* demonstrated 40–45% emission reduction using a CatBoost algorithm in a study in China, 2018–2022 [8]. Andrade *et al.* explored the differences between ethanol and gasoline vehicles, employing unsupervised outlier detection, and digital twin simulations [9]. Sulekha Devi *et al.* introduced IoV-based real-time emission prediction, achieving 11–150 kg/h reduction with speed optimization [10] demonstrating the data efficiency and decision support potential of machine learning-based methods. Nesro *et al.* applied causal ML in the context of Industry 4.0 [11], while Tian *et al.* modeled coal emission with ML, MLR $R^2 = 0.99$ [12], and Wang *et al.* achieved $R^2 = 0.975$ accuracy with RF by adding hazardous driving behavior and road type [13], showing how algorithm diversity, as well as variable selection, affect the results.

In deep learning and explainable artificial intelligence (XAI) approaches, Alam *et al.* introduced carbonMLP, using Canadian data and $R^2 = 0.9938$ [14], while Mobasshir *et al.* applied a hybrid emission benchmarking with AHP-EDAS, highlighting the environmental superiority of hybrid vehicles [15]. Guo *et al.* predicted emissions at a pixel level with night light and terrain data reaching 83.76% accuracy [16] and expanding the diversity of the field. Other contributions include Alazemi-Alazmi, who used real driving data to model petrol and diesel vehicle emissions with bagging [17], and Al-Nefaie-Aldhyani, who used BiLSTM with Kaggle data achieving R^2 =0.9378 [18]. Gürcan conducted a comparison of 18 regression and deep learning algorithms [19], which is important for the comparative evaluation of models. Udoh-Lu utilized UK VCA WLTP data with a decision tree model MAE = 2.20) [20]. Mądziel developed micromodels for LPG and hybrid vehicles using PEMS-OBDII data, applying GPR and GBM [21,22]. Liu *et al.* integrated the Boruta feature into ML with R^2 increase compared to MOVES [23] and Zhang *et al.* proposed the LSTM-DL-DTCEM framework with $R^2 \approx 0.99$ [24], balancing high accuracy with low computational cost, and real-time monitoring. Finally, Natarajan *et al.* utilized Canadian data, emphasizing CatBoost memory efficiency [25]. Li *et al.* explored LSTM-based deep learning, incorporating VSP and slope effect [26], and Moon *et al.* conducted a Euro-7 RDE test, using XGBoost pre-OBM monitoring [27] and highlighting the importance of both model generalizability and environmental policy applications. In this paper, we present the methodological diversity, data sources, and performance results used in emission estimation together, highlighting gaps in the field and areas for future innovation.

3. MATERIAL AND METHOD

In this study, a regression analysis was performed for the prediction of CO_2 emissions from vehicles. Modeling was performed using different machine learning algorithms and the performance of these models were compared. To optimize the model performances, hyperparameter adjustments were made with the grid search method. The open access Kaggle "CO_2 Emission by Vehicles" dataset was used in the study. The dataset was divided into 80% training and 20% testing, and R^2, RMSE, and MAE were used as evaluation criteria.

3.1. Dataset

The "CO_2 Emissions by Vehicles" dataset used in the study was compiled from the Government of Canada's open data portal and published on the Kaggle platform [28]. The dataset contains technical specifications and associated CO_2 emission values for 7385 different vehicles registered in Canada between 2014 and 2020. A total of 12 variables, including engine displacement (cm^3), number of cylinders, fuel type (petrol, diesel, hybrid, electric), horsepower (hp), vehicle weight (kg), transmission type, fuel consumption in liters/100 km and CO_2 emissions (g/km), allow for multidimensional analyses of both numerical and categorical nature. The main variables in the dataset and their descriptions are presented in Table 1 below.

In the data pre-processing stage, characteristics with missing values below 5% were completed using the averaging method; outliers identified using the z-score and interquartile range methods were limited by outlier trimming. The categorical variable 'fuel type' was converted to numerical form using one-hot encoding, and all numerical features were normalized to the interval [0, 1] using a min-max scaler to ensure consistent and comparable performance of the models.

Table 1
Descriptions of the parameters in the dataset

Parameter	Description
Make	Refers to the manufacturer or brand of the vehicle.
Model	Indicates the model name of the vehicle.
Year	The production year of the vehicle used in time-based change analyses.
Engine size	Engine displacement, expressed in liters, and reflecting the relationship between engine power and fuel consumption.
Cylinders	Number of engine cylinders, directly related to engine performance and power output.
Transmission	Indicates the type of transmission (manual, automatic, etc.).
Fuel type	Specifies the type of fuel used in the vehicle, represented by abbreviations: X = Regular gasoline, Z = Premium gasoline, D = Diesel, E = Ethanol (E85), N = Natural gas
Fuel consumption City	Fuel consumption in urban driving conditions; measured in liters/100 km.
Fuel consumption Hwy	Fuel consumption on highways (non-urban); measured in liters/100 km.
Fuel consumption Comb	Combined fuel consumption, calculated as the average of city and highway fuel use and measured in liters/100 km.
CO_2 emissions	The amount of carbon dioxide emitted by the vehicle, expressed in grams/km.

3.2. Machine learning

Machine learning (ML) is a subfield of artificial intelligence that enables computer systems to perform specific tasks by learning from data without being explicitly programmed. This method allows systems to discover patterns and relationships from historical data, make predictions using this information, and improve their performance over time. Unlike traditional programming, machine learning algorithms are based on data-driven learning processes rather than specific rules [29].

3.2.1. Linear regression

Linear regression is a basic statistical method that models the linear relationship between a dependent variable and one or more independent variables [30]. This model is used to understand the effect of changes in independent variables on the dependent variable and to predict future values. Linear regression is a widely preferred technique, especially in data analysis and forecasting processes. In this model, regression coefficients are usually estimated using ordinary least squares (OLS). OLS aims to minimize the sum of squares of the differences between the predicted values and the actual values. In this way, the prediction accuracy of the model is improved and the effects of independent variables on the dependent variable are more accurately determined. Due to its simple structure and interpretability, linear regression is used as an effective tool in many fields, especially in environmental data analysis and emission estimation. The mathematical expression of linear regression is given in (1)

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

3.2.2. Least absolute shrinkage and selection operator (lasso) regression

Lasso regression is a technique used in regression analysis to reduce variable selection and model complexity [31]. This method reduces some coefficients to zero by applying a penalty to the sum of the absolute values of the regression coefficients. In this way, only important variables are retained in the model and unnecessary ones are excluded. This is particularly useful when there are a large number of variables, and some variables have limited influence on the target variable. Lasso regression reduces the risk of model overfitting, increases interpretability, and improves prediction accuracy. Therefore, it is a preferred method in high-dimensional data sets and in analyses where variable selection is important. The mathematical expression of lasso regression is given in (2)

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,. \tag{2}$$

3.2.3. K-nearest neighbors (KNN) regression

KNN regression is a simple and effective supervised learning method used for both classification and regression problems in machine learning [32]. The basic principle of this algorithm is to predict the class or value of a data point by looking at the class or value of its k-nearest neighbors in the training data set. KNN is based on the principle of "like is close to like" and given a new data point, it calculates the distances between that point and all data points in the training set. After identifying the k-neighbors with the smallest distance, it assigns classes based on the majority of these neighbors or performs regression prediction using their average values. The KNN algorithm is an easy-to-implement and intuitive method that can provide effective results, especially for small and medium-sized data sets. However, with large datasets, the computational cost can be high, and the choice of distance metrics can affect performance. The mathematical expression of KNN regression is given in (3)

$$\hat{y}(x^*) = \frac{1}{k} \sum_{i \in N_k(x^*)} y_i \,. \tag{3}$$

3.2.4. Random forest

Random forest is a powerful ensemble learning algorithm used to solve both classification and regression problems. This method works with a combination of multiple decision trees and trains each tree with different subsets of data and randomly selected features. This diversity increases the generalization ability of the model and reduces the risk of overfitting. Random forest

is widely preferred in data science and machine learning applications due to its high accuracy, flexibility, and interpretability [33]. The mathematical expression of random forest is given in (4)

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} h_b(x). \tag{4}$$

3.2.5. CatBoostRegressor

CatBoostRegressor is a gradient-boosting algorithm that shows high performance, especially when working with categorical data [34]. Unlike other boosting methods, CatBoost can process categorical variables directly, which significantly simplifies the data preprocessing process and improves the accuracy of the model. CatBoostRegressor aims to minimize the model errors by building each decision tree in turn. This process increases the generalization ability of the model, reducing the risk of overfitting. It also offers fast training times on large datasets thanks to GPU support. It provides high-accuracy predictions by using categorical and numerical variables together. The mathematical expression of CatBoostRegressor is given in (5)

$$L = \sum_{i=1}^{n} \left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i) \right) + \Omega(f_t). \tag{5}$$

Table 2 compares the advantages and disadvantages of the machine learning methods used in our study.

Table 2
Comparison of advantages and disadvantages of the machine learning methods used

Model	Advantages	Disadvantages
Linear regression	Simple and fast; highly interpretable.	Can only model linear relationships; may suffer from multicollinearity issues.
Lasso regression	Reduces model complexity by performing variable selection; prevents overfitting.	Model accuracy is sensitive to the chosen penalty parameter.
KNN regression	Non-parametric; can model complex relationships.	High computational cost; may perform slowly on large datasets.
Random forest	Provides high accuracy; reduces overfitting; it can handle many variables.	Low interpretability: training time can be long.
CatBoost Regressor	Works effectively with categorical data; offers fast training times.	High model complexity; requires hyperparameter tuning.

3.3. Grid search optimization

Grid search optimization is a common method for the systematic examination of hyperparameter settings in machine learning models [35]. This technique aims to evaluate the model performance of each combination using a finite grid of user-specified hyperparameter values. Thus, the set of hyperparameters that will maximize the overall performance of the model is determined.

In this method, a certain range of hyperparameters is defined and possible values for each hyperparameter are systematically tested. The results are supported by techniques such as cross-validation and evaluated on the overall performance of the model. Grid search is a crucial tool for determining optimal settings to reduce model complexity and the risk of overfitting. However, when the number of hyperparameters is large, the computational cost increases, and the implementation time increases. Therefore, especially in high-dimensional hyperparameter spaces, researchers also consider alternative optimization methods to improve time efficiency. In general, Grid search optimization is a widely preferred method for model selection and tuning due to its robustness and systematic nature.

The grid search method guarantees the best performance set by systematically evaluating all combinations within a narrow and discrete range of hyperparameters. This ensures reproducibility and transparency of the results while keeping the computational cost reasonable. Random search, which is based on random sampling, may not cover the entire space, while Bayesian optimization requires additional pre-modeling and complex updating steps, making the implementation process less straightforward. For these reasons, grid search was preferred in our study.

3.4. Data preprocessing

The data preparation phase is one of the cornerstones of machine learning modeling and involves cleaning, organizing, and transforming raw data. This process is necessary for the model to produce reliable and accurate results. Missing values, outliers, and noise in the raw data are first detected and corrected with appropriate techniques to obtain accurate results. There are many examples in the literature that the data cleaning step significantly improves model performance.

Data preparation involves scaling and normalization of numerical data and coding of categorical data using appropriate methods. In particular, standardization and min-max scaling techniques accelerate the model learning process and minimize the margin of error. Furthermore, the complexity of the model is reduced by appropriately selecting the features in the data set and deriving new variables. At this stage, meticulous data preparation directly affects the overall performance of the final model, thus supporting the academic and practical validity of the results.

3.5. Evaluation metrics

To objectively evaluate the model performance, various statistical metrics were used to measure the agreement between our prediction results and actual values [36]. In this section, the main metrics used are the mean squared error (MSE), explained variance ratio (R-Squared-R^2), root mean squared error (RMSE), and mean absolute error (MAE). These metrics reveal the overall performance of the model by evaluating the magnitude of the prediction errors, the percentage of variance explained, and the

absolute deviations. The formulas of the metrics used are given in equations (6)–(9) [37]

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

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{8}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{9}$$

These equations provide comparability and reliability in the evaluation process by measuring the prediction errors of models from different perspectives [38,39]. The use of these metrics allows an objective measurement of model accuracy and generalizability and supports the choice between different models.

3.6. Proposed model

In this study, a multiple regression-based machine learning model is proposed for the prediction of carbon dioxide (CO_2) emissions from vehicles. The proposed model is based on an architecture that includes different regression algorithms as shown in Fig. 1. The main objective of the model is to accurately estimate CO_2 emissions using technical data of vehicles.

The modeling process starts with the separation of the data set into training and test data. After this stage, necessary preprocessing steps were applied to the data. The preprocessing process includes the removal of missing data, transformation of categorical data, and scaling.

The preprocessed data were fed separately to five different regression algorithms: linear regression, lasso regression, KNN regression, random forest, and CatBoostRegressor. Each of these algorithms is configured as a base learner. The obtained prediction results were subjected to an optimization process to minimize the overall accuracy and error rates of the model. In this stage, hyperparameter tuning was performed using the grid search method, and the optimal parameters for each algorithm were determined. Finally, the optimized prediction values were obtained as the final output. Through this multi-model approach, the overall forecasting performance is improved by leveraging the strengths of each algorithm.

4. RESULTS AND DISCUSSION

In this section, the results of the regression analysis conducted using technical data of vehicles are evaluated. During the modeling process, the dataset was split into 80% training and 20% testing data. The training data was used for the learning process of the models, while the test data was used to assess the generalization capabilities of the models. Model performances were compared using various metrics, and the algorithms yielding the best results were identified.

4.1. Results

To understand the relationships among the variables in the dataset, a correlation analysis was conducted. The correlation matrix in Fig. 2 provides a detailed visualization of the linear relationships between variables in the "CO_2 Emissions by Vehi-

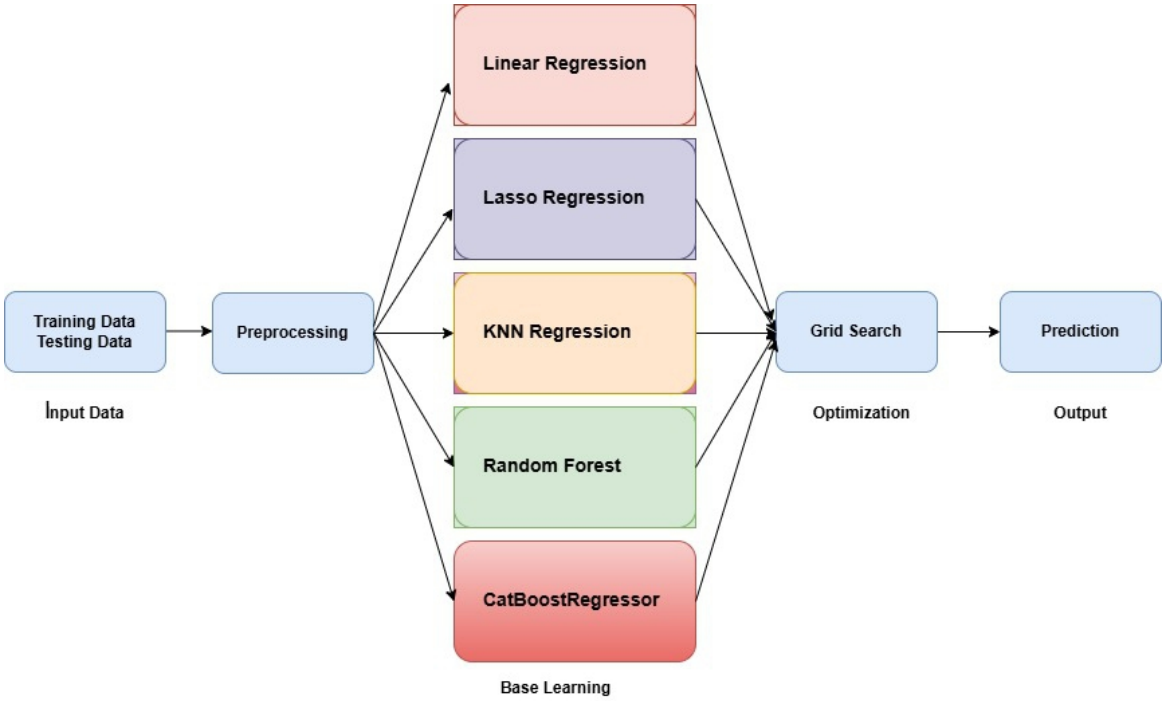

Fig. 1. Proposed multi-model regression architecture

Fig. 2. Correlation matrix

cles" dataset. This dataset contains technical specifications and emissions data for 7385 vehicles registered in Canada between 2014 and 2020. The matrix enhances clarity by visualizing positive correlations in light blue tones and negative correlations in green tones.

The analysis reveals that there is a strong positive correlation between CO_2 emissions and critical technical characteristics such as engine displacement, number of cylinders, and fuel consumption (urban and combined), ranging from 0.92 to 0.99. This significant relationship highlights the significant impact of these variables on emission levels and positions them as key predictors in the regression models used in the study. In addition, a strong negative correlation of −0.93 is observed between fuel efficiency and CO_2 emissions. This finding highlights the role of fuel economy in emission reduction strategies by showing that vehicles with higher fuel efficiency produce lower CO_2 emissions.

Figure 3 presents the scatter plot illustrating the relationship between engine size and CO_2 emissions. Upon examining the

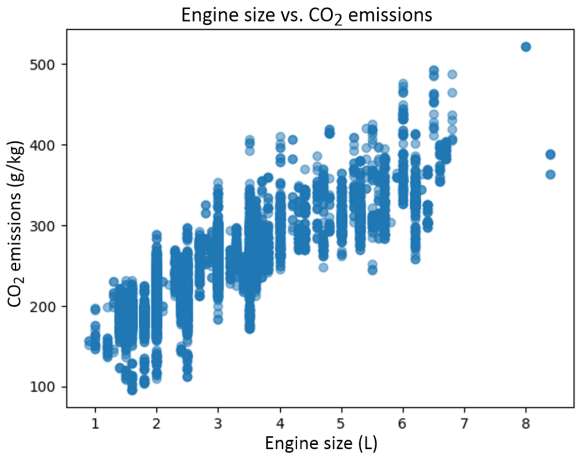

Fig. 3. Scatter plot showing the relationship between engine size and CO_2 emissions

plot, it is observed that CO_2 emission values generally increase as engine size increases. This indicates that engine size is a determining factor in CO_2 emissions and should be considered as a significant input variable in the modeling process. The clear linear trend in the data suggests that regression models can effectively learn this relationship.

The hyperparameters selected for grid search optimization, which was used to identify the most suitable hyperparameters for each model, are presented in Table 3. This table details the systematic hyperparameter tuning process conducted to enhance the performance of each model. The hyperparameter settings were carefully selected to improve the generalization ability of the models and to minimize the risk of overfitting.

Table 3
Model hyperparameter settings

Model	Hyperparameters	Values/ranges
Linear regression	–	Default parameters
Lasso regression	alpha	[0.0001, 0.001, 0.01, 0.1, 1]
KNN	n_neighbors	[1, 5, 10, 20, 50]
Random forest	n_estimators,	[100, 200, 300, 500]
	max_depth	[10, 20, 30, None]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
CatBoostRegressor	Iterations	[100, 200, 300, 500]
	learning_rate	[0.01, 0.1, 0.2, 0.5]
	depth	[2, 4, 6, 8]
	l2_leaf_reg	[1, 3, 5, 7, 9]

The best hyperparameter values obtained through the grid search optimization process are presented in Table 4. This table reflects the results of the systematic search conducted to achieve

Table 4
Hyperparameters determined through optimization

Model	Hyperparameters	Best value
Linear regression	–	Default parameters
Lasso regression	alpha	0.01
KNN	n_neighbors	10
Random forest	n_estimators	300
	max_depth	20
	min_samples_split	2
	min_samples_leaf	1
CatBoostRegressor	iterations	300
	learning_rate	0.1
	depth	6
	l2_leaf_reg	3

Table 5
Comparison of model performances

Model	Mean squared error (MSE)	R-squared (R^2)	Root mean squared error (RMSE)	Mean absolute error (MAE)	Mean diff
Linear regression	5.6604	0.990581	2.379160	3.447754	62.1812
Lasso regression	5.6598	0.990584	2.379033	3.449433	62.1634
KNN	4.5374	0.993948	2.130117	2.871130	58.555
Random forest	3.8818	0.995570	1.970228	2.378100	60.66415
CatBoostRegressor	3.8707	0.995596	1.967409	2.454250	57.1218

optimal model performance. Through a comprehensive search using the grid search optimization method, the hyperparameter combinations that maximize the overall performance of each model were identified. The selected optimal settings play a significant role in enhancing prediction accuracy and preventing overfitting in the models.

According to the comparison in Table 5, linear regression, and lasso regression models, as baseline approaches, performed poorly compared to other models with high error values (MSE, RMSE, MAE) and low R^2 scores. This indicates that these linear models fail to adequately capture the complex and possible non-linear relationships between the independent variables and CO_2 emissions.

On the other hand, KNN, random forest, and CatBoostRegressor models represent the structure of the dataset more successfully thanks to their lower mean square errors and higher explained variance ratios. In particular, CatBoostRegressor provided the best results with 3.8707 MSE, 1.9674 RMSE, and 0.9956 R^2. This model is characterized by its automatic processing of categorical and numerical data, relatively short training time, and generally consistent high accuracy. Random forest achieves a similar level of success, but in some cases, it can show an advantage in error distribution by producing lower error values in the MAE metric.

Despite the success of CatBoostRegressor, its sensitivity to hyperparameter settings and limited interpretability of in-model decision processes are issues that need to be considered in practice. These findings suggest that ensemble and boosting-based methods provide more flexible and reliable results than linear models in vehicle CO_2 emission estimation. Depending on the application context, both performance and usability criteria should be considered together when choosing a model.

The performance comparison results of different regression models are visualized in Fig. 4. The chart presents a visual evaluation of each model based on error metrics (MSE, RMSE, MAE) and the accuracy level (R^2). According to the findings, the CatBoostRegressor model yielded the lowest error rates and the highest R^2 value, demonstrating a strong fit with the dataset. Specifically, it achieved 3.87 MSE, 1.96 RMSE, and 0.9956 R^2. This model is followed by random forest and KNN, respectively. On the other hand, linear regression and lasso regression exhibited higher error values and proved inadequate in modeling complex non-linear relationships. These results highlight the limitations of linear regression approaches and demonstrate

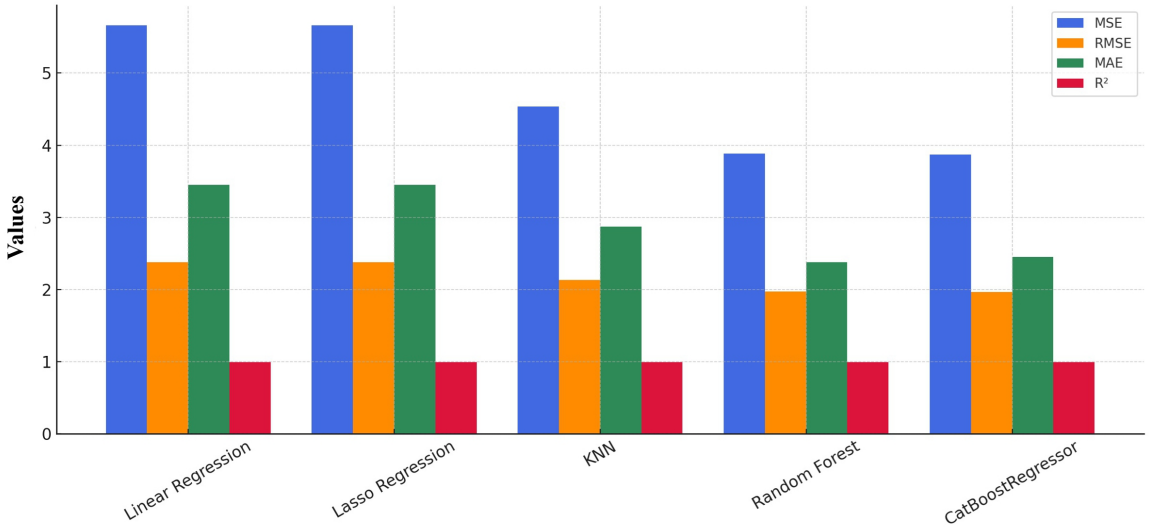

Fig. 4. Model performance graph

that ensemble and non-parametric methods are more effective, particularly in high-dimensional and complex datasets.

Figure 5 depicts the relationship between the predicted CO_2 emission values and the actual values using the CatBoostRegressor model. It is observed that the predictions of the model closely align with the actual values, which supports its high accuracy performance. Especially for low and medium emission levels, the predicted values almost perfectly overlap with the actual data. Although minor deviations are observed at higher emission levels, the overall trend confirms the model success. These results suggest that CatBoostRegressor is a reliable model for predicting vehicle CO_2 emissions and can capture complex data relationships effectively. This performance is reinforced by hyperparameter optimization and the model ability to handle categorical variables.

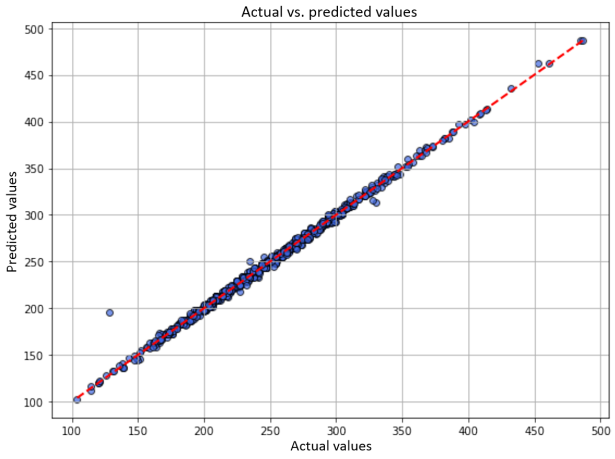

Fig. 5. Comparison of CatBoostRegressor model predictions with actual CO_2 emission values

In Fig. 6, the ROC (receiver operating characteristic) curve plotted to evaluate the performance of the CatBoostRegressor model is presented. The curve position above the diagonal line indicates the model ability to minimize the false positive rate while maximizing the true positive rate. This result confirms that

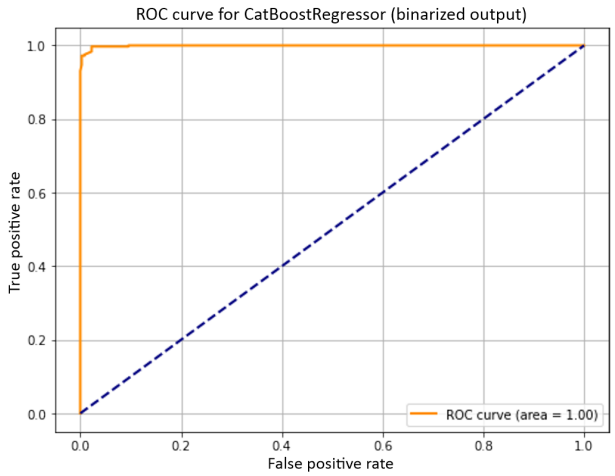

Fig. 6. ROC curve of the CatBoostRegressor model

the CatBoostRegressor is a highly reliable model for predicting vehicle CO_2 emissions and effectively captures the complex relationships within the dataset.

Table 6 presents a comparative overview of the actual CO_2 emission values of vehicles and the predictions made by five different regression models. Upon examining the data, it is evident that the performance of the models varies across different data points. For example, CatBoostRegressor has shown superior performance in modeling complex relationships compared to the other models, providing predictions that are closer to the actual values in some cases. In contrast, linear regression and lasso regression models exhibited higher deviations, highlighting the limitations of their linear approaches. KNN and random forest also performed better in specific scenarios. While CatBoostRegressor stood out for its exceptional prediction ability, it was observed that linear models struggled even with simpler data structures.

Table 6

Actual CO_2 emissions of vehicles and model prediction results

REAL_ VALUE	PRED_ LINEAR	PRED_ LASSO	PRED_ KNN	PRED_ RF	PRED_ CGB
370.0	483.52	483.47	471.5	478.51	448.90
242.0	294.78	294.77	292.0	294.25	266.66
284.0	414.68	414.66	354.0	366.24	387.78
240.0	282.41	282.39	283.0	286.52	255.08
220.0	262.62	262.61	264.5	263.25	233.86
206.0	238.16	238.15	240.0	239.98	219.96
212.0	255.51	255.51	255.0	254.88	249.60
234.0	279.97	279.96	281.5	282.92	262.12
185.0	208.43	208.42	214.5	209.49	203.86
256.0	314.19	314.17	308.0	316.04	298.02

4.2. Discussion

This study focuses on the prediction of vehicle-related CO2 emissions using machine learning-based regression models, systematically comparing the performance of different algorithms. The "CO_2 Emission by Vehicles" dataset from the Kaggle platform was used to evaluate linear regression, lasso regression, KNN regression, random forest, and CatBoostRegressor models. Grid search optimization was applied for the model hyperparameters, and their performances were measured using MSE, RMSE, and R^2 metrics. The results showed that the CatBoostRegressor model provided high prediction accuracy with MSE = 3.8707, RMSE = 1.9674, and R^2 = 0.9956. The random forest and KNN models also demonstrated acceptable accuracy, while the linear models (linear regression and lasso regression) were inadequate in modeling complex data relationships. These findings suggest that ensemble learning and boosting-based methods are more effective, especially for heterogeneous and high-dimensional datasets.

A comparison of this study with existing literature offers valuable insights in terms of methodological approaches and performance metrics. Table 7 below compares the results of this study with eight prominent studies in the literature, focusing on MSE, R^2, and RMSE metrics. This comparison highlights the position and contributions of our study within the literature.

Table 7
Performance comparison of this study with existing studies in the literature

Study	Model	MSE	R^2	RMSE
This study	CatBoost Regressor	3.8707	0.9956	1.9674
Tian *et al.* [5]	Multiple linear regression	3.2000	0.9900	1.7889
Alam *et al.* [7]	CarbonMLP	0.0002	0.9938	0.0141
Wang *et al.* [6]	Random forest	–	0.9750	13.2930
Al-Nefaie and Aldhyani [11]	BiLSTM	5.0000	0.9378	2.2361
Udoh and Lu [13]	Decision tree	4.8400	–	2.2000
Mądziel [22]	Gradient boosting	0.7700	0.6100	0.8775
Zhang *et al.* [17]	LSTM (DL-DTCEM)	0.0278	0.9860	0.1650
Li *et al.* [19]	LSTM	–	0.9860	0.1650

The CatBoostRegressor model in this study demonstrates a balanced performance when compared to many studies in the literature (MSE = 3.8707, RMSE = 1.9674, R^2 = 0.9956). For example, Tian *et al.* [5] report a slightly lower error rate using a multiple linear regression model (MSE = 3.2000, R^2 = 0.9900, RMSE = 1.7889), but it is limited in handling complex datasets as it can only model linear relationships. In contrast, CatBoostRegressor effectively handles non-linear relationships and categorical data, offering broader applicability.

Alam *et al.* [7] with their CarbonMLP model (MSE = 0.0002, R^2 = 0.9938, RMSE = 0.0141) achieved extraordinarily low error rates, but the high computational complexity of this deep learning-based model and data preprocessing requirements may pose practical limitations. CatBoostRegressor, with its lower computational cost and faster training times, overcomes such constraints, making it a more suitable alternative for industrial applications.

Wang *et al.* [6] present the random forest model (R^2 = 0.9750, RMSE = 13.2930), which provides a high R^2 value but with an unusually high RMSE, indicating that the model lags behind others in terms of prediction accuracy. This suggests that even with additional variables like driving behaviors, the generalization capacity of the model may be limited. In this study, however, a robust foundation was established by analyzing the high correlation (0.92–0.99) between core variables such as engine size, cylinder count, and fuel consumption.

The BiLSTM model by Al-Nefaie and Aldhyani [11] (MSE = 5.0000, R^2 = 0.9378, RMSE = 2.2361) shows a lower R^2 value, making it less successful than the CatBoostRegressor in capturing complex data relationships. Similarly, Udoh and Lu [13] with their decision tree model (MSE = 4.8400, RMSE = 2.2000) provide a similar error rate to this study, but the lack of R^2 value makes it difficult to evaluate the generalization capability of the model.

Mądziel [22] reports a gradient boosting model (MSE = 0.7700, R^2 = 0.6100, RMSE = 0.8775), which draws attention due to its low R^2 value, reflecting the limitations of micro-scale modeling specific to LPG vehicles. In contrast, the CatBoostRegressor model in this study, working with a larger dataset, provides higher explanatory power.

Zhang *et al.* [17] and Li *et al.* [19] with their LSTM-based models (R^2 = 0.9860, RMSE = 0.1650) show high accuracy with extremely low RMSE values, but these models are known to have risks of overfitting on small datasets and carry high computational costs. The MSE of Zhang *et al.* [17] is reported as 0.0278, supporting the high accuracy of the model, while the lack of MSE for Li *et al.* [19] limits the comparison. CatBoostRegressor in this study demonstrates a balanced performance in terms of both high accuracy and practical applicability.

This study differs from other works in the literature in several key aspects:

- **Ensemble approach**: While studies like Alam *et al.* [14] or Zhang *et al.* [24] focus on a single algorithm, this study compares five different models, evaluating the advantages and disadvantages of each holistically.
- **Hyperparameter optimization**: Systematic optimization via grid search, which is absent in studies like Tian *et al.* [12] or Mądziel [22], has significantly enhanced model performance in this study.
- **Depth of data analysis**: Through correlation analysis, the relationships between engine size, cylinder count, and fuel consumption with CO_2 emissions have been thoroughly explored, providing a broader applicability compared to more specific approaches like Wang *et al.* [13].
- **Practical applicability**: While Alam *et al.* [14] offer deep learning models with high accuracy, these models come with high computational costs. In this study, CatBoost provides high accuracy with lower computational costs, making it a more suitable option for industrial applications.

The main limitations of this study are that the dataset focuses on a single geographical region and does not include dynamic variables such as driving conditions. For example, Wang *et al.* [6] modeled driving behaviors, while Zhang *et al.* [24] included road gradients. In the future, adding such variables could enhance the generalization ability of the model. Additionally, the interpretability limitations of complex models like CatBoost can be addressed using XAI methods, as done by Alam *et al.* [14]. The use of broader datasets from different regions and vehicle types would strengthen the global applicability of the study.

This study highlights the superior performance of CatBoostRegressor in CO_2 emission prediction, showcasing the contribution of machine learning to environmental sustainability policies. A comparison with the literature reinforces the methodological rigor and practical value of the study, while also offering new avenues for future research.

5. CONCLUSIONS

This study presents a systematic review of machine learning-based regression models for predicting vehicle-induced CO_2 emissions. Using the "CO_2 Emission by Vehicles" dataset from the Kaggle platform, linear regression, lasso regression, KNN regression, random forest, and CatBoostRegressor algorithms were compared. Hyperparameter optimization was carried out using grid search, and model performance was evaluated using metrics such as mean squared error (MSE = 3.8707), root mean squared error (RMSE = 1.9674), mean absolute error (MAE = 2.4543), and R-squared (R^2 = 0.9956). CatBoostRegressor emerged as the top performer with superior prediction accuracy, while random forest and KNN showed notable results. In contrast, linear models were limited in modeling complex data relationships. Correlation analysis confirmed a high correlation (0.92–0.99) between engine size, cylinder count, and fuel consumption with CO_2 emissions, highlighting the critical role of these variables in prediction models. A comparison with the literature revealed that a multi-model approach in the study, thorough data preprocessing, and systematic optimization set it apart. However, the geographical limitation of the dataset and the absence of dynamic variables such as driving conditions constrain the model's generalization capacity. Future research can overcome these limitations by integrating XAI methods, larger datasets, and dynamic variables. This study strengthens the role of machine learning in environmental sustainability policies by emphasizing the high accuracy and applicability of CatBoostRegressor and introduces methodological innovation into the literature. It is planned to prototype the developed model in the form of a web service or mobile application and develop real-time CO_2 prediction software.

REFERENCES

[1] E. Aslan, "Prediction and Comparative Analysis of Emissions from Gas Turbines Using Random Search Optimization and Different MachineL earning Based Algorithms," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 72, p. 151956, Sep. 2024, doi: 10.24425/BPASTS.2024.151956.

[2] C. Magazzino, A. Costantiello, L. Laureti, A. Leogrande, and T. Gattone, "Greenhouse gas emissions and road infrastructure in Europe: A machine learning analysis," *Transp. Res. D Transp. Environ.*, vol. 139, p. 104602, Feb. 2025, doi: 10.1016/J.TRD.2025.104602.

[3] F. Alpsalaz, "Fault Detection in Power Transmission Lines: Comparison of Chirp-Z Algorithm and Machine Learning Based Prediction Models," *Eksploat. Niezawodn.–Mainten. Reliab.*, 2025, doi: 10.17531/ein/203949.

[4] G. Çinarer, M.K. Yeşilyurt, Ü. Agbulut, Z. Yilbaşi, and K.I. Kiliç, "Application of various machine learning algorithms in view of predicting the CO_2 emissions in the transportation sector," *Sci. Technol. Energ. Transit.*, vol. 79, p. 15, 2024, doi: 10.2516/STET/2024014.

[5] E. Aslan, "Araçlarda CO_2 Emisyonlarının FarklıYapay Sinir AğıModelleri Kullanılarak Tahminlerinin Karşılaştırılması," *Çukurova Üniversitesi Mühendislik Fakültesi Dergisi*, vol. 39, no. 2, pp. 309–324, doi: 10.21605/cukurovaumfd.1513998.

[6] N. Niroomand and C. Bach, "Integrating Machine Learning for Predicting Internal Combustion Engine Performance and Segment-Based CO_2 Emissions Across Urban and Rural Settings," *IEEE Access*, vol. 12, pp. 66223–66236, 2024, doi: 10.1109/ACCESS.2024.3399025.

[7] E. Aslan, "Temperature Prediction and Performance Comparison of Permanent Magnet Synchronous Motors Using Different Machine Learning Techniques for Early Failure Detection," *Eksploat. Niezawodn.–Mainten. Reliab.*, vol. 27, no. 1, pp. 1–16, Aug. 2025, doi: 10.17531/EIN/192164.

[8] G. Zhou, L. Mao, T. Bao, and F. Zhuang, "Machine learning-driven CO2 emission forecasting for light-duty vehicles in China," *Transp. Res. D Transp. Environ.*, vol. 137, p. 104502, Dec. 2024, doi: 10.1016/J.TRD.2024.104502.

[9] M. Andrade *et al.*, "On the Use of Biofuels for Cleaner Cities: Assessing Vehicular Pollution through Digital Twins and Machine Learning Algorithms," *Sustainability*, vol. 16, no. 2, p. 708, Jan. 2024, doi: 10.3390/SU16020708.

[10] A. Sulekha Devi *et al.*, "Internet-of-Vehicles Network for CO_2 Emission Estimation and Reinforcement Learning-Based Emission Reduction," *IEEE Access*, vol. 12, pp. 110681–110690, 2024, doi: 10.1109/ACCESS.2024.3441949.

[11] V.M. Nesro, T. Fekete, and H. Wicaksono, "Leveraging Causal Machine Learning for Sustainable Automotive Industry: Analyzing Factors Influencing CO2 Emissions," *Procedia CIRP*, vol. 130, pp. 161–166, Jan. 2024, doi: 10.1016/J.PROCIR.2024.10.071.

[12] L. Tian *et al.*, "Predicting Energy-Based CO_2 Emissions in the United States Using Machine Learning: A Path Toward Mitigating Climate Change," *Sustainability*, vol. 17, no. 7, p. 2843, Mar. 2025, doi: 10.3390/SU17072843.

[13] Z. Wang, M. Mae, S. Nishimura, and R. Matsuhashi, "Vehicular Fuel Consumption and CO_2 Emission Estimation Model Integrating Novel Driving Behavior Data Using Machine Learning," *Energies*, vol. 17, no. 6, p. 1410, Mar. 2024, doi: 10.3390/EN17061410.

[14] G.M.I. Alam *et al.*, "Deep learning model based prediction of vehicle CO_2 emissions with eXplainable AI integration for sustainable environment," *Sci. Rep.*, vol. 15, no. 1, pp. 1–28, Jan. 2025, doi: 10.1038/s41598-025-87233-y.

[15] M. Mobasshir *et al.*, "Analyzing vehicle emissions using a hybrid machine learning approach using weighted average based k-means clustering for sustainable transportation decision-making," *Green Technol. Sustain.*, vol. 3, no. 3, p. 100163, Jul. 2025, doi: 10.1016/J.GRETS.2024.100163.

[16] W. Guo, Y. Li, X. Cui, X. Zhao, Y. Teng, and A. Rienow, "Mapping Spatiotemporal Dynamic Changes in Urban CO_2 Emissions in China by Using the Machine Learning Method and Geospatial Big Data," *Remote Sens.*, vol. 17, no. 4, p. 611, Feb. 2025, doi: 10.3390/RS17040611.

[17] F. Alazemi, A. Alazmi, M. Alrumaidhi, and N. Molden, "Predicting Fuel Consumption and Emissions Using GPS-Based Machine Learning Models for Gasoline and Diesel Vehicles," *Sustainability*, vol. 17, no. 6, p. 2395, Mar. 2025, doi: 10.3390/SU17062395.

[18] A.H. Al-Nefaie and T.H. H. Aldhyani, "Predicting CO_2 Emissions from Traffic Vehicles for Sustainable and Smart Environment Using a Deep Learning Model," *Sustainability*, vol. 15, no. 9, p. 7615, May 2023, doi: 10.3390/SU15097615.

[19] F. Gurcan, "Forecasting CO_2 emissions of fuel vehicles for an ecological world using ensemble learning, machine learning, and deep learning models," *Peer J. Comput. Sci.*, vol. 10, p. e2234, Aug. 2024, doi: 10.7717/PEERJ-CS.2234/SUPP-2.

[20] J. Udoh, J. Lu, and Q. Xu, "Application of Machine Learning to Predict CO_2 Emissions in Light-Duty Vehicles," *Sensors*, vol. 24, no. 24, p. 8219, Dec. 2024, doi: 10.3390/S24248219.

[21] M. Mądziel, "Liquified Petroleum Gas-Fuelled Vehicle CO_2 Emission Modelling Based on Portable Emission Measurement System, On-Board Diagnostics Data, and Gradient-Boosting Machine Learning," *Energies*, vol. 16, no. 6, p. 2754, Mar. 2023, doi: 10.3390/EN16062754.

[22] M. Mądziel, A. Jaworski, H. Kuszewski, P. Woś, T. Campisi, and K. Lew, "The Development of CO_2 Instantaneous Emission Model of Full Hybrid Vehicle with the Use of Machine Learning Techniques," *Energies*, vol. 15, no. 1, p. 142, Dec. 2021, doi: 10.3390/EN15010142.

[23] R. Liu *et al.*, "Integrated MOVES model and machine learning method for prediction of CO_2 and NO from light-duty gasoline vehicle," *J. Clean. Prod.*, vol. 422, p. 138612, Oct. 2023, doi: 10.1016/J.JCLEPRO.2023.138612.

[24] R. Zhang *et al.*, "A Deep Learning Micro-Scale Model to Estimate the CO_2 Emissions from Light-Duty Diesel Trucks Based on Real-World Driving," *Atmosphere*, vol. 13, no. 9, p. 1466, Sep. 2022, doi: 10.3390/ATMOS13091466.

[25] S. Marco *et al.*, "Forecasting Carbon Dioxide Emissions of Light-Duty Vehicles with Different Machine Learning Algorithms," *Electronics*, vol. 12, no. 10, p. 2288, May 2023, doi: 10.3390/ELECTRONICS12102288.

[26] S. Li, Z. Tong, and M. Haroon, "Estimation of transport CO2 emissions using machine learning algorithm," *Transp. Res. D Transp. Environ.*, vol. 133, p. 104276, Aug. 2024, doi: 10.1016/J.TRD.2024.104276.

[27] S. Moon, J. Lee, H.J. Kim, J.H. Kim, and S. Park, "Study on CO_2 Emission Assessment of Heavy-Duty and Ultra-Heavy-Duty Vehicles Using Machine Learning," *Int. J. Automot. Technol.*, vol. 25, no. 3, pp. 651–661, Jun. 2024, doi: 10.1007/S12239-024-00051-5.

[28] "CO2 Emission by Vehicles." [Online]. Available: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles (Accessed: Apr. 14, 2025).

[29] M. Singh and R. Dubey, "Deep Learning Model Based CO2 Emissions Prediction Using Vehicle Telematics Sensors Data," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 768–777, Jan. 2023, doi: 10.1109/TIV.2021.3102400.

[30] J. Cha, J. Park, H. Lee, and M.S. Chon, "A Study of Prediction Based on Regression Analysis for Real-World CO_2 Emissions with Light-Duty Diesel Vehicles," *Int. J. Automot. Technol.*, vol. 22, no. 3, pp. 569–577, Jun. 2021, doi: 10.1007/S12239-021-0053-Z.

[31] H.T.T. Vu and J. Ko, "Effective Modeling of CO_2 Emissions for Light-Duty Vehicles: Linear and Non-Linear Models with Feature Selection," *Energies*, vol. 17, no. 7, p. 1655, Mar. 2024, doi: 10.3390/EN17071655.

[32] A. Vishnu Priya *et al.*, "CO_2 emissions: machine learning models for assessing the economic & environmental impact of fossil fuels and electric vehicles," *Green Mach. Learn. Big Data Smart Grids-Pract. Appl.*, pp. 99–112, Jan. 2025, doi: 10.1016/B978-0-443-28951-4.00008-3.

[33] Y. Wen *et al.*, "A data-driven method of traffic emissions mapping with land use random forest models," *Appl. Energy*, vol. 305, p. 117916, Jan. 2022, doi: 10.1016/J.APENERGY.2021.117916.

[34] R.F. Melo, N.M. de Figueiredo, M.S.G. Tobias, and P. Afonso, "A Machine Learning Predictive Model for Ship Fuel Consumption," *Appl. Sci.*, vol. 14, no. 17, p. 7534, Aug. 2024, doi: 10.3390/APP14177534.

[35] J. Seo and S. Park, "Optimizing model parameters of artificial neural networks to predict vehicle emissions," *Atmos. Environ.*, vol. 294, p. 119508, Feb. 2023, doi: 10.1016/J.ATMOSENV.2022.119508.

[36] Y. Özüpak, "Machine learning-based fault detection in transmission lines: A comparative study with random search optimization," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 73, no. 2, p. 153229, 2025, doi: 10.24425/BPASTS.2025.153229.

[37] M. Çinar, E. Aslan, and Y. Özüpak, "Comparison and optimization of machine learning methods for fault detection in district heating and cooling systems," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 73, no. 3, p. 154063, 2025, doi: 10.24425/BPASTS.2025.154063.

[38] A.C. Ari, "Mechanical and hydrophobic properties determination of epoxy/ignimbrite/pine waste composites," *Polym. Compos.*, pp. 1–14, 2025, doi: 10.1002/PC.29662.

[39] F. Alpsalaz, Y. Özüpak, E. Aslan, and H. Uzel, "Classification of maize leaf diseases with deep learning: Performance evaluation of the proposed model and use of explicable artificial intelligence," *Chemometrics Intell. Lab. Syst.*, vol. 262, p. 105412, 2025, doi: 10.1016/j.chemolab.2025.105412.