International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Machine Learning-Based CO2 Emission Prediction to Support Sustainable Urban Development

Fatema Tuj Johora [1,*], Md Badhan Ahmed Topu [2], Md Mostafizur Rahman [1] and Kazi Tausin Islam [1]

[1] *Urban and Rural Planning Discipline, Khulna University, Khulna-9208, Bangladesh.*
[2] *Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh.*

International Journal of Science and Research Archive, 2025, 17(03), 939-956

Publication history: Received on 12 November 2025; revised on 25 December 2025; accepted on 27 December 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.17.3.3344

Abstract

Accurate identification of CO_2 emissions from vehicle has become important for sustainable urban planning strategies aimed at mitigating global warming. This study presents a framework for predicting CO_2 emissions using data collected from Canada's complete vehicle fleet. We have applied a voting-based ensemble approach that aggregates five feature selection algorithms such as SelectKBest, Lasso, Recursive Feature Elimination, Random Forest importance, and mutual information to identify the most influential predictors in the dataset. Subsequently, we have evaluated the predictive performance of various machine learning (ML) and deep learning (DL) models using the six highest-ranked features, including combined fuel consumption, engine size, and fuel type. Our analysis shows that the Random Forest Regressor substantially outperforms competing models, achieving a R^2 value of 0.9976 including the lowest root mean squared error (RMSE) 2.847 g/km. These results highlight the strength of the ensemble framework in generating precise CO_2 emission estimates. For sustainable urban transportation planning, the suggested method offers a practical and data-driven framework for decision making. This application can help planners and policymakers to formulate comprehensive strategies to reduce carbon emissions and encourage low-carbon urban development.

Keyword: CO_2 Emissions; Low-Carbon; Random Forest; Sustainable Urban Planning; Transport; Machine Learning; Voting Ensemble.

1. Introduction

The global energy demand and the rapid process of industrialization have resulted in the acute growth of atmospheric greenhouse gas (GHG) concentrations, the major contributor of which is carbon dioxide (CO_2). The rapid rise in global energy demand and fossil-fuel consumption has significantly increased atmospheric CO_2 concentrations, intensifying climate change and imposing serious environmental and socio-economic consequences [1,2]. One of the biggest and most rapidly expanding sources of these emissions is the transportation sector, which is largely powered by fossil fuel and generates around 16.2% of total CO_2 emissions in the world [3]. As transportation remains one of the fastest-growing contributors to these emissions, accurate prediction and monitoring of vehicle-based CO_2 output have become essential for shaping effective mitigation policies and advancing sustainable urban planning [4]. As a result, forecasting and tracking of CO_2 emissions of vehicles can no longer be a only technical issue but a necessary process in the development of effective policies to reduce climate change and help countries achieve such essential goals as carbon neutrality [5,6].

Machine learning (ML) offers a powerful tool to uncover emission patterns and guide cities toward data-driven, low-carbon mobility strategies [1]. Governments and international organizations, including Canada and its target of 40-45% decline in CO_2 emissions by 2030 are actively trying to find effective and trustworthy instruments that could be used to

* Corresponding author: Fatema Tuj Johora

regulate activities, instigate more eco-friendly technologies, and do urban design [3,6,7,8]. Conventional methods of emission prediction, frequently based on models of statistics and time-series, often have drawbacks of limited efficiency, scalability and the capacity to reflect the complex and non-linear interactions of a variety of vehicle operating conditions and specifications [2,7,9].

Over the last few years, the convergence of big data accessibility in the automotive sector and the development of Artificial Intelligence (AI) has placed the concepts of Machine Learning (ML) and Deep Learning (DL) as strong competitors to emission prediction [2,4,10]. Investigations have been able to use the parameters of vehicle design (e.g. engine size, fuel consumption, and type of transmission) and operational dynamics (e.g. speed and acceleration) to construct high-precision predictive models [4,6,11]. The studies have shown the predictive power of advanced ML models, such as Support Vector Machines (SVM), Random Forest (RF), and ensemble approaches such as XGBoost to be higher than older statistical methods, with coefficients of determination (R^2) less often less than 0.95, but frequently near-perfect scores [3,9,12,13].

Although these sophisticated ML and DL models have proven to be highly predictive, there is a serious problem that needs to be resolved which involves determining the most ideal subset of features that will be most useful in the prediction exercise. Some frameworks have assimilated feature selection methods (e.g. the use of Neighborhood Mutual Information (NMI) in stacking ensembles [14] or are implicit (e.g. Lasso Regression) [3], they are usually based on a single selection algorithm or a deterministic, fixed procedure.

2. Voting-Based Ensemble Feature Selection (VBE-FS) Framework

In an effort to overcome the aforementioned instability and possible bias of the single-method feature selection, the current paper argues for incorporating a new approach: Voting-Based Ensemble Feature Selection (VBE-FS) Framework [15,16,17]. This framework is created to utilize the aggregate knowledge of several, heterogeneous feature ranking algorithms prior to preparation of the ultimate prediction model.

When predicting carbon dioxide (CO_2) emissions by the transportation sector, it has changed considerably in the last several decades. The shift towards advanced methods of Artificial Intelligence (AI) and Machine Learning (ML) has compelled by the demands of increased accuracy, as well as the presence of large-scale automotive data, which has pushed us out of our conventional approaches based on statistical modelling. Combining several ranking algorithms, the idea of our method is to remove noise better than alternative single-method selection, which will improve the performance and interpretability of the resulting Ensemble Learning predictors.

Traditionally, the forecasting of CO_2 emissions was based on the univariate time-series analysis and statistics. Techniques like the Autoregressive Integrated Moving Average (ARIMA) and the Grey Models (GM) were the norms in terms of predicting the national emission patterns. As an example, [9] and Kumari and Singh [18] compared these conventional statistical models with the modern ML algorithms based on time-series data. Both papers were consistent in concluding that the statistical models such as SARIMAX and Holt-Winters forecasted baselines but failed to capture non-linear trends and multi-dependencies found to exist within the environment data.

These weaknesses of these statistical approaches, namely, their inefficiency and generality, led to the emergence of AI-based modelling. The shift has been emphasized by Sidana [2], who has shown that ML models optimized via methods such as Grid Search may be much more successful at reducing the Mean Squared Error (MSE) and the R^2. In the same manner, [19] applied Support Vector Machines (SVM) and Regression Trees to the Canadian vehicle data, and SVM produced almost per cent accuracy (R^2 =100%), but at the expense of longer training times. These results all confirmed the fact that ML algorithms provide a stronger alternative to the modelling of the complex relationship between vehicle engine characteristics and exhaust emissions.

3. Deep Learning for Emission Forecasting

With the growth of computational power, Deep learning (DL) models became one of the potent tools to process high-dimensional data. Al-Nefaie & Aldhyani [4] examined the effectiveness of the LSTM and BiLSTM models based on a Kaggle data set of vehicle specifications. A better R^2 of 93.78 was attained using their BiLSTM model, which is effective at the policy-level forecasting. Similarly, LSTM on real-world data obtained using light-duty diesel trucks through portable emission measurement system (PEMS) and determined that deep learning has the capacity to predict dynamic variables, including vehicle speed, road slope, and acceleration ($R^2 > 0.98$) (S. Li et al., 2024).

To support the approaches of DL further, Chukwunonso [7] and Nassef [5] reported other superior neural network architectures such as Layer Recurrent Neural Networks (L-RNN) and Feed-Forward Neural Networks (FFNN). Their model showed that it had a Mean Absolute Percentage Error (MAPE) of 1.187% with L-RNN, and LSTM is more effective than other models individually with regard to forecasting the emissions of Saudi Arabia. Nonetheless, due to their high level of accuracy, deep learning models have been criticized in that they tend to be rather black-box and require a significant number of computational resources.

There is a paradigm shift with regard to Ensemble Learning, which is the combination of many weak learners to form a strong learner. This strategy has long been shown to have a better accuracy to computation efficiency ratio than standalone DL models. The comparative studies of up to 18 algorithms by Gurcan [20] found that the ensemble methods, i.e. Extreme Gradient Boosting (XGBoost), Random Forest (RF), and LightGBM, were characterized by the most accurate predictions with lower error rates than deep learning models, such as CNNs and GRUs. On the same note, Natarajan [6] and Chadha [3] emphasized the success of boosting algorithms. The study by Natarajan [6] on CatBoost showed that it could make accurate predictions with small inputs and was memory-efficient, and the research of Chadha and his associates on XGBoost and RF showed the accuracy went to 99.8%. A two-stage forecasting process involving SVR and ANN and RF resulted in a forecast error decreased by more than 43 percent through the hybrid models over single-stage forecasting models [21]. Zhang [14] developed a complex decomposition-ensemble method and stacking strategy with the lowest MAPE value of 5.45% on multi-step prediction.

Although much effort has been put into the predictive capability of classifiers, the optimization of input features has been of great importance but under-researched. The majority of current works either employ the conventional preprocessing methods or promote the feature selection abilities inherent to such algorithms as Lasso Regression or the importance scores of Tree-based models [3,22].

4. Methodology

The methodology follows a comprehensive workflow, beginning with data acquisition and proceeding through preprocessing, advanced feature selection, model training, and evaluation. The entire process was implemented in Python, utilizing libraries such as Pandas for data manipulation and Scikit-learn for machine learning. The overall workflow is shown in Figure 1.

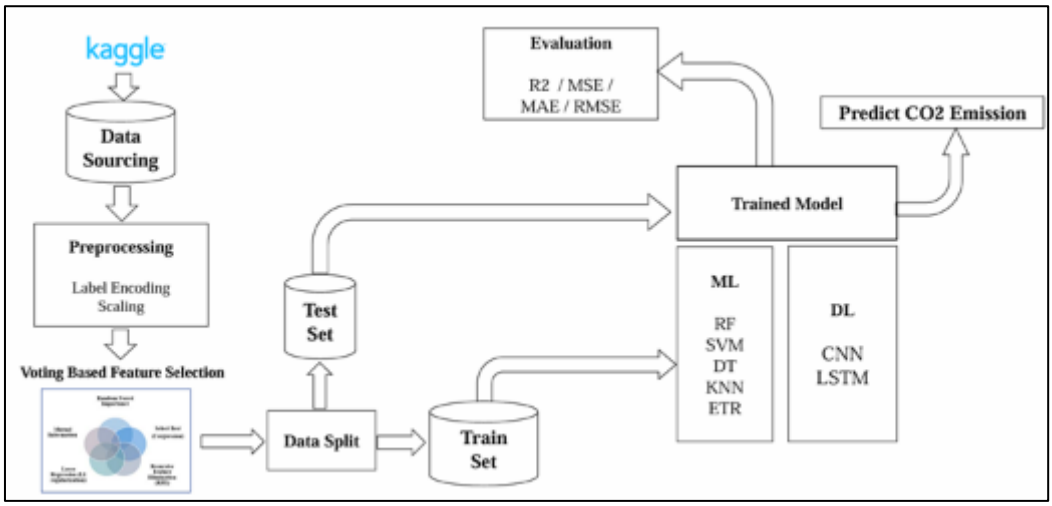

Figure 1 Conceptual Understanding of Research**.**

4.1. Dataset Description

This study utilizes the " CO_2 Emissions_Canada" dataset, a publicly available resource sourced from the Canadian government's open data portal (Debajyoti Podder, n.d.). The dataset contains 7,385 instances and 12 features detailing vehicle specifications and their corresponding CO_2 emission ratings. The key features include categorical attributes such as Make, Model, Vehicle Class, Transmission, and Fuel Type, as well as numerical attributes like Engine Size(L), Cylinders, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), and Fuel Consumption Comb (L/100 km). The target variable for prediction is CO_2 Emissions(g/km). The Figure 2 shows the top 5 rows of the dataset.

Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.7	10.6	27	244
ACURA	RLX	MID-SIZE	3.5	6	AS8	Z	11.9	7.7	10.0	28	230
ALFA ROMEO	4C	TWO-SEATER	1.8	4	AM6	Z	9.7	6.9	8.4	34	193
ASTON MARTIN	DB9	MINICOMPACT	5.9	12	A6	Z	18.0	12.6	15.6	18	359
ASTON MARTIN	V8 VANTAGE	TWO-SEATER	4.7	8	AM7	Z	17.4	11.3	14.7	19	338

Figure 2 Dataset Snapshot

Figure 3 displays the Frequency distribution and Kernel density estimation curve of the key numerical variables of the data set and the variable of interest CO_2 Emissions. On a further analysis of the structural attributes, one can observe that Engine Size and Cylinders have multi-modal, discrete distributions. Cylinders plot: The three clear peaks at 4, 6 and 8 cylinders are the common engine configurations, and Engine Size has common capacity clusters indicating that these variables are useful as categorical proxies in spite of their numerical values.

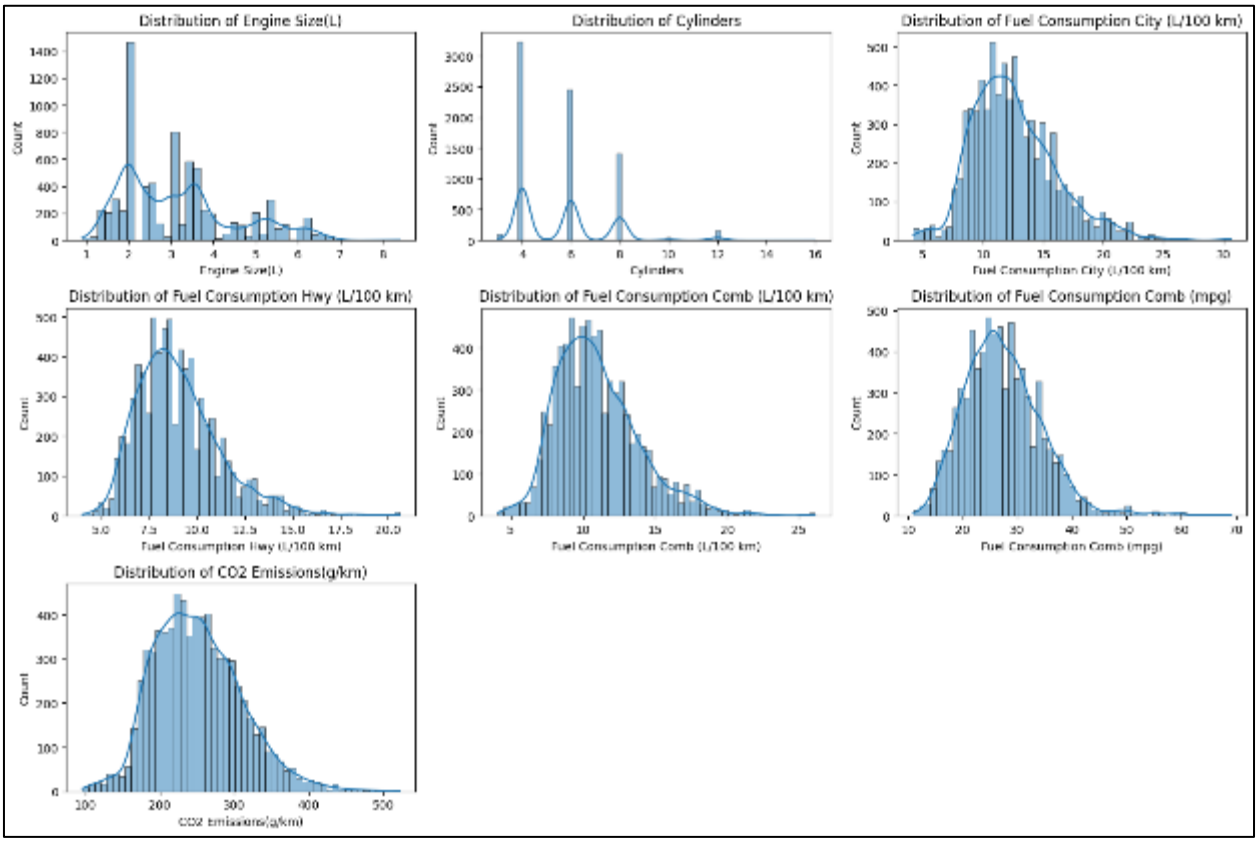

Figure 3 Distribution of Numerical Features

The metrics of fuel consumption, in their turn, such as Fuel Consumption City, Fuel Consumption Hwy and Fuel Consumption Comb are characterized by a strong right-skewed distribution. This shows that although most vehicles are moderately fuel-efficient (8-12 L/100 km) there is a big long tail of vehicles that is high consuming vehicles, which is probably made up of the heavy-duty or high-performance types. More importantly, CO_2 Emissions (g/km) is the target variable, which is going to be symmetrically distributed with a mean value of 250 g/km. This normality is very beneficial in regression modeling since it means that the predictive models will not be overly invalidated by extreme outliers concerning the output space.

A granular analysis of the probability distribution of the target variable, CO_2 Emissions has been showed in figure 4. The histogram, with the quantity of the emission on the x-axis and the ratio of the frequency on the y-axis, proves that the target data is approximately of a Gaussian (normal) distribution with the mean at 250 g/km.

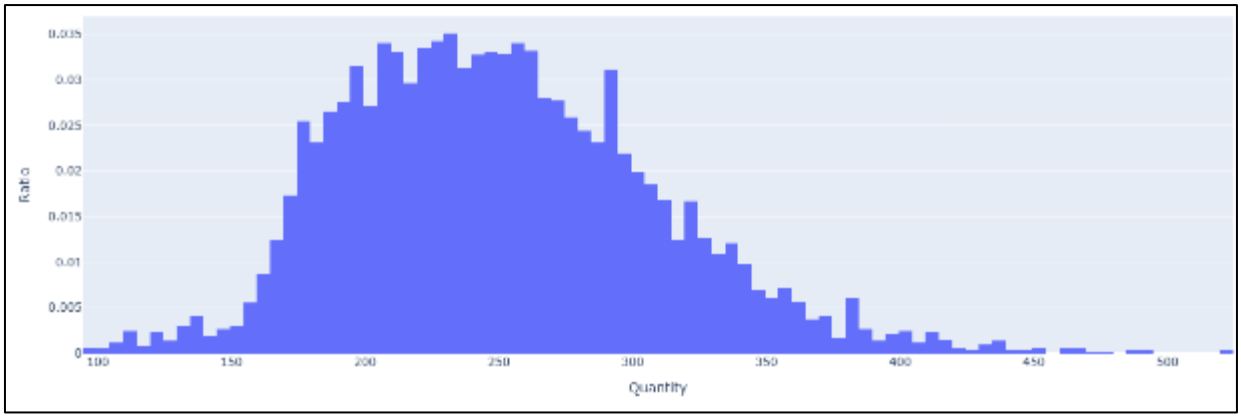

Figure 4 Co_2 Distribution

Although this is mostly symmetric i.e. a good property to reduce bias in regression algorithms, there exists a small positive skew in the distribution which is shown by the long tail to the direction of 450-500 g/km. This tail indicates a minority category of vehicles that has high emissions and this implies that although this model will probably work well on average vehicles, it may need the few attenuating properties of non-linear models to effectively predict the outlier values on heavy-duty or inefficient engines.

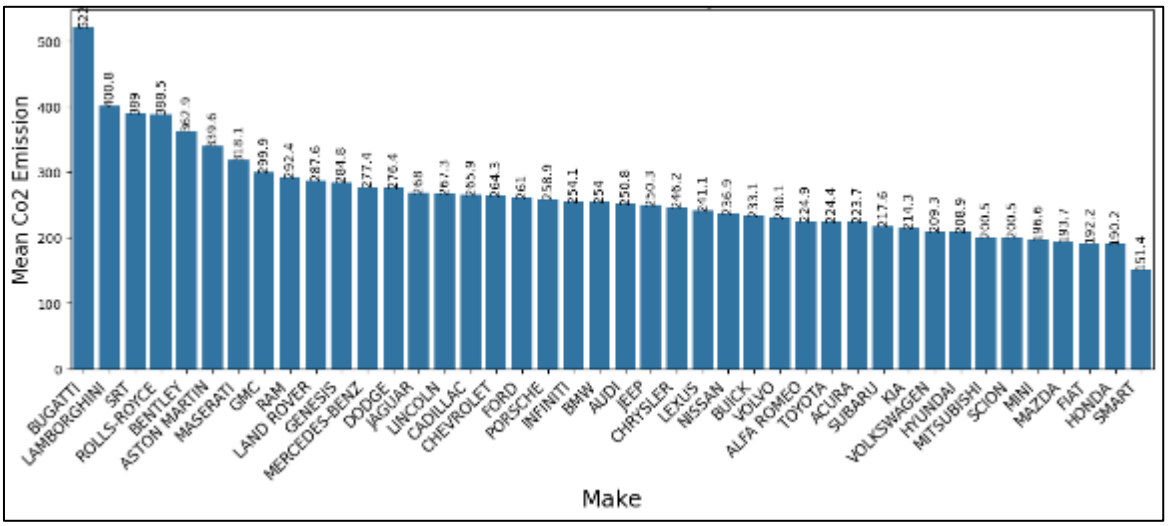

Figure 5 Co_2 Emission by Make**.**

Figure 5 is a comparative analysis of the mean CO_2 emissions broken down by vehicle manufacture (Make). The bar chart shows severe heterogeneity in the carbon output of the various automotive industries and that is mostly influenced by the market segment and engineering philosophy of the respective brands. The leading automobile companies in the high-performance and ultra-luxury category, namely Bugatti, Lamborghini, and SRT, occupy the highest point of the emissions, and Bugatti has an astonishing average at around 522 g/km. This can be owed to the fact that they are powered by massive displacing multi-cylinder engines that are not efficient. The economy-oriented brands Smart, Honda, and Fiat, in their turn, occupy the lower range, ranging below 200 g/km. The fact that the difference between these groups is substantial (as low as 151.4 g/km of Smart down to over 500 g/km of Bugatti) highlights the high predictive power of the "Make" feature (and the related engine features) to predict the environmental impact of a vehicle.

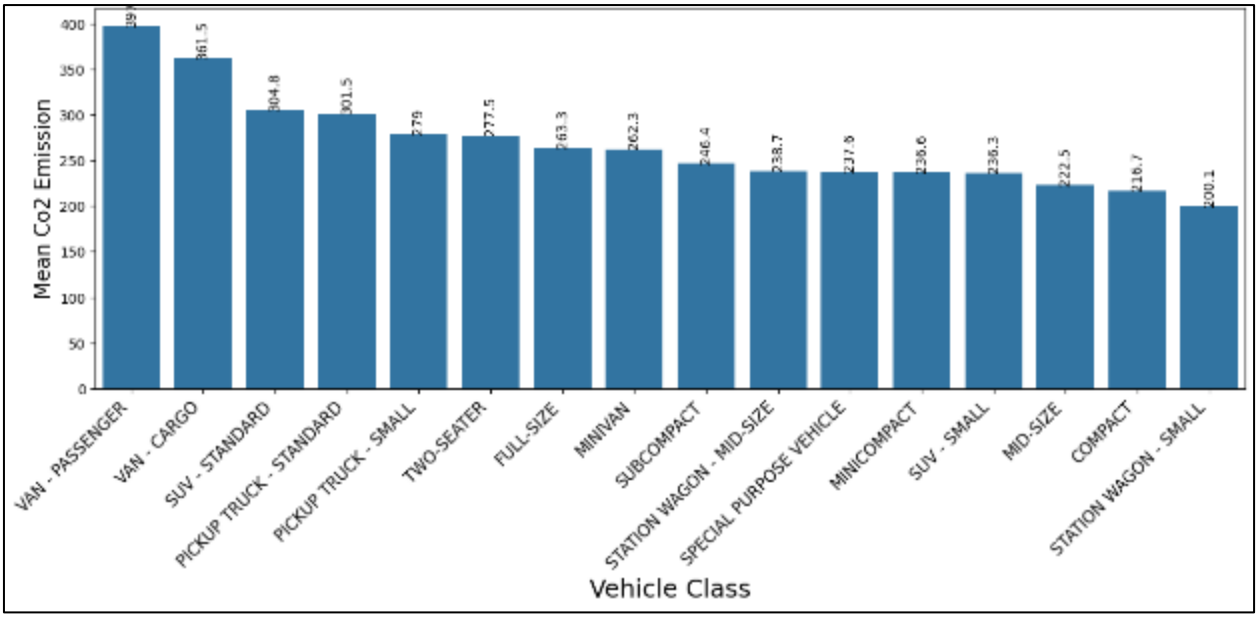

Figure 6 Co_2 Emission by Vehicle Class.

Figure 6 shows the average emission of CO_2 of the various classes of vehicles showing a very considerable influence of utility and size of the vehicle on its environmental performance. It is also evident that passenger vans and cargo vans represent the most emitting (397.2 g/km and 361.5 g/km, respectively) because they have heavier chassis and are not as aerodynamic, which requires more powerful engines. Smaller passenger types, such as Compact cars and Small Station Wagons, on the other hand, have the lowest carbon footprint and the latter has an average of about 200.1 g/km, which is almost half the amount of the heaviest type. This clear division into utility classes (Vans, Pickups) and commuter classes (Compact, Mid-size) supports the fact that "Vehicle Class" is a very strong high-level discriminator in terms of predicting emissions.

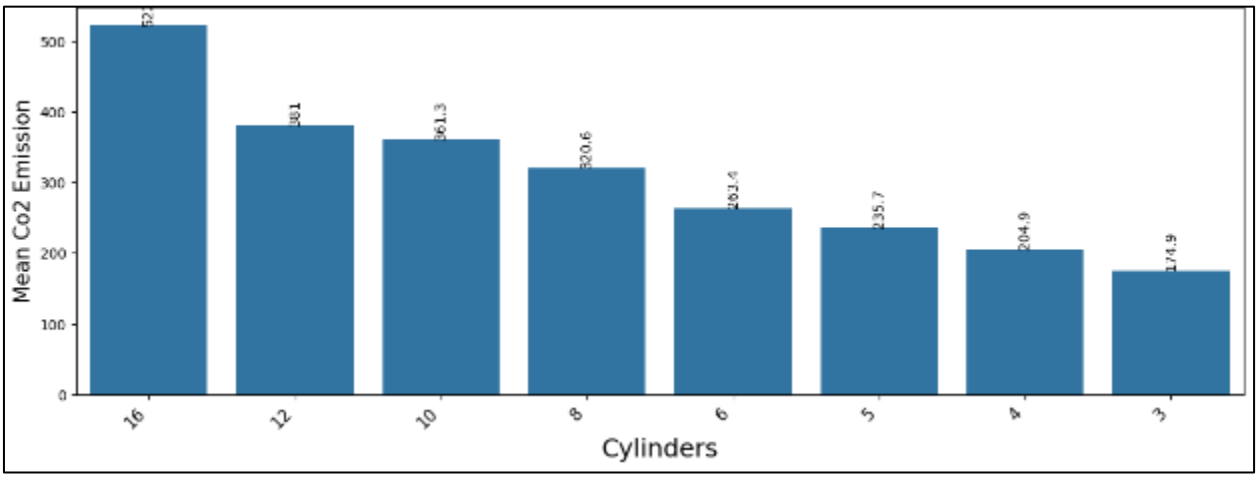

Figure 7 Co_2 Emission by Cylinder.

Figure 7 illustrates that the number of engine cylinders has a direct and monotonic relationship with the mean CO_2 emissions. The bar chart indicates a positive correlation in the whole: the higher the number of cylinders, the higher the level of average emissions. The top of the chart is occupied by vehicles with huge engines with 16-cylinders which emit about 522 g/km which is quite high considering the high fuel consumption of the large-displacement engines. Conversely, 3-cylinder engine prevalent in efficiency-driven modern cars has the lowest average emissions of 174.9 g/km- about a third of the 16-cylinders. This step by step increase (4 cylinders at 204.9 g/km to 8 cylinders at 320.6 g/km) has proven the fact that the number of cylinders is not only a structural element but one of the main physical determinants of fuel consumption and, accordingly, carbon emission. The high score our Voting-Based Ensemble Feature Selection framework attaches to the "Cylinders" feature warrants this high importance score.

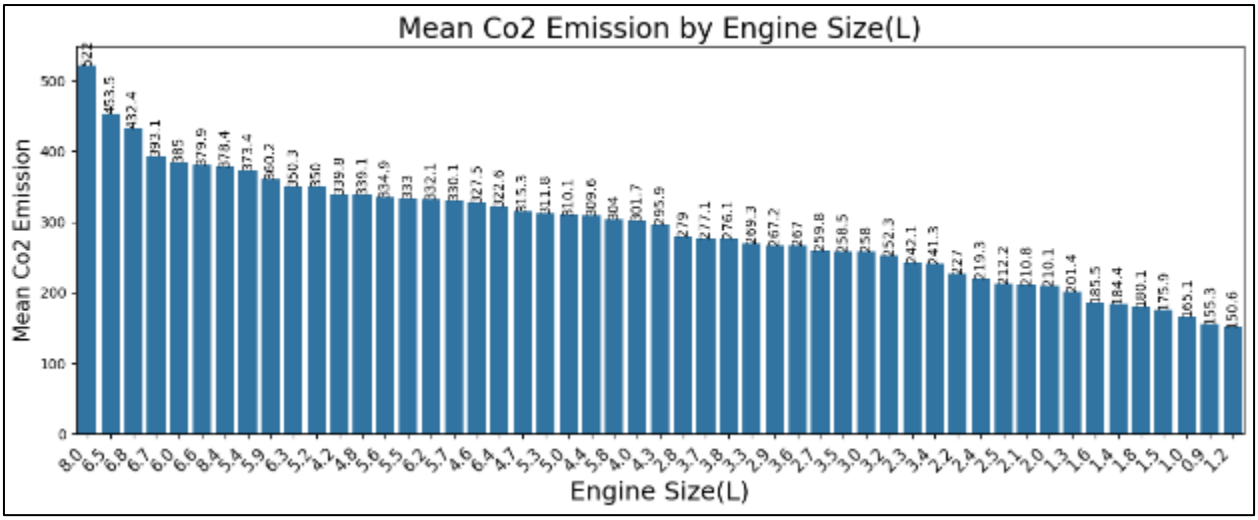

Figure 8 Co_2 Emission by Engine Size

Figure 8 also explains the mechanical determinants of pollution by making a plot of the average CO_2 emissions against Engine Size (in Liters). The statistics show a very high positive relationship of displacement with carbon production, which is similar to that of the number of cylinders. The 8.0L engine configuration is pegged at the top end of the chart with the highest mean emissions of 522 G/km, showing clearly the high fuel consumption of large-displacement performance engines. A steady decline in emissions is seen as the engine size is reduced, and the lowest level of emission of about 150.6 g/km is recorded in the 1.2L engine category- about 1/3rd of the volume of emissions produced by the 8.0L engine. This is a strong linear relationship that Engine Size is a superior continuous predictor of the model, which underpins granular information to supplement the discrete Cylinders attribute.

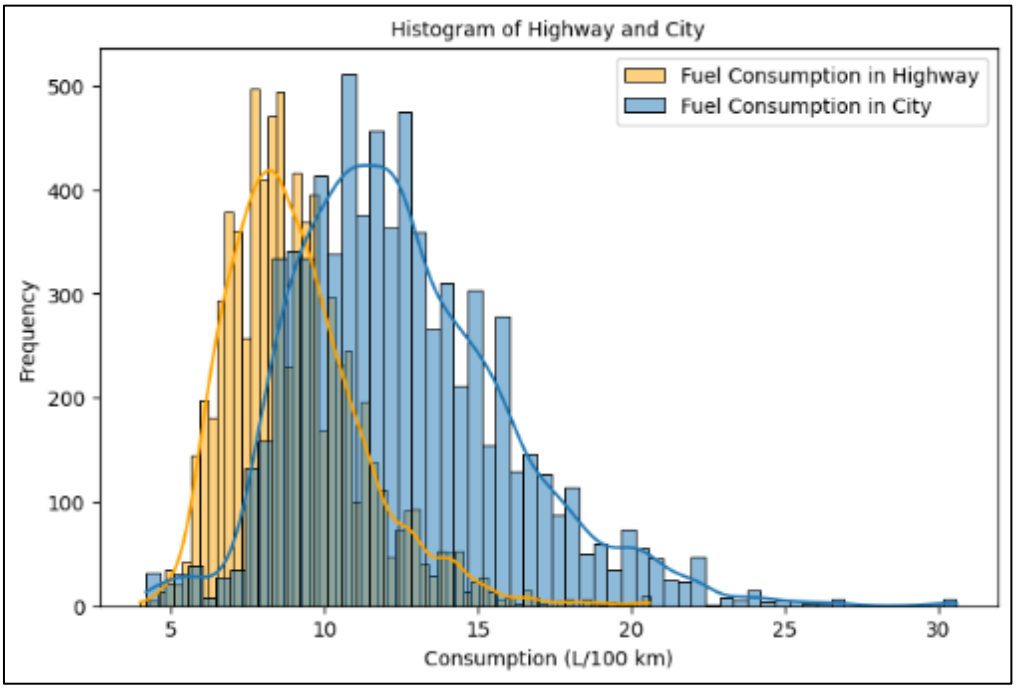

Figure 9 Fuel Consumption in Highway vs City

Figure 9 gives a comparative histogram of the rates of fuel consumption (l/100 km) differentiated by the driving cycle: City (Blue) and Highway (Orange). The visual inspection shows an expected and apparent discrepancy in efficiency profiles. Highway consumption distribution is significantly left skewed with highest value in the lower part of roughly 8-9 L/100 km indicating the high efficiency of highway driving due to frequent speeds and less braking frequences. On the other hand, the location of the City consumption is right-shifted and has a greater mean mode of about 10-12 L / 100 km and a much longer tail that covers up to 20 L/100 km. Such broader distribution and mean within the City cycle

highlight the effects of stop and go traffic and idling to fuel consumption. These are separated into separate input features by the model, enabling it to better explain the varying load on the engine under various operating conditions, and thus improving its prediction of emissions.

The results of a multi-variate scatter plot representing the relationship between Fuel Consumption (Combined) and CO_2 Emissions, color coded to distinguish between Vehicle Classes, are shown in Figure 10. The visualization presentation would also indicate an almost perfect positive line of relationship between the fuel consumption and emissions, which is a physical fact that carbon production is a direct proportion of fuel burns.

More importantly, the plot shows clear stratification in terms of utility of vehicles. Smaller, efficiency-oriented classes like Compact, Subcompact, and Mid-Size vehicles (presented in pinks and oranges) mainly tend to concentrate at the lower-left quadrant and they have a low consumption rate of less than 10 L/100km and low emissions of less than 250g/km. Conversely, classes with heavy-utility such as Passenger Vans and Standard Pickup Trucks (which are painted in blues) dominate the upper-right quadrant, with consumption of over 15 L/100km and emissions of over 350g/km. The parallel, isolated, linear bands indicate that the relationship fuel-emission is a universal one, though it is slightly disturbed by the type of fuel (e.g. Ethanol against Gasoline) or the engine calibration due to specific vehicle class. This confirms the fact that Fuel Consumption Combined is the most powerful predictor in the dataset whilst Vehicle Class offers relevant contextual variance.

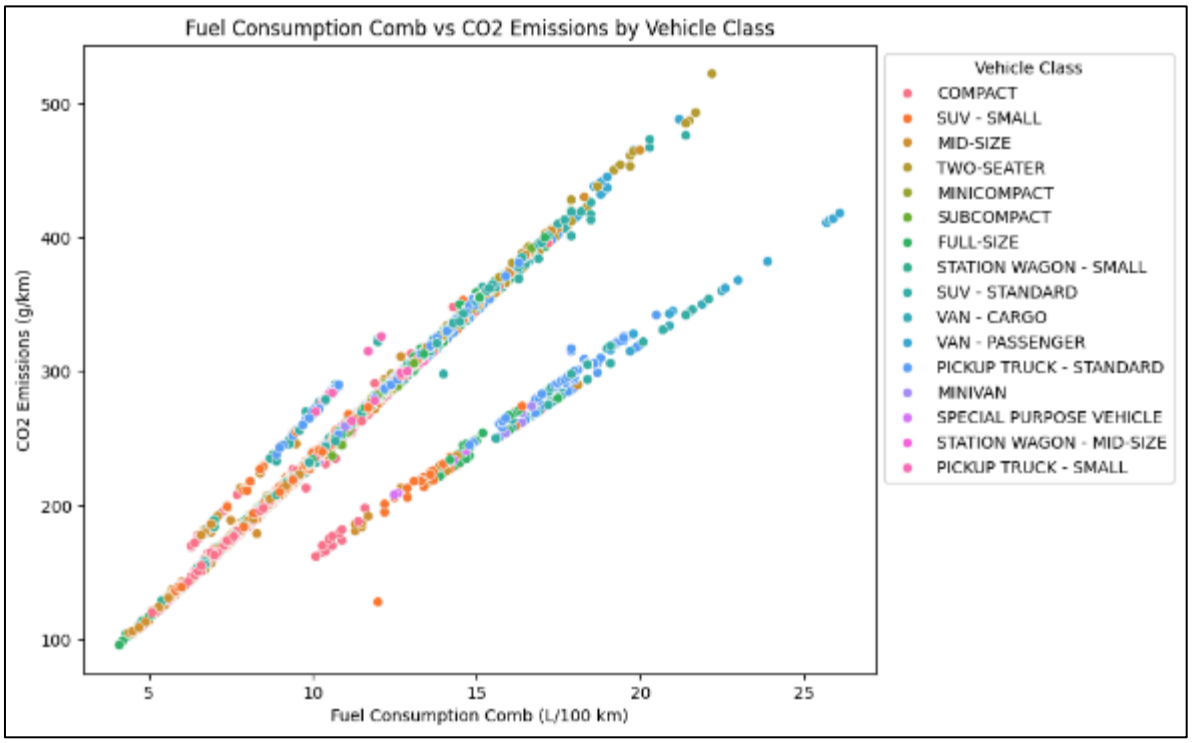

Figure 10 Fuel Consumption vs **CO_2** Emission by Vehicle Class

4.2. Data Preprocessing

A multi-step preprocessing phase was conducted to prepare the data for machine learning.

4.2.1. Categorical Data Encoding

All features in the category and object type (Make, Model, Vehicle Class, Transmission, Fuel Type) were transformed into numbers with the help of Scikit-learn LabelEncoder. This procedure is necessary to supply the machine learning algorithms with the data. The dataset is demonstrated in Figure 11, which is processed with the help of data encoding.

make	model	vehicle_class	engine_size	cylinders	transmission	fuel_type	fuel_cons_city	fuel_cons_hwy	fuel_cons_comb	fuel_cons_comb_mpg	co2
15	1648	6	27	3	3	1	19.3	14.3	17.1	17	284
3	248	2	8	1	17	4	10.8	7.3	9.2	31	211
24	1009	12	29	4	15	4	15.7	11.7	13.9	20	320
21	915	1	8	1	16	3	9.7	7.0	8.3	34	197
9	1653	6	34	4	3	1	20.7	14.4	17.9	16	315

Figure 11 Dataset After Encoding

4.2.2. Creating New Feature

To capture the relationship between urban and highway fuel efficiency, a new feature, city_hwy_diff, was engineered by calculating the difference between fuel_cons_city and fuel_cons_hwy.

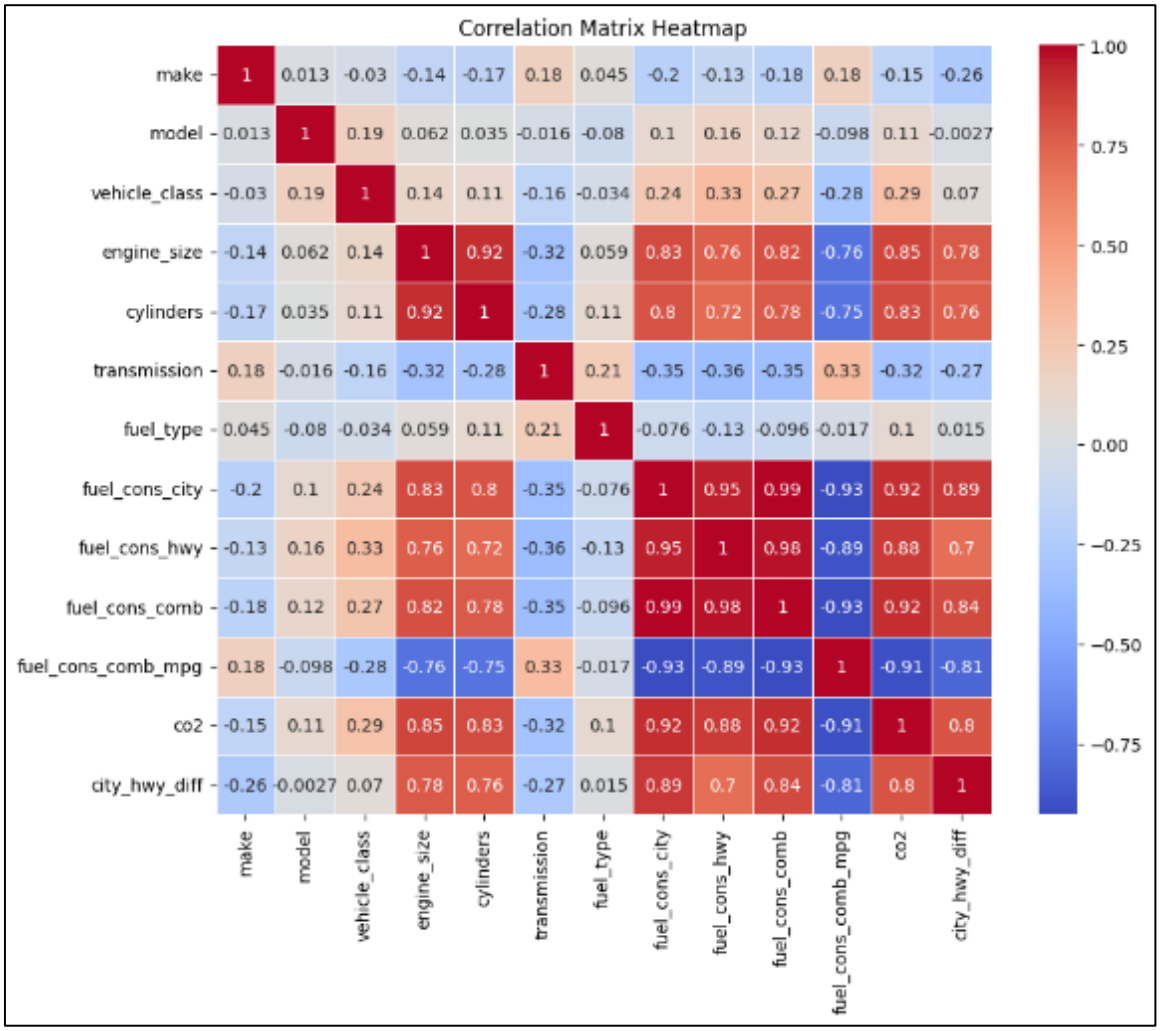

Figure 12 Correlation Heatmap After Creating City Highway Diff

The feature engineering stage is followed by a correlation heatmap, which is shown in Figure 12. The additional variable (CityHighwayDiff) has a positive correlation with the target variable, CO_2 emissions, that is significantly high (0.80). This high correlation means that the difference in efficiency between the urban and highway performance is a predictor of the total carbon footprint of a vehicle; those with a greater gap between the urban and highway performance are heavier polluters. Moreover, the CityHighwayDiff has strong correlations with the physical engine features, including Engine Size and Cylinders. This indicates that more powerful engines have a bigger efficiency penalty in urban stop and go traffic than in highway running, a unique behavioral pattern which this feature manages to capture well on behalf of the model.

3.2.3 Data Splitting

The dataset was first separated into features (X) and the target variable (y). To determine an optimal split, a preliminary test was run using a simple Linear Regression model on various test-to-train ratios (10%, 20%, 25%, 30%). A test size of 0.20 (20%) yielded the highest R-squared score, and was therefore adopted for the main experiment. The data was split into training (80%) and testing (20%) sets. Table 1 shows the R^2 value found in each split.

Table 1 Result of Different Split of The Dataset

Ratio	Model	R^2
70:30	Linear Regression	0.9108
75:25	Linear Regression	0.9096
80:20	Linear Regression	0.9117
90:10	Linear Regression	0.9086

4.2.3. Feature Scaling

To ensure that all features contribute equally to model performance and to normalize their different scales, the StandardScaler from Scikit-learn was employed. The scaler was fit only on the training data and then used to transform both the training and testing sets, preventing data leakage.

4.3. Voting-Based Feature Selection

One of the new additions of the research is that voting-based ensemble feature selection method is assumed. Multicollinearity and irrelevant noise are common in datasets in the field of vehicle emission prediction. Applying one feature selection method may give biased results; e.g., the linear methods may fail to detect any non-linear dependencies, and tree-based measures of feature importance may be biased towards high-cardinality variables.

In order to alleviate these shortcomings and guarantee the choice of a strong, generalizable feature subset, we combine the results of five different, complementary feature selection algorithms in our framework. It is a "Democratic" or "Majority Voting" system, based on the idea of consensus: a feature can only be regarded as important when it is independently confirmed by a number of algorithms based on various mathematical assumptions (statistical correlation, information theory, and impurity reduction).

The five constituent techniques that are applied within this ensemble are explained below:

4.3.1. Random Forest Importance (Gini Impurity)

The algorithm of the Random Forest gives an automatic value of the feature importance depending on the decrease in impurity when building the decision trees. In the current research, we have used the featureimportances attribute of a trained RandomForestRegressor.

- Mechanism**:** Every data tree in the forest will contain an internal node that divides the data between two labeled by a particular feature to optimize the reduction in the variance (in the case of regression) or Gini impurity. The significance of a feature is computed as the average reduction in impurity that is made on each and every tree in the forest.

- Justification**:** This is specifically a very effective method in capturing non-linear interactions among vehicle specifications. In contrast with linear models, Random Forest does not presuppose a linear relation between features and the target variable (CO_2 emissions), which allows it to be much more effective in detecting non-monotonic, complex features that dominate the emissions.

4.3.2. SelectKBest (Univariate Statistical Selection)

In order to get strong linear dependency, we used SelectKBest algorithm with the F-regression scoring function. It is a univariate filter based on the selection of features and it compares each individual feature against the target variable.

- Mechanism**:** This technique calculates F-score of every feature, which is used to test the null hypothesis that the coefficient of a certain feature in a linear regression model is zero. It is a proper measure of linear dependence of two random variables.

- Justification**:** The ensemble needs a sanity check, which is provided through this method, though it is rather simple. It makes sure that features that have high direct statistical correlations with CO_2 emissions are retained. It is used to offset the more complicated model-based selectors by ranking statistically significant linear predictors.

4.3.3. Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a wrapper technique which follows a backward selection scheme to determine the most optimal set of features. As our base estimator to the RFE process, we used a RandomForestRegressor.

- Mechanism**:** RFE is initiated by training the estimator using the full range of original features and finding the importance of each. Then it removes the least significant feature of the set and re-trains the model. It is an iterative process that is continued until the required quantity of features is left.

- Justification**:** RFE takes into account feature interaction unlike univariate methods (such as SelectKBest) that consider features individually. It eliminates the weakest contributors in an iterative fashion to take into account feature dependencies and to makes the remaining subset cooperate to form the least error in prediction.

4.3.4. Lasso (Least Absolute Shrinkage and Selection Operator)

We added L1-regularization through the Least Absolute Shrinkage and Selection Operator with Cross-Validation (LassoCV) algorithm. Lasso is incorporated to be used as a selection procedure.

- Mechanism**:** Lasso reduces the remainder sum of squares under the sum of the absolute values of the coefficients is less than a constant. This penalty term compels those coefficients of factors of lower significance or redundancy to contract to the point of zero.

- Justification**:** This approach is extraordinarily strong in the case of multicollinearity, which is likely to occur in vehicle datasets, such as the interrelations of such features as Fuel Consumption City and Fuel Consumption Highway. Lasso is able to select features automatically by reducing coefficients to zero, only one variable is left in a set of highly correlated predictors and simplifies the model.

4.3.5. Mutual Information Regression

We used Mutual Information (MI) regression to derive arbitrary dependencies because they are not necessarily linear or monotonic.

- Mechanism**:** The Mutual Information is a measurement of the amount of information that can be gained on one random variable (CO_2 emissions) under the condition that another variable (a vehicle feature) is observed. It measures the decrease in uncertainty of the variable of interest provided the knowledge of the feature. The use of K-nearest neighbor distances is to estimate entropy using continuous data.

- Justification**:** MI is not like F-tests or correlation coefficients since it identifies any type of relationship, including periodic or complex non-linear relationships. Its representation in the voting ensemble is what makes sure that the strong predictive power features, which can be overlooked by linear filters such as Lasso or SelectKBest are included.

4.3.6. Consensus Voting and Final Selection

The final stage of the framework involved a hard voting mechanism. Each of the five methods described above produced a ranking of the top candidate features. A binary voting matrix was constructed where:

$$V_{i,j} = 1 \qquad \text{(i)}$$

if algorithm j selected feature i, and 0 otherwise.

$$C_i = \sum_{j=1}^{5} V_{i,j} \quad \text{(ii)}$$

Features were ranked based on C_i, and a consensus threshold was applied. The analysis resulted in the selection of the top six features, which demonstrated the highest stability across varying selection methodologies.

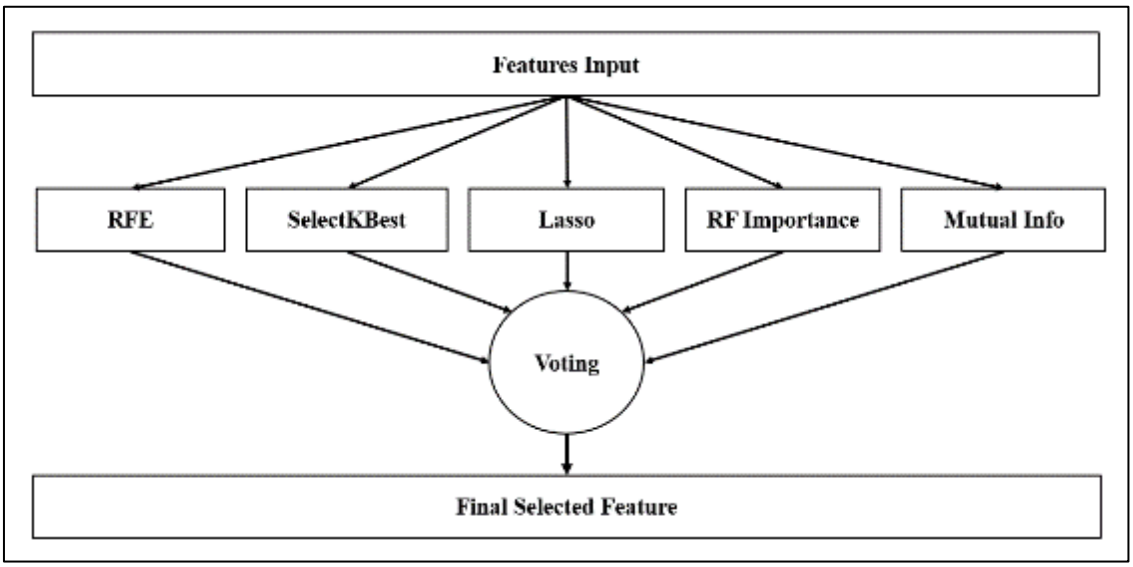

Figure 13 Architecture of Proposed Voting Based Feature Selection**.**

Figure 13 illustrates the architectural process of the proposed Voting-Based Ensemble Feature Selection (VBE-FS) framework. The diagram represents a parallel processing pipeline with the original high-dimensional Features Input being shared among five heterogeneous feature selection algorithms: Recursive Feature Elimination (RFE), SelectKBest (F-regression), Lasso (L1 Regularization), Random Forest Importance and Mutual Information. This parallel implementation makes sure that the feature space will be assessed in a variety of statistical and model-fashion ways. The best rankings of these algorithms coincide at the middle Voting node, which is the mechanism of aggregation of the ensemble. In this case, the majority rule is used to override noise and instability, and what remains is the Final Selected Feature subset, the solid and optimized set of predictors that will be the basis of the following predictive modeling step.

4.4. Machine Learning Models

The selected feature subset was used to train and evaluate four distinct machine learning regression models.

4.4.1. Random Forest (RF) Regressor

An ensemble model that operates by constructing a multitude of decision trees at training time. The model was configured with optimized hyperparameters: n_estimators=750, max_depth=9, and max_features=5.

4.4.2. K-Neighbors (KNN) Regressor

A non-parametric, instance-based learning algorithm that predicts a value based on the average of its 'k' nearest neighbors. The model was set with an optimal n_neighbors=3.

4.4.3. Support Vector Regression (SVR)

A regression variant of Support Vector Machines that finds a hyperplane to fit the data, while allowing for a specified margin of error. The model was tuned with a C parameter of 50.

4.4.4. Extra Trees (ET) Regressor

An ensemble method, similar to Random Forest, that introduces additional randomness by splitting nodes on random subsets of features and thresholds. It was configured with n_estimators=100 and random_state=42.

4.5. Deep Learning Models

While not implemented in the final comparison for this study, deep learning models represent a significant area of research for this problem, as noted in the literature [2, 3, 4]. The following models are often considered for their ability to capture complex, non-linear patterns.

4.5.1. Convolutional Neural Network (CNN)

Image processing 1D CNNs can be applied to tabular and sequence data though they are mostly associated with image processing. In this issue, the input data (the 6 chosen features) may be reshaped to a 1D vector. The CNN convolutional layers would introduce filters over these features and would thus be really efficient in learning local interactions of feature.

Figure 14 illustrates the schematic architecture of the proposed Convolutional Neural Network (CNN) which is used to learn the non-linear patterns that are more complex in the feature space. The design is sequential feed forward with the first stage being the Input Layer which receives the pre-processed feature vectors. This is then succeeded by a Convolutional Layer which proves useful in isolating high-level, local feature dependencies- a trick commonly used to eliminate noise and accentuate salient patterns in the input data. The resulting feature maps are flattened and processed by a deep regression head comprising of three straight Fully Connected (Dense) Layers. The layers are used to project the latent features into a higher dimensional space so that the network learns the complex relationships and eventually converges at the Output Layer which provides the final continuous prediction of the CO_2 emissions.

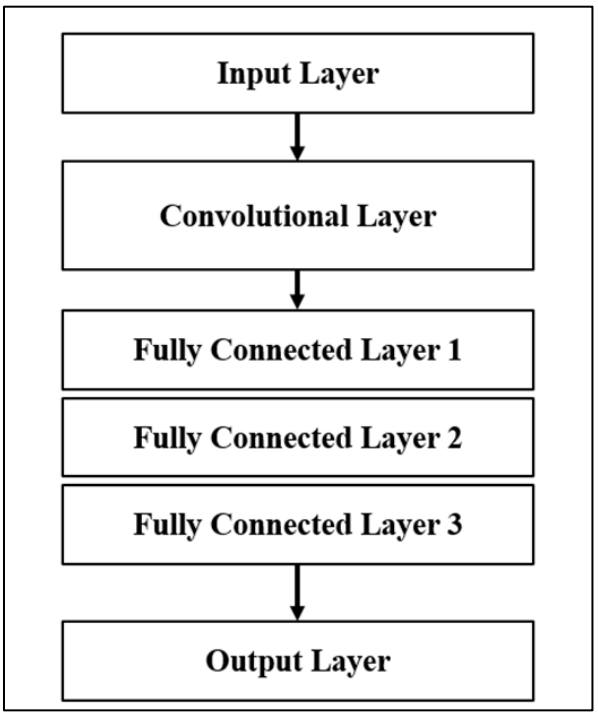

Figure 14 Architecture of Proposed CNN Model.

4.5.2. Long Short-Term Memory (LSTM)

LSTMs are recurrent neural network (RNNs) that are used to process long-term dependencies in sequence data. Although such data is not necessarily time-series, LSTMs and their variants (such as BiLSTM) are becoming the potent non-linear regressors [2]. Through the feature set being treated as a sequence, an LSTM is able to capture the complex relationships and higher-order interactions between inputs, and generate one and accurate prediction of CO_2.

4.6. Evaluation Metrics

To quantitatively assess and compare the performance of the trained models, a dedicated evaluation function was used to calculate four standard regression metrics for both the training and testing datasets.

4.6.1. R-squared (R^2)

The coefficient of determination is abbreviated as R^2 and it is the ratio of the amount of the variance in the dependent variable (CO_2 emissions) which can be predicted by the independent variables. It gives an indication of the goodness with which the model reproduces the actual results with a value of 1.0 being a perfect fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad \text{(iii)}$$

Where,
y_i is the actual value.
$\hat{y}_i$ is the predicted value.
$\bar{y}_i$ is the mean of the actual values.
n is the total number of observations.

4.6.2. Mean Absolute Error (MAE)

MAE averages the predicated and actual differences across the values. It gives a linear score with all individual differences being weighted equally to give an easy interpretation of the average error magnitude in the same units as the target variable (g/km).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{(iv)}$$

4.6.3. Mean Squared Error (MSE)

MSE is a measure of the mean squared error in the predicted and actual values. MSE gives more weight to large errors than to small ones and therefore makes it a useful measure in determining models that can give large deviations at times.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{(v)}$$

4.6.4. Root Mean Squared Error (RMSE)

The square root of MSE is the RMSE. It is also convenient since it can convert the error measure back to a comparable unit as the target variable (g/km) so that it is easier to compare with the MAE. RMSE is outlier sensitive and also gives an understanding of the standard deviation of the error in prediction (residuals).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{(vi)}$$

5. Results and Discussion

This section presents the results of the experimental analysis, beginning with the outcome of the voting-based feature selection process and followed by a detailed performance comparison of the four machine learning models.

5.1. Feature Selection Results

According to the methodology, the voting-based ensemble of five various feature selection methods was used to select the most important predictors of CO_2 emissions. Table 2 displays the final number of votes in each feature. Six out of the

top features, including fuel_cons_comb, fuel_cons_city, engine_size, fuel_cons_hwy, fuel_type, and cylinders were chosen to train the model. The findings are consistent with the accepted principles of automotive engineering: fuel_cons_comb (combined fuel consumption) was unanimously voted, which is the most important predictor, which is reasonable since CO_2 emissions are a direct byproduct of fuel combustion. The remaining fuel consumption measures (city and hwy) and the main engine variables (engine_size, cylinders) were also found to be very relevant. The addition of the fuel_type indicates the fact that various fuel ratios influence the level of emissions in distinct ways.

Table 2 Voting-Based Feature Selection Results

Feature	SelectKBest	Lasso	RFE	RF Importance	Mutual Info	Total Votes
fuel_cons_comb	1	1	1	1	1	5
fuel_cons_city	1	1	1	1	1	5
engine_size	1	1	1	1	1	5
fuel_cons_hwy	1	0	1	1	1	4
fuel_type	0	1	1	1	0	3
cylinders	1	0	0	1	0	3
city_hwy_diff	1	0	0	0	1	2
model	0	1	0	0	1	2
make	0	1	0	0	0	1
vehicle_class	0	1	0	0	0	1

5.2. Model Performance Comparison

The number of characteristics which were selected was 6, and these characteristics formed the basis of six different models. R-squared (R2), Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were used to measure the performance of each model on the test dataset. All of the metric comparisons of the models are explained in detail in Table 3. The model employed in every case exhibited great capacity to predict CO_2 emissions with the R2 value of above 0.995. A comparison of the R 2 values and the Root Mean Squared Errors (RMSE) of both models demonstrates that the Random Forest (RF) Regressor was the most satisfactory model in general. RF Regressor was the best model in terms of the highest R2 (0.9976) and the lowest RMSE (2.847 g/km) of the test data. The Support Vector Regressor (SVR) and the Convolutional Neural Network (CNN) were almost the same; they resulted in the R2 of 0.9972 and the RMSE of 3.052 and 3.115, respectively. Long Short-Term Memory (LSTM) model was a good performer (R2 = 0.9970). As much as the Extra Trees (ET) Regressor performed very well (R2 = 0.9967), it also ranked as the highest RMSE of the best performing models, yet, it ranked second lowest in the Mean Absolute Error (MAE) (1.642). K-Neighbors (KNN) Regressor did not cope with the rest in the group in terms of R2 (0.9958) and RMSE (3.759).

Table 3 Comparative Performance of ML And Dl Models.

Model	R-squared (R^2)	MAE	MSE	RMSE
Random Forest	99.76	2.087	8.107	2.847
K-Neighbors	99.58	2.125	14.134	3.759
Support Vector Regressor	99.72	2.123	9.315	3.052
Extra Tree Regressor	99.67	1.642	11.011	3.318
CNN	99.72	2.269	9.706	3.115
LSTM	99.70	2.272	10.367	3.219

The strong performance of the Random Forest model is further visualized in the Figure 15, which shows a near-perfect linear alignment between the actual and predicted values.

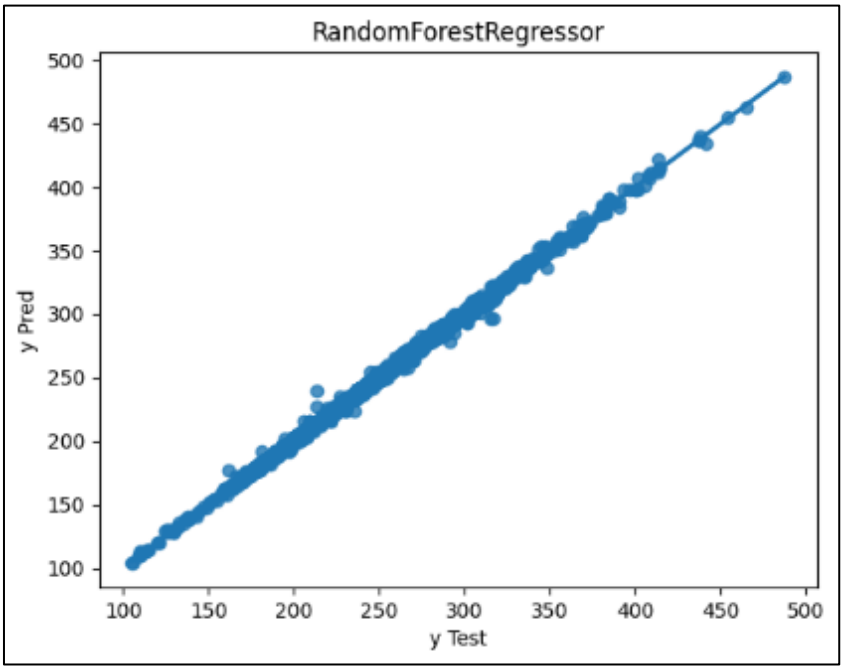

Figure 15 Prediction of Random Forest Model**.**

6. Conclusion

A new, robust machine learning method was developed in this research, and effectively tested against a large, diverse database from across Canada, for the prediction of the CO_2 emission levels of vehicles. The major innovation in this research was the use of a novel voting-based feature selection ensemble method, where five different methods of selecting features (SelectKBest, Lasso, Recursive Feature Elimination, Random Forest importance, and Mutual Information) were used to vote on the most important variables to select when building the predictive model.

It was demonstrated that the CO_2 emission level of a vehicle could be predicted to an exceptionally high degree of accuracy by utilizing only six significant features: fuel_cons_comb, fuel_cons_city, engine_size, fuel_cons_hwy, fuel_type, and cylinders. In addition, among the models tested it was found that the Random Forest Regressor was the best performing model. The Random Forest Regressor had the highest test R-squared value (R^2) of 0.9976 and the lowest Root Mean Squared Error (RMSE) of 2.861 g/km, which demonstrates that the Random Forest Regressor has a strong ability to generalize to unseen data and produce accurate predictions. Overall, these findings support the idea that ensemble methods, both for feature selection and for regression, can be very powerful for problems like this one. The study demonstrates that advanced machine learning models can generate highly reliable CO_2 emission forecasts, enabling planners and policymakers to make informed decisions that reduce urban transport emissions. By integrating these predictive insights into planning and governance, cities can accelerate progress toward sustainable, low-carbon and resilient urban futures.

Future Scope

Although the proposed model has very high accuracy, there are many potential areas to study in the future. The first area to be studied is the validation of the model's robustness through its use on various additional data sets. This would allow the model to be used on more real-world driving scenarios versus just those created in laboratory testing environments. The second area to be studied is the application of Explainable AI (XAI) to the model using a technique like SHAP (SHapley Additive exPlanations). This will allow the researchers to have a better insight into the decision-making process of the model and therefore a greater ability to understand the contribution of each feature to the final model prediction.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Sethi M. Climate Change and Urban Settlements: A Spatial Perspective of Carbon Footprint and Beyond. 1st ed. London: Routledge; 2017.

[2] Sidana S. Grid Search Optimized Machine Learning based Modeling of CO_2 Emissions Prediction from Cars for Sustainable Environment. Int J Curr Sci Res Rev. 2024;7(9):37.

[3] Chadha AS, Shinde Y, Sharma N, De PK. Predicting CO_2 Emissions by Vehicles Using Machine Learning. Singapore: Springer; 2022. p. 197-207.

[4] Al-Nefaie AH, Aldhyani THH. Predicting CO_2 Emissions from Traffic Vehicles for Sustainable and Smart Environment Using a Deep Learning Model. Sustainability. 2023;15(9):7615.

[5] Nassef AM, Olabi AG, Rezk H, Abdelkareem MA. Application of Artificial Intelligence to Predict CO_2 Emissions: Critical Step towards Sustainable Environment. Sustainability. 2023;15(9):7648.

[6] Natarajan Y, Wadhwa G, Sri Preethaa KR, Paul A. Forecasting Carbon Dioxide Emissions of Light-Duty Vehicles with Different Machine Learning Algorithms. Electronics. 2023;12(10):2288.

[7] Chukwunonso BP, AL-Wesabi I, Shixiang L, AlSharabi K, Al-Shamma'a AA, Farh HMH, Saeed F, Kandil T, Al-Shaalan AM. Predicting carbon dioxide emissions in the United States of America using machine learning algorithms. Environ Sci Pollut Res. 2024;31(23):33685-707.

[8] Yeasmin S, Syed SNJ, Shmais LA, Dubayyan RA. Artificial Intelligence-based CO_2 Emission Predictive Analysis System. 2020 International Conference on Artificial Intelligence & Modern Assistive Technology (ICAIMAT). 2020:1-6.

[9] Li X, Zhang X. A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China. Environ Sci Pollut Res. 2023;30(55):117485-502.

[10] Shah S, Thakar S, Jain K, Shah B, Dhage S. A Comparative Study of Machine Learning and Deep Learning Techniques for Prediction of CO_2 Emission in Cars [Internet]. arXiv; 2022. Available from: http://arxiv.org/abs/2211.08268

[11] Li S, Tong Z, Haroon M. Estimation of transport CO_2 emissions using machine learning algorithm. Transp Res Part D Transp Environ. 2024;133:104276.

[12] Author C, Gurcan F. Forecasting CO_2 emissions of fuel vehicles for an ecological world using ensemble learning, machine learning, and deep learning models. [Publication details unavailable]; n.d.

[13] Sayed GI, Hassanien AE. Prediction of CO_2 Emission in Cars Using Machine Learning Algorithms. 2023. p. 85-97.

[14] Zhang B, Ling L, Zeng L, Hu H, Zhang D. Multi-step prediction of carbon emissions based on a secondary decomposition framework coupled with stacking ensemble strategy. Environ Sci Pollut Res. 2023;30(27):71063-87.

[15] Akhy SA, Mia MB, Mustafa S, Chakraborti NR, Krishnachalitha KC, Rabbany G. A Comprehensive Study on Ensemble Feature Selection Techniques for Classification. 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom). 2024:1319-24.

[16] Cai J, Luo J, Liang C, ShengYang. A Novel Information Theory-Based Ensemble Feature Selection Framework for High-Dimensional Microarray Data. Int J Performability Eng. 2017;13(5).

[17] Drotar P, Gazda M, Gazda J. Heterogeneous ensemble feature selection based on weighted Borda count. 2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE). 2017:1-4.

[18] Kumari S, Singh SK. Machine learning-based time series models for effective CO_2 emission prediction in India. Environ Sci Pollut Res. 2022;30(55):116601-16.

[19] Sayed GI, Hassanien AE. Prediction of CO_2 Emission in Cars Using Machine Learning Algorithms. 2023. p. 85-97.

[20] Podder D. CO_2 emission by vehicles [Internet]. Kaggle; n.d. Available from: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles

[21] Wang C, Li M, Yan J. Forecasting carbon dioxide emissions: application of a novel two-stage procedure based on machine learning models. J Water Clim Change. 2023;14(2):477-93.

[22] Li Y, Sun Y. Modeling and predicting city-level CO_2 emissions using open access data and machine learning. Environ Sci Pollut Res. 2021;28(15):19260-71.