Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 7, No.4, October 2025, pp. 02504019-01 ~ 02504019-015
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v7i4.2097

Boosting-Based Machine Learning Models and Hyperparameter Tuning for Predicting Vehicle Carbon Dioxide Emission

Firman Ridwan Petervan Siburian*, Suharjito

Industrial Engineering Department, BINUS Graduate Program - Master of Industrial Engineering, Bina Nusantara University, Jakarta 11480, Indonesia

*firman.siburian@binus.ac.id

Abstract. Sustainable development and climate change are central agendas in global policy and research. This study examines and compares three ensemble learning models using Gradient Boosting Machine, Categorical Boosting, and Extreme Gradient Boosting for forecasting vehicle carbon dioxide (CO_2) emission. Data preprocessing with Interquartile Range (IQR) and median imputation is among the methods used to address missing values in CO_2 rating and smog rating variables. SHAP and PDP were employed for feature importance analysis and model interpretability. The findings from the third experiment demonstrate that Extreme Gradient Boosting (XGBoost) outperformed other models achieving a Coefficient Determination of 0.9988, Root-Mean-Square Error of 2.1696, Mean-Absolute Error of 0.4977, and Mean-Absolute-Percentage Error of 0.0019. The primary predictive features included combined fuel consumption (liters/100 km), city and highway fuel consumption, ethanol fuel consumption, model year, engine size and diesel consumption. The findings suggest the potential of boosting-based models for supporting sustainable transport planning, policy for emission reduction, and evidence-based policy making.

Keywords: Vehicle Carbon Dioxide Emission, Machine Learning Models, SHAP, Comparative Study, Environmental Sustainability.

(Received 2025-06-06, Revised 2025-09-18, Accepted 2025-09-18, Available Online by 2025-10-09)

1. Introduction

The issue of global warming has become a prominent subject of discussion, particularly concerning the alignment of net zero targets with Sustainable Development Goals (SDGs) [1]. It is driven by the cumulative release of greenhouse gases emissions—more particularly carbon dioxide (CO_2)—into the atmosphere that continues to increase the global temperature for as long as the net emissions exceed zero. Increased energy consumption and unconstrained consumption of fossil fuels has aggravated climate change, exhaustion of resources, and environmental pollution [2]. Net-zero emissions, a situation of equilibrium between the amount of greenhouse gases produced and the amount extracted from the atmosphere, are thus essential in mitigating global warming [3] . The objective of sustainable

development established in 2015 through the Paris Climate Agreement is to set an ambitious target to limit global temperature below 1.5°C compared to pre-industrial levels, highlighting the urgency of addressing climate change by achieving net zero emissions between 2045 and 2060 [4]. Current emission trends exceed this threshold underscoring the necessity to attain net-zero by the latter half of the century. The massive scale of consumption generated from human intervention including excessive use of vehicles, industrial development, the combustion of fossil fuels, agricultural practices and forestry which capture greenhouse gases especially CO_2 [5].

The expansion of the transportation sector in urban regions is a critical factor that substantially contributes to the rise in CO_2 emission, which is increasingly under pressure due to the need for greater mobility and more stringent environmental regulations [6]. Taking into account a range of emission-specific characteristics such as CO, CO_2, hydrocarbons (HC), nitrogen oxides (NOx), and particulate matter (PM) are key factors in the rise of particulate air pollution. With transport demand set to increase, it is essential to evaluate and reduce its environmental impact. This initiative corresponds with the sustainable development pillars associated with goal 7 regarding Affordable and Clean Energy, goal 11 focused on Sustainable Cities and Communities, goal 13 emphasizing Climate Action, and goal 15 concerning Life on Land [7,8]. In this context, machine learning methods have increasingly been utilized in the implementation of data-driven environmental policy [9] where machine learning-based regression models were employed to model urban transport emissions.

Globally, carbon emissions from transportation are increasing, driven by motorization and demographic trends. From 2018 to 2023 several countries including China, the United States, India, Canada, and Indonesia reflected both increases and decreases in CO_2 emission. For instance, China's emission rose by 15.19% from 10.333 to 11.903 $MtCO_2$, while that of the United States declined by 8.68% from 5.378 to 4.911 $MtCO_2$. India's emissions grew by 18.09% from 2.593 to 3.062 $MtCO_2$, and Indonesia's grew by 23.40% from 594 to 733 $MtCO_2$ [10]. In the midst of this global trend, Canada also experienced a 5.18% reduction from 579 to 549 $MtCO_2$. However, the emission continuously improves in alignment with road transport growth based on fossil fuels and making it a suitable region of study for predictive work on emissions with national vehicle datasets [11].

Several implementations of the Machine learning algorithms have been utilized to address the issue of CO_2 emission with enhanced prediction. Among these, boosting algorithms including the Extreme Gradient Boosting mode, followed by Categorical Boosting and Gradient Boosting Machine demonstrate superior predictive capability by effectively learning from complex nonlinear relationships and high dimensional datasets [12–14]. Low error values have been recorded in the literature for these models, with XGBoost recording an RMSE of 2.6554 [12], CatBoost 1.9 [13], and GBM 3.3633 [14]. However, there remains room for improvement in the model's performance, particularly with more optimized hyperparameter tuning and more sophisticated feature interpretation. To fill these gaps, this study makes the following contributions including comparative model assessment with XGBoost, CatBoost, and GBM are contrasted in predictive performance founded on grid search technique and fivefold cross-validation to ensure robustness. The second contribution is using feature importance analysis to enhance feature transparency in decision making involves the use of SHapley Additive exPlanations (SHAP) values to rank and identify the most influential features in predicting CO_2 emission and Partial Dependence Plot (PDP) to examine the relationship between each feature and the expected outcome [15].

2. Material and Method

The work adopts a systematic methodological process with four major phases: (i) data collection, (ii) SHAP value and Partial Dependence Plot-based feature selection, (iii) model development and hyperparameter tuning, and (iv) measurement of performance.

2.1. Data collection

The data employed in the present work is drawn from publicly published data submitted by the Government of Canada and accessed from the Kaggle repository [16]. The data set is widely used in

emission modeling research and contains 26.076 light-duty vehicles recorded between 1995 and 2022. Missing values were removed to ensure the accuracy and consistency of data.

The dataset comprises both numerical and categorical variables. The attributes that can be classified into distinct categories include make (vehicle brand), model, vehicle class, transmission, and fuel type. The model year, engine size, number of cylinders, city and highway fuel consumption (liters/100 km), combined fuel consumption in both liters/100 km and miles per gallon (mpg), CO_2 emission (g/km), CO_2 rating, and smog rating are all numerical features to consider. Units of measurement adhere to transportation standards: liters/100 km indicates liters consumed per 100 kilometers, while mpg signifies miles per gallon. Fuel types include Ethanol (E), Diesel (D), Natural Gas (N), Regular Gasoline (X), and Premium Gasoline (Z). Table 1 and Table 2 give a few sample records from the dataset, detailing technical specifications along with corresponding CO_2 emission and ratings.

Table 1. Sample of CO_2 Vehicle Emission Data

Model Year	Vehicle Brand	Vehicle Class	Engine Size	Cylinders	Transmission	Fuel Type
2001	Chevrolet	Pickup Truck-Small Size	2.2	4	A4	E
2004	Chevrolet	Van-Cargo	6	8	A4	N
2008	Mercedez	Mid-Size	3	6	A7	D
2020	Ford	Pickup Truck: Standard	2.7	6	S10	X
2022	Bugatti	Two-Seater	8	16	AM7	Z

Table 2. Sample of CO_2 Vehicle Emission Data (Continue)

Fuel Consumption City (liters/100 km)	Fuel Consumption Highway (liters/100 km)	Comb (liters/100 km)	Comb (miles per gallon)	CO_2 Emission	CO_2 Rating	Smog Rating
17.7	12.2	15.2	19	243	0	0
20.2	13.8	17.3	16	327	0	0
9	6.1	7.7	37	208	0	0
12	8.9	10.6	27	249	4	5
30.3	20.9	26.1	11	608	1	1

2.2. Feature importance using SHAP and PDP

This work employed two parametric analyses utilizing SHAP and PDP technique for each of three models to improve model transparency (i.e., model-specific utilization). SHAP is an eXplainable AI (XAI) unifying framework used to explain machine learning predictions by representing the marginal contribution of separate input features [17] .

Global interpretability encompasses Partial Dependence technique used to investigate the join effect of predictors, offering deeper insights into feature interactions [18]. These techniques provide a comprehensive insight into the model's operational mechanism and facilitate a precise evaluation of each feature's actual contribution to the output [19]. In order to ensure unit consistency, fuel consumption (mpg) was not included even though it is significant. Smog rating and CO_2 rating were left out because the features are imputed and may bias.

2.3. Experimental setup and hyperparameter tuning

Tree-based boosting methods offer a more adaptable and robust approach for addressing complex nonlinear relationships. This study, thus, did not apply normalization or standardization techniques, as recommended by previous research [20]. In experimental setting, One-hot encoding was applied in this work to convert the categorical variable of fuel type into numerical features for Extreme Gradient Boosting and Gradient Boosting Machine Model.

The data were split at random into 80% train (n = 20.860) and 20% test (n = 5.215) sets. The same fixed random seed (random_state = 42) was used in all to ensure reproducibility of the results. Hyperparameters were optimized by applying grid search methodology, an exhaustive approach that tests pre-defined sets of parameter values to identify the most suitable model specification. This procedure enables comparison on equal terms for all three algorithms under the same evaluation conditions. The experimental framework was segmented into three distinct experiments focused on feature selection. In the initial experiment, five variables were examined: variable Comb (liters / 100 km), vehicle fuel usage in the city (liters/100 km), Fuel consumption on the highway (liters / 100 km), Ethanol (E), and Model year. The second experiment employed the same set of variables, with the addition of engine size as a new variable. The third experiment involved the same variables as the second experiment with the addition of Diesel, resulting in a total of seven variables.

Table 3 compiles the hyperparameter range and literature sources used to inform tuning for each boosting model. These ranges were chosen based on both empirical experimentation and existing literature [21–23]. Three models were configurated with important hyperparameter and determined with upper and lower boundaries. The range value of GBM model is established from optimal values derived according to previous literature, thereafter, refined through fine-tuning around the baseline as the final phase of optimization. The hyperparameter range of three models was evaluated based on training time efficiency, dataset size, and the number of features utilized. From these search ranges, there are 324 combinations for Extreme Gradient Boosting (XGBoost), 16 combinations for Categorical Boosting (CatBoost), and 72 combinations for Gradient Boosting Machine (GBM) utilizing 5-fold cross-validation.

Table 3. Hyperparameter Range for Model Tuning

Model	Parameter	Range Value	Reference
Extreme Gradient Boosting (XGBoost)	N_estimators	100,300,500,700	[21]
	Max_depth	6, 7, 9	
	Subsample	0.5, 0.7, 1.0	
	Gamma	0, 0.1, 0.2	
	Learning rate	0.1, 0.01, 0.001	
Categorical Boosting (CatBoost)	Iterations	40, 200	[22]
	Learning rate	0.01, 0.5	
	Depth	2, 10	
	L2 leaf reg	0.01, 1	
Gradient Boosting Machine (GBM)	N_estimators	190, 200, 210	[23]
	Learning_rate	0.08, 0.1, 0.12	
	Max_depth	5	
	Min_samples_split	2, 3	
	Min_samples_leaf	1, 2	
	Subsample	0.9, 1.0	

Both the models were trained in a fivefold cross-validation setup to minimize overfitting and identify the grid through the sampling of value combinations to achieve optimal parameters. The best parameters were selected based on validation fold performance. Several factors in the implementation of boosting technique as evidenced by prior studies [24].

1. Parallelization and Scalability
   Boosting algorithms are engineered for parallel execution and scalability. Models originating from the boosting category can leverage many processor cores or distributed computing systems to accelerate the data training process and effectively manage large-scale datasets.
2. Improving model performance
   Ensemble learning approach significantly enhances prediction accuracy relative to conventional machine learning algorithms. Boosting operates by combining numerous weak learner models into a singular robust ensemble model, hence facilitating the capture of intricate patterns within the data. By persistently concentrating on challenging-to-predict data, boosting incrementally enhances the model's overall accuracy.
3. Robust against overfitting
   Boosting techniques are effective in mitigating the risk of overfitting. Boosting employs techniques such as regularization to inhibit the model from acquiring excessive detail from the training data. Consequently, boosting models can excel on previously unseen data and deliver consistent outcomes across various datasets.

2.4. Evaluation metrics

This section outlines four standard performance metrics used to evaluate regression models, including:

- R^2 (Coefficient Determination): Proportion of variance in the target variable explained by the model.
- RMSE (Root-Mean-Square Error): Refers to a penalty for larger errors and provides an estimate of model accuracy that is in the same units as the target.
- MAE (Mean-Absolute Error): Estimates average absolute prediction errors.
- MAPE (Mean-Absolute-Percentage Error): Calculates the mean percentage error, providing a normalized measure of accuracy.

These were selected to quantify absolute and relative prediction quality across models according to best practice in Machine learning-based regression analysis [25].

3. Results and Discussion

This section outlines the findings from three regression models—XGBoost, CatBoost, and GBM—for forecasting vehicle CO_2 emission. The process entails Exploratory Data Analysis (EDA), feature selection through SHAP and PDP, hyperparameter adjustment through grid search, and model performance measurement through standard regression metrics.

3.1 Exploratory data analysis

This section defines the preliminary exploratory examination of the data, which was employed to assess the distribution, central tendency, dispersion, and integrity of the data. Table 4 presents the summary of the descriptive statistics for this study, highlighting the key numeric variables.

The data set contained missing data for CO_2 rating (n=18.905) and Smog rating (n=20.015). These were imputed using median imputation. Outliers in numeric attributes including engine size, fuel usage in city and highway, Comb (liters/100 km) and mpg, CO_2 emission, and CO_2 rating were removed to prevent bias in model training.

Table 4. Exploratory Data Analysis

	Mean	Min	Standard Deviation	Median	Max	Skewness	Kurtosis
Model year	2009.576	1995.0	7.726	2010.0	2022.0	-0.200	-1.026
Engine size	3.355	0.8	1.342	3.0	8.4	0.629	-0.416
Cylinders	5.844	2.0	1.797	6.0	16.0	0.891	1.154

	Mean	Min	Standard Deviation	Median	Max	Skewness	Kurtosis
Fuel consumption city	13.036	3.5	3.560	12.6	30.6	0.694	0.841
Fuel consumption highway	9.163	3.2	2.398	8.7	20.9	0.887	0.803
Comb (l/100 km)	11.294	3.6	2.997	10.8	26.1	0.753	0.748
Comb (mpg)	26.784	11.0	7.307	26.0	78.0	1.171	3.569
CO_2 emission	256.679	83.0	63.062	248.0	608.0	0.606	0.616
CO_2 rating	4.660	1.0	1.638	5.0	10.0	0.388	0.323
Smog rating	4.674	1.0	1.791	5.0	8.0	-0.378	-0.682

3.2 Interpretability model-based feature importance

SHapley Additive exPlanations (SHAP) examination was utilized to provide predictions of every variable's contribution towards CO_2 emission in table 5 presents SHAP values-based feature ranking allows for greater transparency and is consistent with explainable AI (XAI) principles [26]. SHAP values not only allow model-specific interpretability but also were found to be more stable and global explainability than other feature attribution methods, as pointed out by Sahraei et al. [27].

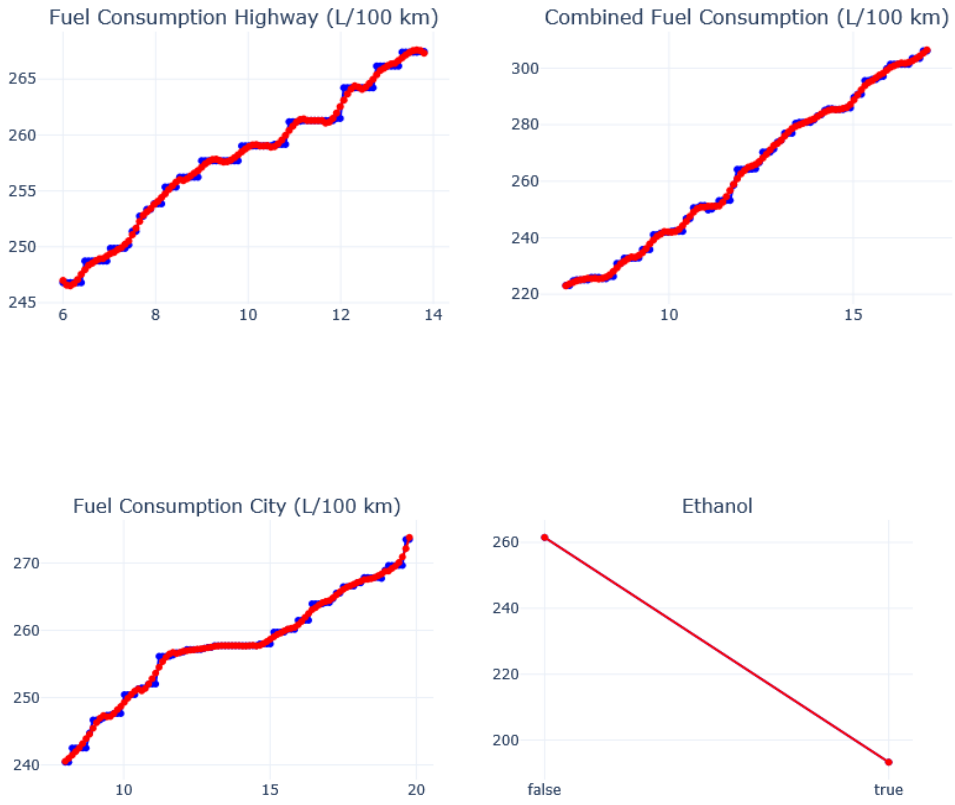
This study utilizes SHAP interpretability to describe model decisions and identify important features from most to least influential that were ranked quantitatively by the average SHAP value. Those that were contributing little were the focus for exclusion in subsequent iterations to help reduce dimensionality and improve model generalization. The model results indicate that the primary factors influencing CO_2 emission in the regression analysis are combined liters and miles per gallon, fuel consumption in city and highway conditions, and ethanol while the elements regular and premium gasoline have minimal impact on the outcome. The use of SHAP is consistent with increasingly required interpretable AI in environmentally impactful decision-making.

Table 5. Importance Feature based on SHAP Values

Variable	Variable Importance based SHAP
Comb (liters / 100 km)	22.582797
Comb (mpg)	11.855192
Fuel consumption city (liters / 100 km)	9.389006
Fuel consumption highway (liters /100 km)	6.658591
Ethanol (E)	4.801257
Model year	1.463873
Engine size (L)	1.127038
Diesel (D)	0.937435
CO_2 rating	0.650921
Natural gas	0.178506
Cylinders	0.097256
Regular gasoline (X)	0.060661
Premium gasoline (Z)	0.018739
Smog rating	0.012613

The specified value weighting using SHAP was examined through one-way partial dependence plots to assess the average relationship between each input variables and the output. Figure 1 presents Partial Dependence Plot (PDP) for seven input variables used in experimental conditions highlighting the effect of each feature to be examined on the predicted target variable and maintaining all other features constant as indicated by the original PDP (represented by the blue diagonal line) and smoothed PDP (represented by the red diagonal line) more distincly illustrates this pattern and emphasizing the global

trend by identifying upward and downard movement. First, Fuel consumption highway (liters/100 km), fuel consumption city (liters/100 km) and comb (liters/100 km) increases with higher values, indicating a correlation with elevated forecasts. The range of highway fuel consumption spans from 6 liters/100 km where a vehicle consumes 6 liters of fuel per 100 km resulting in CO_2 emission of 245 g/km and 14 liters/100 km which indicates higher fuel consumption for the same distance and a notable increase in CO_2 emission to 268 g/km. This contrasts with the trend noted in fuel consumption city where the value exhibits a gradual and consistent increase within the range of 11 to 15 liters/100 km. Second, The next analysis to fuel type where the prediction for diesel grows with rising values indicating a positive correlation with CO_2 emission. Conversely, Ethanol exhibit a decline as the value increases, signifying an inverse relationship with CO_2 emission. Lastly, the analysis of engine size reveals a stable target value of approaximately 256-257 suggesting that variations in engine size have minimal impact on emissions. The significant rise on the blue diagonal line at approximately 261 contrasts with the value indicated on the red diagonal line suggesting the smoothed increase pattern between the feature and predicted outcome.



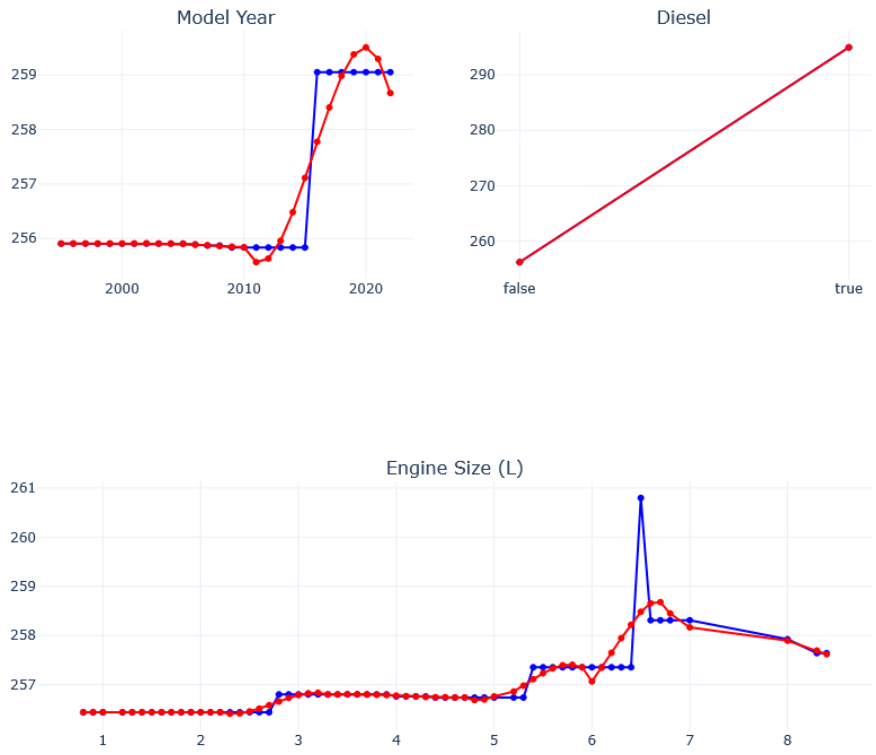


Figure 1. PDP analysis results for input

3.3 Model performance and hyperparameter tuning

Three experiments were performed based on different subsets of the features selected. Grid search-based tuning was performed with each model. Summary of best parameters is provided in Table 6 and model performance metrics—RMSE, MAE, R^2, and MAPE—are presented in Table 7.

Table 6. Optimal Hyperparameters Selected

No.Experiment	Optimal Parameter	Optimal Value
XGBoost (Experiment 1)	N Estimators	300
	Max Depth	6
	Subsample	0.5
	Gamma	0
	Learning Rate	0.1
XGBoost (Experiment 2)	N Estimators	500
	Max Depth	6
	Subsample	0.7
	Gamma	0
	Learning Rate	0.1
XGBoost (Experiment 3)	N Estimators	700
	Max Depth	6
	Subsample	0.5
	Gamma	0.1
	Learning Rate	0.1

No.Experiment	Optimal Parameter	Optimal Value
CatBoost (Experiment 1)	Depth	10
	Iterations	200
	L2 leaf reg	1
	Learning rate	0.5
CatBoost (Experiment 2)	Depth	10
	Iterations	200
	L2 leaf reg	1
	Learning rate	0.5
CatBoost (Experiment 3)	Learning rate	0.5
	Iterations	200
	L2 leaf reg	1
	Depth	10
GBM (Experiment1)	Learning rate	0.12
	Max Depth	5
	Min Samples Leaf	0.12
	Min samples split	3
	N estimators	210
	Subsample	0.9
GBM (Experiment 2)	Learning rate	0.12
	Max Depth	5
	Min Samples Leaf	1
	Min samples split	2
	N estimators	210
	Subsample	0.9
GBM (Experiment 3)	Learning rate	0.1
	Max Depth	5
	Min Samples Leaf	1
	Min samples split	2
	N estimators	210
	Subsample	0.9

Table 7. Model Performance Metrics

Algorithm	Experiment	RMSE		R-Squared		MAE		MAPE	
		Train data	Test data	Train data	Test data	Train data	Test data	Train data	Test data
CatBoost	1	2.3568	4.2476	0.9986	0.9954	0.8897	1.3669	0.3628	0.5452
	2	1.3104	3.2876	0.9996	0.9973	0.6061	1.0009	0.0025	0.0040
	3	0.6888	2.2072	0.9999	0.9988	0.3664	0.6155	0.0015	0.0024
XGBoost	1	2.9763	4.4567	0.9977	0.9949	1.0339	1.3992	0.0041	0.0055
	2	1.4342	3.2471	0.9994	0.9973	0.5457	0.8794	0.0022	0.0034
	3	0.6657	2.1696	0.9998	0.9988	0.2866	0.4977	0.0011	0.0019
GBM	1	3.4231	4.5991	0.9970	0.9946	1.1268	1.4209	0.0045	0.0055
	2	2.4244	3.5438	0.9985	0.9968	0.8979	1.1108	0.0036	0.0044
	3	1.2445	2.2519	0.9996	0.9987	0.5438	0.6586	0.0021	0.0025

The optimum model was XGBoost (Experiment 3), with the most minimal RMSE and MAPE, reflecting high accuracy and minimal bias. Table 8 presents ANOVA test results for CO_2 emission and according to the ANOVA test, F_{stat} = 622, 141 in XGBoost, 601,105 in CatBoost, and 577,419 in GBM. While $F_{critical}$ = 2.0113 in the same three model. Indicating $F_{stat} > F_{critical}$. This finding indicates that null hypothesis should be rejected despite being true (H0: No significant difference between the models) and accepts the alternative hypothesis (H1: Significant differences exist between the models) as indicated by a probability value (P-value) of less than 0.05 which provides significant differences among the models [28].

Table 8. ANOVA Test with Feature Selection

Source of Variance	Sum of Squares	Degree of Freedom	Mean Square	F- Statistic	P-Value	F-Critical
XGBoost						
Between Groups (Regression)	20,531	7	2,933	622, 141	1.110e-16	2.0113
Within Groups (Residual)	24,548	5207	4.7146	–	–	–
Total	20,556	5214	–	–	–	–
CatBoost						
Between Groups (Regression)	20,531	7	2,933	601,105	1.110e-16	2.0113
Within Groups (Residual)	25,406	5207	4.8793	–	–	–
Total	20,556	5214	–	–	–	–
GBM						
Between Groups (Regression)	20,529	7	2,932	577,419	1.110e-16	2.0113
Within Groups (Residual)	26,447	5207	5.0792	–	–	–
Total	20,556	5214	–	–	–	–

3.4 Overfitting assessment and avoidance

The model's behaviour can be assessed through the range of values produced during both testing and training, which can be classified as either overfitting or underfitting. Overfitting arises from extended computation time, leading the model to concentrate excessively on the training data [26]. This results from inadequate regularization contributions or the application of an unsuitable and irrelevant learning rate for the data variables. Figure 2 illustrates possible overfitting identified in the third experiment of XGBoost, evidenced by the RMSE test value exceeding the RMSE train value.

To avoid this:

- Early stopping parameters were recommended to prevent overfitting or excessive training by diminishing the period of iterative training while maintaining prediction accuracy.
- Including additional heterogeneous training samples and drop-out mechanisms were recommended to reduce variance.

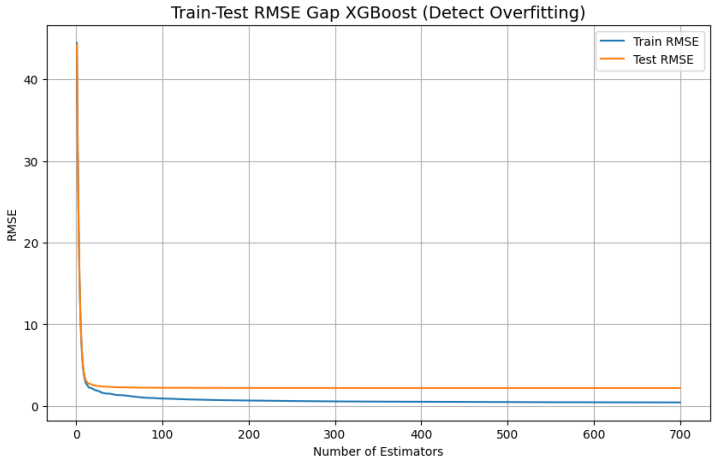


Figure 2. RMSE graph to detect overfitting using XGBoost

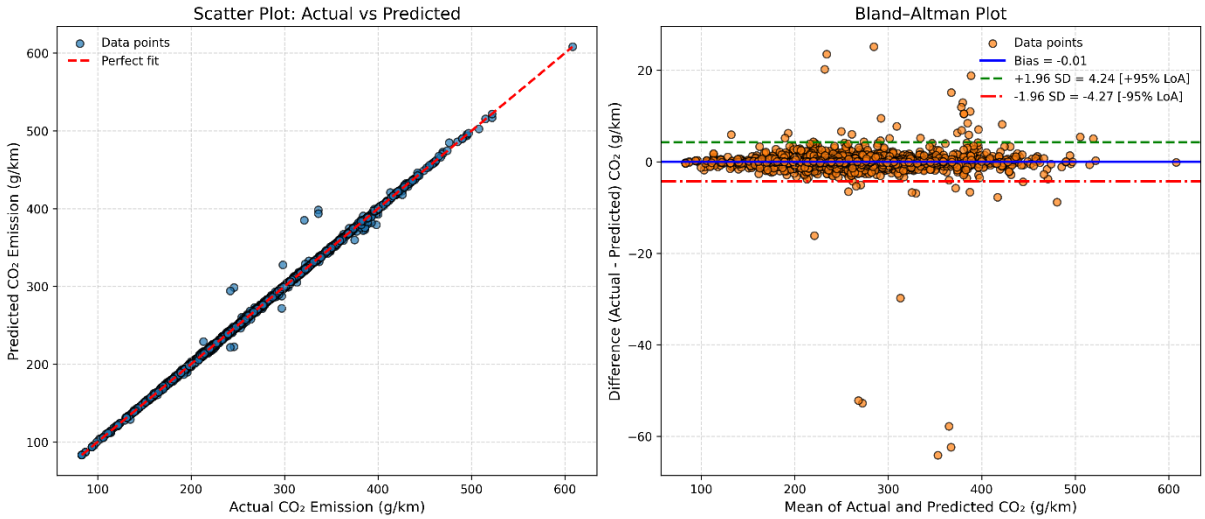


Figure 3. Actual and predicted CO_2 emission of XGBoost model

Figure 3 presents the actual CO_2 versus the predicted values using scatter plot graphs (a) and Bland-Altman analysis (b) generated by XGBoost that was identified as the optimal model. The principle of a scatter plot visualizes the alignment of the model's prediction to fit the actual data. A set of data points are nearly aligned with the diagonal line represents the optimal relationship between the difference of actual and predicted values to CO_2 emission. In direct comparison, bland-Altman analysis was divided into four categories such as differences (data points), bias (= mean difference) was found as -0.01 with the minimum and maximum 95% limits of agreement between 2 methods at -4.27 and 4.24. Most data points within the baseline area of [−95% LoA, +95% LoA] or ± 1.96 standard deviations indicating with no significant difference between the reliability [29] [30]. Bland-Altman plot is good to use to validate machine learning-based regression [31].

3.5 Cross-validation for robustness

To check generalizability validity, fivefold cross-validation was performed on the optimal XGBoost model. Mean values suggest consistent performance across all folds (Table 9).

Table 9. Fivefold Cross-Validation Results (XGBoost, Third Experiment)

Folds	R^2	RMSE	MAE	MAPE
Fold-1	0.9985	2.4844	0.5623	0.21
Fold-2	0.9994	1.5264	0.5050	0.19
Fold-3	0.9977	3.0185	0.5881	0.22
Fold-4	0.9990	2.0093	0.5517	0.20
Fold-5	0.9988	2.2120	0.5392	0.21
Mean	2.2501	0.9987	0.5493	0.21

Model robustness refers to a model's capacity to sustain accurate performance throughout diverse situations, particularly when evaluated with new data. It is essential to illustrate that the model is genuinely stable and useful. This study's measurements can discover the optimal model that ensures both predictive accuracy and robustness to varying experimental settings. A key metric employed R^2, MAE, MAPE and RMSE where a high r-squared indicating that the model effectively captures the significant factors affecting performance on carbon dioxide prediction. XGBoost performed consistently and showed high R^2 values of around 0.998 and 0.999 for most folds. A value of R-squared near 1 suggest a robust correlation between independent and dependent variables, indicating that the model demonstrates a strong fit. Meanwhile RMSE and MAE have a similar purpose, the closer these error metric are to 0, the greater the alignment between predicted outcomes and actual data, thereby providing a consistent error metric relative to the target variable and enhancing understanding when assesing predictive accuracy.

3.6 Model comparison to the previous literature

To make the results of this study valid and put them into perspective alongside other research, a comparison was drawn with other research that had utilized similar boosting-based machine learning approaches to predict CO_2 emissions. Table 10 consolidates model performance data from various sources including recent literature benchmark values for XGBoost, CatBoost, and GBM. Mądziel [32] performed an extensive benchmarking of the boosting algorithms with different dataset and experiment on environmental prediction issues and concluded that XGBoost performs better in general compared to other ensemble models regarding predictive accuracy, stability, and computational expense. The findings of their work justify the conclusions of this work, where XGBoost not only yielded the better MSE and R^2 result, but also had superior generalization capability across diverse feature sets. This study's grid search-tuned and fivefold cross-validation-tested XGBoost model recorded an RMSE of 2.1696, R^2 of 0.9988, and MAPE of 0.0019, outperforming the cited literature's best-performing results. This study demonstrates a significantly improved MAPE of 0.0019, indicating a high level of forecasting accuracy compared to previous studies [33].

Table 10. Model Benchmark Comparison with Previous Studies

Reference	Best Model	RMSE	MAE	R-Squared	MAPE
[12]	XGBoost	2.6554	Not Available	0.9973	24.3
[13]	CatBoost	1.9	2.41	99.6	Not Available
[14]	GBM	3.3633	2.2706	0.9973	0.8854
This study	XGBoost	2.1696	0.4977	0.9988	0.0019

In addition to addressing technical issues, data-driven management is essential for decision-making in creating operational excellence. Key managerial insight highlights the importance of integrating AI-prediction methods with large-scale datasets specifically recommending the use of XGBoost model for CO_2 emission forecasting. This study identifies three performance indicators for light-duty vehicles including economic indicator which encompass vehicle maintenance and fuel usage costs. Secondly, environmental indicator encompass the volume of CO_2 emission generated by each vehicle, quantified

in liters per 100 kilometers. Lastly, vehicle indicator encompass the number of trips quantified by distance (measured on a daily basis), the type of fuel, and the vehicle engine classified in Euro emission standards. In Indonesia context, ethanol serves a viable alternative fuel potentially leading to a reduction in carbon dioxide emission by 30-70 percent relative to gasoline, as supported by existing literature [34]. Potential alternative fuel options for implementation in Indonesia encompass fuel blends such as B15, B20, and B30 which consist of 15 percent, 20 percent, and 30 percent biofuel, respectively [35]. This study proposes policy recommendation aimed at enhancing the environmental system in the transportation sector, which are categorized into three spesific area. The initial policy implementation encourages light-duty vehicle users to consider environmentally friendly characteristics when purchasing a vehicle such as certified emission reductions, carbon offsets, or renewable energy certificates in accordance with regional regulations. This approach assists users in estimating the mileage utilized by the vehicle while also considering cost associated with fuel consumption. The second policy implementation for automotive manufacturing is to emphasize the importance of developing vehicle type that prioritize fuel economy and to adopt Artificial Intelligence of Things - powered predictive systems with a visualization dashboard for real-time emissions monitoring. The third policy implementation for fuel producers should identify other options for increasing the production of environmentally friendly fuels through cross-industry collaboration to produce biodiesel, ethanol, and hydrogen fuel in supporting the implementation of road transport emission reduction strategies and providing robust guidance for decision-makers to achieve sustainability outcomes.

4. Conclusion

The study demonstrates the efficacy of boosting-based machine learning algorithms specifically using Extreme Gradient Boosting-XGBoost, Categorical Boosting-CatBoost, and Gradient Boosting Machine-GBM in predicting vehicle carbon dioxide emission from actual Canadian vehicle data. Through systematic SHAP and PDP-based feature selection, the model interpretation was reduced from 19 to 7 predictor variables, significantly enhancing interpretability without any loss in predictive accuracy. Among the three algorithms, the third XGBoost configuration did best with R^2 = 0.9988, RMSE = 2.1696, MAE = 0.4977, and MAPE = 0.0019, with its best performance in CO_2 emission prediction. The highest value feature variables were combined fuel consumption (liters/100 km), fuel consumption city and highway, engine size, model year, ethanol, and diesel use. The findings confirm that model interpretability and hyperparameter tuning method such as SHAP and PDP can significantly help model transparency, simplicity, and daily applications in emission prediction. However, overfitting risk, especially in small or imbalanced regression of data sets, requires the use of regularization methods.

Author contribution

Suharjito: Responsible for research conceptualization, funding acquisition, supervision, methodology, review and data validation. Firman Ridwan Petervan Siburian: Handled for data collection, wrote the initial draft, editing, data visualization and analysis. All authors have reviewed and consented to the study content in the final version of the manuscript for publication.

Data availability

The dataset is accessible through the official open data portal of the Canadian government and the Kaggle repository. https://www.kaggle.com/datasets/abhikdas2809/canadacaremissions
https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64

References

[1] Govindan K. How digitalization transforms the traditional circular economy to a smart circular economy for achieving SDGs and net zero. Transp Res E Logist Transp Rev 2023;177. https://doi.org/10.1016/j.tre.2023.103147.

[2] Cuce E, Cuce PM, Riffat S. Thin film coated windows towards low/zero carbon buildings: Adaptive control of solar, thermal, and optical parameters. Sustainable Energy Technologies and Assessments 2021;46:101257. https://doi.org/10.1016/J.SETA.2021.101257.
[3] Keys DL. Getting to Net Zero by 2050. Using NEPA to Combat Global Warming 2024:129–56. https://doi.org/10.1007/978-3-031-69316-8_7.
[4] van Soest HL, den Elzen MGJ, van Vuuren DP. Net-zero emission targets for major emitting countries consistent with the Paris Agreement. Nat Commun 2021;12. https://doi.org/10.1038/s41467-021-22294-x.
[5] Regufe MJ, Pereira A, Ferreira AFP, Ribeiro AM, Rodrigues AE. Current developments of carbon capture storage and/or utilization–looking for net-zero emissions defined in the paris agreement. Energies (Basel) 2021;14. https://doi.org/10.3390/en14092406.
[6] Singh S, Kulshrestha MJ, Rani N, Kumar K, Sharma C, Aswal DK. An Overview of Vehicular Emission Standards. Mapan - Journal of Metrology Society of India 2023;38:241–63. https://doi.org/10.1007/s12647-022-00555-4.
[7] Bachmann N, Tripathi S, Brunner M, Jodlbauer H. The Contribution of Data-Driven Technologies in Achieving the Sustainable Development Goals. Sustainability (Switzerland) 2022;14. https://doi.org/10.3390/su14052497.
[8] Kwilinski A, Lyulyov O, Pimonenko T. Environmental Sustainability within Attaining Sustainable Development Goals: The Role of Digitalization and the Transport Sector. Sustainability (Switzerland) 2023;15. https://doi.org/10.3390/su151411282.
[9] Zhu B, Hu S, Chen X (Michael), Roncoli C, Lee DH. Uncovering driving factors and spatiotemporal patterns of urban passenger car CO2 emissions: A case study in Hangzhou, China. Appl Energy 2024;375. https://doi.org/10.1016/j.apenergy.2024.124094.
[10] Zhu L. Comparative evaluation of CO2 emissions from transportation in countries around the world. J Transp Geogr 2023;110:103609. https://doi.org/10.1016/J.JTRANGEO.2023.103609.
[11] Andrew RM. A comparison of estimates of global carbon dioxide emissions from fossil carbon sources. Earth Syst Sci Data 2020;12:1437–65. https://doi.org/10.5194/ESSD-12-1437-2020.
[12] Gurcan F. Forecasting CO2 emissions of fuel vehicles for an ecological world using ensemble learning, machine learning, and deep learning models. PeerJ Comput Sci 2024;10. https://doi.org/10.7717/PEERJ-CS.2234.
[13] Natarajan Y, Wadhwa G, Sri Preethaa KR, Paul A. Forecasting Carbon Dioxide Emissions of Light-Duty Vehicles with Different Machine Learning Algorithms. Electronics (Switzerland) 2023;12. https://doi.org/10.3390/electronics12102288.
[14] Guo X, Kou R, He X. Towards Carbon Neutrality: Machine Learning Analysis of Vehicle Emissions in Canada. Sustainability (Switzerland) 2024;16. https://doi.org/10.3390/su162310526.
[15] Meddage P, Ekanayake I, Perera US, Azamathulla HMd, Md Said MA, Rathnayake U. Interpretation of Machine-Learning-Based (Black-box) Wind Pressure Predictions for Low-Rise Gable-Roofed Buildings Using Shapley Additive Explanations (SHAP). Buildings 2022;12:734. https://doi.org/10.3390/buildings12060734.
[16] Vehicle Emissions and Smog Rating Classification n.d. https://www.kaggle.com/datasets/abhikdas2809/canadacaremissions (accessed August 18, 2025).
[17] Abdollahi A, Pradhan B. Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. Science of the Total Environment 2023;879. https://doi.org/10.1016/j.scitotenv.2023.163004.
[18] Tudor C, Sova R, Stamatiou P, Vlachos V, Polychronidou P. Future-Proofing EU-27 Energy Policies with AI: Analyzing and Forecasting Fossil Fuel Trends. Electronics (Switzerland) 2025;14. https://doi.org/10.3390/electronics14030631.

[19] Ukwaththa J, Herath S, Meddage DPP. A review of machine learning (ML) and explainable artificial intelligence (XAI) methods in additive manufacturing (3D Printing). Mater Today Commun 2024;41. https://doi.org/10.1016/j.mtcomm.2024.110294.
[20] Arora G, Kumar D, Singh B. Tree based Regression Models for Predicting the Compressive Strength of Concrete at High Temperature. IOP Conf Ser Earth Environ Sci, vol. 1327, Institute of Physics; 2024. https://doi.org/10.1088/1755-1315/1327/1/012015.
[21] Laghmati S, Hamida S, Hicham K, Cherradi B, Tmiri A. An improved breast cancer disease prediction system using ML and PCA. Multimed Tools Appl 2024;83:33785–821. https://doi.org/10.1007/s11042-023-16874-w.
[22] Xiao W, Wang C, Liu J, Gao M, Wu J. Optimizing Faulting Prediction for Rigid Pavements Using a Hybrid SHAP-TPE-CatBoost Model. Applied Sciences (Switzerland) 2023;13. https://doi.org/10.3390/app132312862.
[23] Yazici C, Domínguez-Gutiérrez FJ. Machine learning techniques for estimating high–temperature mechanical behavior of high strength steels. Results in Engineering 2025;25. https://doi.org/10.1016/j.rineng.2025.104242.
[24] Kharazi Esfahani P, Peiro Ahmady Langeroudy K, Khorsand Movaghar MR. Enhanced machine learning—ensemble method for estimation of oil formation volume factor at reservoir conditions. Sci Rep 2023;13. https://doi.org/10.1038/s41598-023-42469-4.
[25] Mustapha IB, Abdulkareem M, Jassam TM, AlAteah AH, Al-Sodani KAA, Al-Tholaia MMH, et al. Comparative Analysis of Gradient-Boosting Ensembles for Estimation of Compressive Strength of Quaternary Blend Concrete. Int J Concr Struct Mater 2024;18. https://doi.org/10.1186/s40069-023-00653-w.
[26] Hakim S Bin, Adil M, Acharya K, Song HH. Decoding Android Malware with a Fraction of Features: An Attention-Enhanced MLP-SVM Approach, 2025, p. 187–209. https://doi.org/10.1007/978-981-96-3531-3_10.
[27] Sahraei MA, Li K, Qiao Q. A Multi-Stage Feature Selection and Explainable Machine Learning Framework for Forecasting Transportation CO2 Emissions. Energies 2025, Vol 18, Page 4184 2025;18:4184. https://doi.org/10.3390/EN18154184.
[28] Alsaadi N. Comparative analysis and statistical optimization of fuel economy for sustainable vehicle routings. Sustainability (Switzerland) 2022;14. https://doi.org/10.3390/su14010064.
[29] Ecker H, Adams NB, Schmitz M, Wetsch WA. Feasibility of real-time compression frequency and compression depth assessment in CPR using a “machine-learning” artificial intelligence tool. Resusc Plus 2024;20. https://doi.org/10.1016/j.resplu.2024.100825.
[30] Peng Y, Luo Y, Yan J, Li W, Liao Y, Yan L, et al. Automatic measurement of fetal anterior neck lower jaw angle in nuchal translucency scans. Sci Rep 2024;14. https://doi.org/10.1038/s41598-024-55974-x.
[31] Danza L, Belussi L, Meroni I, Mililli M, Salamone F. Hourly calculation method of air source heat pump behavior. Buildings 2016;6. https://doi.org/10.3390/buildings6020016.
[32] Mądziel M. Predictive methods for CO2 emissions and energy use in vehicles at intersections. Sci Rep 2025;15. https://doi.org/10.1038/s41598-025-91300-9.
[33] Meng Y, Noman H. Predicting CO2 Emission Footprint Using AI through Machine Learning. Atmosphere (Basel) 2022;13. https://doi.org/10.3390/atmos13111871.
[34] Utilization of Biopertalite for Fuel Efficiency and Reduction in CO and CO2 Gas Emissions in Four-Wheel Motor Vehicles. Makara Journal of Technology 2022;26:117–23. https://doi.org/10.7454/mst.v26i3.1603.
[35] Ramadhan R, Mon MT, Tangparitkul S, Tansuchat R, Agustin DA. Carbon capture, utilization, and storage in Indonesia: An update on storage capacity, current status, economic viability, and policy. Energy Geoscience 2024;5. https://doi.org/10.1016/j.engeos.2024.100335.