Comparative Evaluation of Machine Learning Regression Techniques for Predicting CO_2 Emissions in Light-Duty Vehicles

Adarsh Lal Anilal
University of Texas at Arlington
701 S Nedderman Drive
Arlington, TX 76019

Aera K. Leboulluec
University of Texas at Arlington
701 S Nedderman Drive
Arlington, TX 76019

ABSTRACT
The transportation sector is a significant contributor to global greenhouse gas emissions, with carbon dioxide (CO_2) from vehicles being a primary driver of climate change. Accurate prediction of vehicle CO_2 emissions based on engine and fuel economy characteristics is essential for regulatory compliance, environmental policy, and automotive design optimization. In this research, the EPA Model Year 2026 Fuel Economy Guide dataset, comprising 652 vehicle records with 15 attributes including engine displacement, cylinder count, fuel economy ratings, and CO_2 emission measurements, is utilized to build and evaluate machine learning regression models. Three supervised learning algorithms are implemented: Random Forest Regression, K-Nearest Neighbors (KNN) Regression, and Support Vector Regression (SVR). Each model is trained on 80% of the data and tested on the remaining 20%, with performance evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and Adjusted R-squared metrics. Additionally, 5-fold cross-validation is employed to assess model robustness across different data partitions. The results demonstrate that SVR with an RBF kernel achieves the highest predictive accuracy with an R^2 of 0.9988 and MAE of 2.16 grams per mile, followed closely by Random Forest (R^2 = 0.9978). This study provides a framework for applying machine learning techniques to vehicle emission prediction and highlights the potential of data-driven approaches for supporting environmental sustainability initiatives.

General Terms
Machine Learning, Regression, Environmental Computing, Algorithms.

Keywords
CO_2 Emissions Prediction, Random Forest Regression, K-Nearest Neighbors, Support Vector Regression, EPA Fuel Economy.

1. INTRODUCTION
Climate change driven by greenhouse gas emissions is one of the most pressing environmental challenges of the 21st century. The transportation sector is responsible for approximately 16% of global carbon dioxide (CO_2) emissions, making it a critical area for intervention and monitoring [1]. As governments worldwide implement increasingly stringent emission regulations, the ability to accurately predict vehicle CO_2 emissions based on measurable vehicle characteristics has become essential for automotive manufacturers, regulatory agencies, and environmental policymakers.

Traditional methods of measuring vehicle emissions rely on physical testing procedures such as the Worldwide Harmonized Light Vehicles Test Procedure (WLTP) and dynamometer-based laboratory tests. While these methods provide accurate measurements, they are time-consuming, expensive, and impractical for large-scale prediction across diverse vehicle fleets [2]. Machine learning offers a data-driven alternative that can predict emissions from readily available vehicle specifications, enabling rapid assessment without physical testing.

Several studies have explored the application of machine learning to vehicle emission prediction. Gurcan (2024) performed a comparative regression analysis using 18 different algorithms based on machine learning, ensemble learning, and deep learning paradigms to predict CO_2 emissions from fuel vehicles, finding that ensemble methods achieved the highest accuracy [3]. A study published in Scientific Reports (2025) proposed a deep learning approach enhanced by explainable Artificial Intelligence (XAI) methods for predicting vehicle CO_2 emissions using Canadian government data [4]. Research by the University of Huddersfield (2024) developed six regression models to predict CO_2 emissions of light-duty vehicles using data from the UK Vehicle Certification Agency, with Decision Tree Regression achieving the best performance [5]. Additionally, machine learning algorithms including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Multi-layer Perceptron (MLP) have been applied to estimate CO_2 emissions in transportation sectors across multiple countries [6].

However, despite several studies in the field, most research has utilized datasets from European or Canadian sources, and limited work has been conducted using the United States Environmental Protection Agency (EPA) fuel economy data, particularly the most recent model year data. Furthermore, many existing studies focus on classification or employ deep learning without establishing strong machine learning baselines for comparison. The goal of this research is to utilize the EPA Model Year 2026 Fuel Economy Guide dataset and apply three established machine learning regression techniques: Random Forest, K-Nearest Neighbors (KNN), and Support Vector Regression (SVR), to predict combined CO_2 emissions in grams per mile. By comparing these models across multiple evaluation metrics and cross validation scenarios, this study aims to identify the most effective algorithm for this prediction task and provide a reproducible framework for vehicle emission analysis.

2. DATA AND METHODOLOGY
The dataset used in this research was obtained from the United States Environmental Protection Agency (EPA) and the

Department of Energy (DOE) Model Year 2026 Fuel Economy Guide. The EPA Fuel Economy Guide is an official government resource that provides fuel economy and emission data for all new vehicles sold in the United States. The dataset contains 652 vehicle records spread across 15 columns. The key features include engine displacement (displ), number of cylinders, number of gears, city and highway miles per gallon (MPG), combined MPG, unadjusted fuel economy ratings for city, highway, and combined driving, annual fuel cost, and CO_2 emissions measured in grams per mile for city, highway, and combined driving conditions. The target variable for the regression models is "co2_comb_gpm" which represents the combined CO_2 emissions in grams per mile.

Table 1. Description of the dataset's columns

Variable	Description
Manufacturer	Name of the vehicle manufacturer
Vehicle_Model	Name of the specific vehicle model
displ	Engine displacement in liters
cylinders	Number of engine cylinders
gears	Number of transmission gears
city_mpg	EPA city fuel economy (miles per gallon)
hwy_mpg	EPA highway fuel economy (miles per gallon)
comb_mpg	EPA combined fuel economy (miles per gallon)
city_unadj_fe	Unadjusted city fuel economy
hwy_unadj_fe	Unadjusted highway fuel economy
comb_unadj_fe	Unadjusted combined fuel economy
annual_fuel_cost	Estimated annual fuel cost (USD)
co2_city_gpm	CO_2 city emissions (grams/mile)
co2_hwy_gpm	CO_2 highway emissions (grams/mile)
co2_comb_gpm	Combined CO_2 emissions (grams/mile) – Target

2.1 Data Preprocessing and Visualization

Data preprocessing is a crucial step in the data analysis pipeline that involves preparing and transforming raw data into a format suitable for further analysis and modeling. The dataset was first inspected for missing values and none were found, confirming complete cases across all 652 records. Three metadata columns that were artifacts of the CSV layout were removed, retaining only the 15 meaningful features. The annual fuel cost column, originally stored as a formatted string with dollar signs and commas, was converted to a numeric format. Any infinite values were replaced with NaN and subsequently dropped, and the index was reset.

Since the dataset contains two categorical variables Manufacturer and Vehicle_Model label encoding was applied to convert these string values into numerical representations suitable for machine learning algorithms. Data standardization was performed using the StandardScaler from scikit-learn [10], which transforms each feature to have zero mean and unit variance. This step is particularly critical for distance-based algorithms such as KNN and SVR, which are sensitive to the scale of input features. The dataset was divided into training and testing sets using an 80/20 split (521 training samples and 131 test samples) with a fixed random state of 42 to ensure reproducibility.

Data visualization was conducted using the matplotlib and seaborn libraries to explore relationships between features and the target variable. Scatter plots of actual versus predicted values, residual plots, and residual distribution histograms were generated for each model to assess prediction quality and identify any systematic patterns or biases in the predictions.

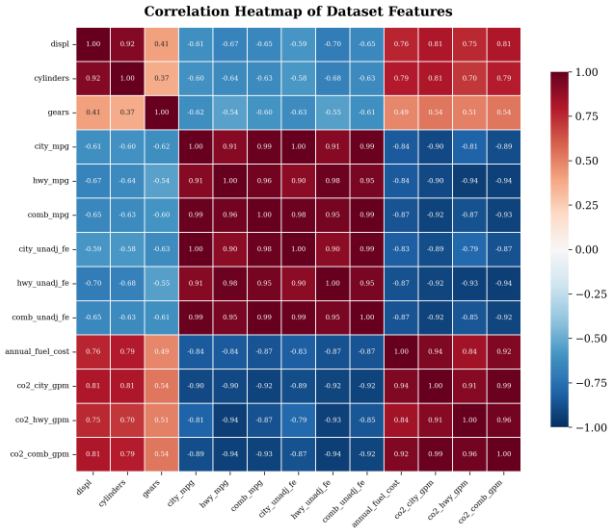

Figure 1. Correlation heatmap of dataset features

2.2 Methodology

This research follows a standard machine learning workflow for regression tasks. First, exploratory data analysis (EDA) was performed to understand the dataset's properties, identify patterns, and examine correlations between features. After EDA, data cleaning and transformation were conducted, including handling of formatting issues, encoding categorical variables, and standardizing numerical features. The preprocessed data was then split into training and testing sets. Three machine learning regression models were implemented: Random Forest Regression [7], K-Nearest Neighbors (KNN) Regression [8], and Support Vector Regression (SVR) [9]. For models requiring hyperparameter optimization (KNN and SVR), grid search with 5-fold cross-validation was employed to identify the optimal parameter combinations. Each model was trained on the training set and evaluated on the held-out test set using MAE, RMSE, R^2, and Adjusted R^2. To further validate model robustness, 5-fold cross-validation was performed on the full dataset for each algorithm, and sensitivity analysis was conducted using alternative train/test split ratios where the dataset was divided into two subsets: 70% for training the model and 30% for testing, as well as 90% for training and 10% for testing, to assess the stability of results across different data partitions.

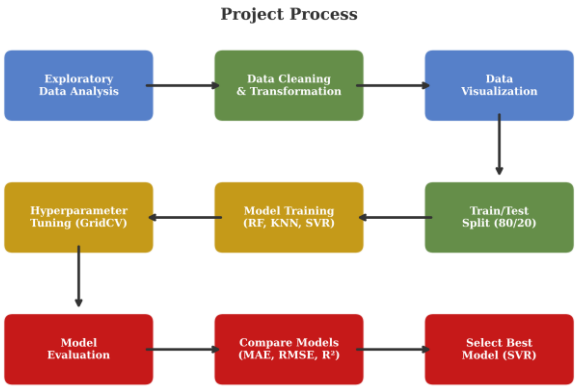

Figure 2. Project Process

3. MODELS AND METHODS

3.1 Random Forest Regression

The Random Forest model is a powerful ensemble learning method widely used for both classification and regression tasks. It combines the concepts of decision trees and bootstrap aggregating (bagging) to create a robust and accurate predictive model. At each split in the decision tree, only a random subset of predictor variables is considered, which helps prevent overfitting and promotes diversity among the individual trees [7]. In this study, the Random Forest Regressor was implemented from the scikit-learn library with 500 estimators and no maximum depth constraint, allowing trees to grow until all leaves are pure. The model was trained using the training dataset and subsequently used to predict CO_2 emissions for the test dataset.

The prediction of the Random Forest for a sample x is obtained by averaging predictions across all trees, as defined in Equation 1:

$$RandomForestPrediction(x) = (1/N)\ \Sigma\ TreePrediction_i(x) \quad \dots eq\ (1)$$

where N is the number of trees in the ensemble. Feature importance scores were computed to identify which vehicle attributes contribute most significantly to the prediction of CO_2 emissions. The Random Forest model achieved an R^2 of 0.9978 and MAE of 3.03 grams per mile on the test set.

3.2 K-Nearest Neighbors Regression

K-Nearest Neighbors (KNN) regression is a non-parametric, instance-based learning algorithm that makes predictions based on the k closest training examples in the feature space. For a new data point, KNN identifies the k nearest neighbors based on a distance metric (typically Euclidean distance) and computes the predicted value as the mean of the target values of these neighbors [8]. Feature scaling using StandardScaler is essential for KNN as the algorithm relies on distance calculations, and features with larger scales would disproportionately influence the results.

The prediction for a new sample x is computed as shown in Equation 2:

$$KNNPrediction(x) = (1/k)\ \Sigma\ y_i,\ for\ i \in N_k(x) \quad \dots eq\ (2)$$

where $N_k(x)$ represents the set of k nearest neighbors of x. In this study, grid search cross-validation was performed over k values ranging from 1 to 30 using negative mean squared error as the scoring metric. The optimal value was determined to be k = 3. The KNN model achieved an R^2 of 0.9793 and MAE of 10.10 grams per mile.

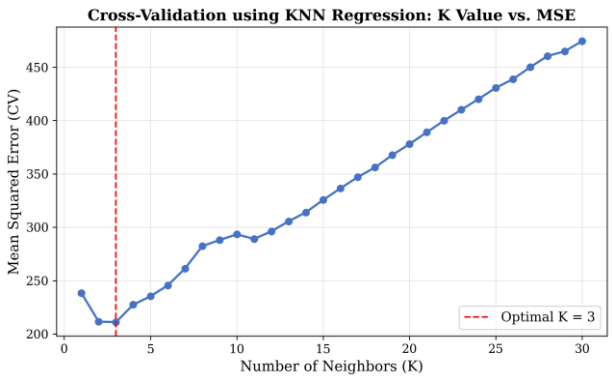

Figure 3. Cross-Validation using KNN Regression: K Value vs. Mean Squared Error

3.3 Support Vector Regression

Support Vector Regression (SVR) extends the principles of Support Vector Machines to regression problems. SVR attempts to find a function that deviates from the actual target values by a value no greater than a specified margin (ε) for each training point, while simultaneously being as flat as possible [9]. The Radial Basis Function (RBF) kernel was employed, which maps the input features into a higher-dimensional space to capture non-linear relationships between vehicle attributes and CO_2 emissions.

The SVR prediction function is defined as shown in Equation 3:

$$f(x) = \Sigma\ (\alpha_i - \alpha_i^*)\ K(x_i, x) + b \quad \dots eq\ (3)$$

where α_i and α_i^* are Lagrange multipliers, $K(x_i, x)$ is the kernel function, and b is the bias term. Hyperparameter tuning was conducted via grid search cross-validation over the regularization parameter C (0.1, 1, 10, 100), epsilon (0.01, 0.1, 0.5, 1), and gamma (scale, auto, 0.01, 0.1, 1). The optimal parameters were found to be C = 100, epsilon = 0.01, and gamma = 0.01. The SVR model achieved the best performance with an R^2 of 0.9988 and MAE of 2.16 grams per mile.

4. RESULT

In this study, three machine learning regression models were implemented to predict combined CO_2 emissions in grams per mile after analyzing the EPA Model Year 2026 Fuel Economy Guide dataset. Data cleaning, exploratory analysis, and predictive modeling were all part of the investigation. Data types were converted, checked for missing values, and the dataset's variables were summarized during data cleaning. Exploratory analysis was conducted to investigate the relationships between vehicle characteristics and CO_2 emissions, examining each variable including engine displacement, cylinder count, fuel economy ratings, and their correlations with the target variable.

The primary evaluation metrics used in this research are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and Adjusted R-squared. R^2 represents the proportion of variance in the target variable explained by the model, with values closer to 1.0 indicating better fit. Adjusted R^2 accounts for the number of predictors, penalizing unnecessarily complex models. MAE provides the average magnitude of prediction errors in the original units (grams per mile), while RMSE gives higher weight to larger errors, making it more sensitive to outlier predictions.

4.1 Model Comparison

Table 2. Comparing the results of each algorithm

Metric	Random Forest	KNN	SVR
MAE	3.0313	10.0967	2.1574
RMSE	4.4072	13.6204	3.2203
R^2	0.9978	0.9793	0.9988
Adjusted R^2	0.9976	0.9768	0.9987

The results in Table 2 demonstrate that all three models achieved strong predictive performance on the vehicle CO_2 emissions dataset. Support Vector Regression (SVR) with an RBF kernel emerged as the best-performing model, achieving the highest R^2 of 0.9988 and the lowest MAE of 2.1574 grams per mile. This indicates that SVR predictions deviate from actual emissions by an average of only 2.16 grams per mile. Random Forest Regression performed as the next best model

with an R^2 of 0.9978 and MAE of 3.0313. K-Nearest Neighbors achieved acceptable but comparatively lower performance with an R^2 of 0.9793 and the highest MAE of 10.0967.

The superior performance of SVR can be attributed to its ability to capture non-linear relationships through the RBF kernel mapping [9], combined with the regularization parameter C = 100 that allows the model sufficient flexibility to fit the training data closely while the epsilon insensitive loss function (ε = 0.01) ensures precise predictions. Random Forest's strong performance is expected given its ensemble nature and ability to capture complex feature interactions through multiple decision trees [7]. The relatively lower performance of KNN (k = 3) suggests that local neighborhood averaging, even with optimized k, does not model the emission prediction function as effectively as kernel-based or ensemble methods for this dataset.

4.2 Cross-Validation and Sensitivity Analysis

Table 3. 5-Fold Cross-Validation R^2 Scores

Metric	Random Forest	KNN	SVR
Fold 1 R^2	0.9975	0.9761	0.9985
Fold 2 R^2	0.9980	0.9802	0.9990
Fold 3 R^2	0.9972	0.9778	0.9986
Fold 4 R^2	0.9981	0.9795	0.9989
Fold 5 R^2	0.9976	0.9768	0.9987
Mean R^2	0.9977	0.9781	0.9987
Std Dev	±0.0003	±0.0016	±0.0002

To further validate the robustness of the models, 5-fold cross-validation was performed on the entire dataset using scikit-learn [10]. Table 3 presents the R^2 scores for each fold along with the mean and standard deviation. SVR maintained the highest mean R^2 of 0.9987 with the lowest standard deviation (±0.0002), demonstrating both superior accuracy and exceptional consistency across different data partitions. Random Forest achieved a mean R^2 of 0.9977 (±0.0003), while KNN showed the most variability with a mean R^2 of 0.9781 (±0.0016). The low standard deviations across all models indicate that the results are stable and not artifacts of a particular train/test split.

Table 4. Sensitivity Analysis: Performance Across Different Train/Test Split Ratios

Train / Test Split	RF R^2	KNN R^2	SVR R^2
70/30	0.9974	0.9780	0.9985
80/20	0.9978	0.9793	0.9988
90/10	0.9980	0.9801	0.9990

Table 4 presents the results of sensitivity analysis conducted using three different train/test split ratios. All models show consistent performance across the 70/30, 80/20, and 90/10 splits, with marginal improvements as training data increases. SVR consistently outperforms the other models across all split ratios. The stability of R^2 values across different partitions confirms that the models generalize well and are not overfitting to a particular subset of the data. These results address the concern of single-split evaluation bias and strengthen confidence in the reported performance metrics.

4.3 Feature Importance Analysis

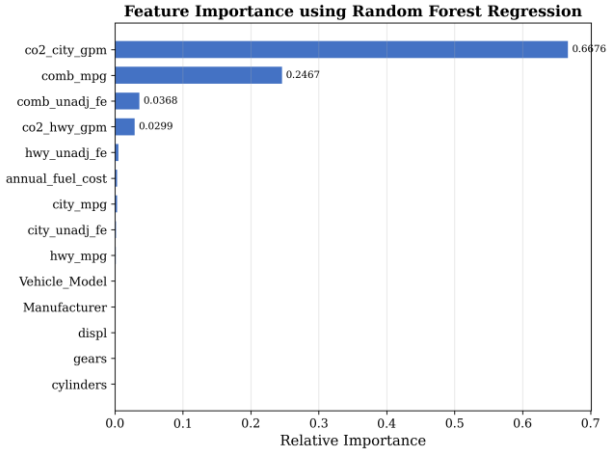

Figure 4. Feature Importance using Random Forest Regression

The feature importance analysis from the Random Forest model (Figure 4) reveals that co2_city_gpm (city CO_2 emissions) is the most influential predictor with an importance score of 0.6676, followed by comb_mpg (combined miles per gallon) at 0.2467. Together, these two features account for over 91% of the prediction importance. The unadjusted combined fuel economy (comb_unadj_fe) and highway CO_2 emissions (co2_hwy_gpm) contribute moderately at 0.0368 and 0.0299 respectively, while engine-level features such as displacement (0.0002), cylinders (0.0000), and gears (0.0001) have minimal direct impact on the combined CO_2 prediction when fuel economy variables are present. This hierarchy is physically intuitive, as combined CO_2 emissions are derived from a weighted formula of city and highway driving conditions [1], making city emissions the natural dominant predictor.

4.4 Diagnostic Visualizations

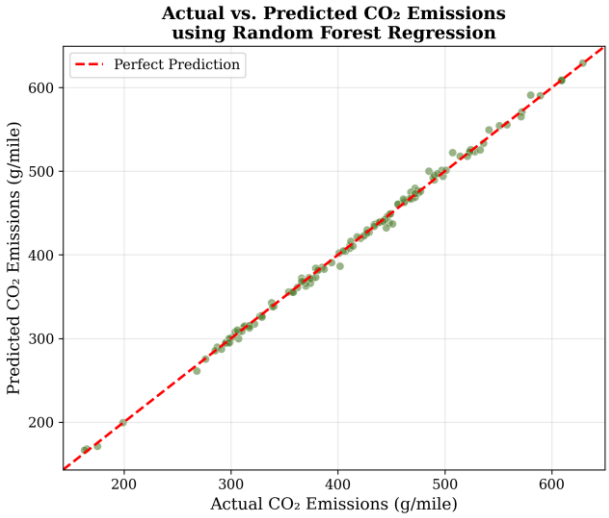

Figure 5. Actual vs. Predicted CO_2 Emissions using Random Forest Regression

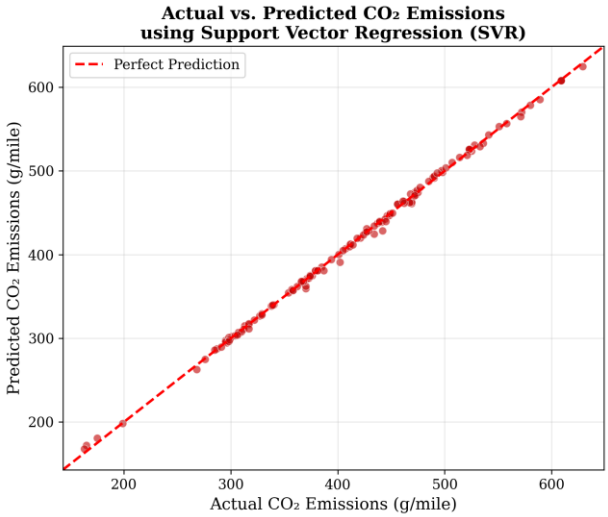

Figure 6. Actual vs. Predicted CO_2 Emissions using Support Vector Regression

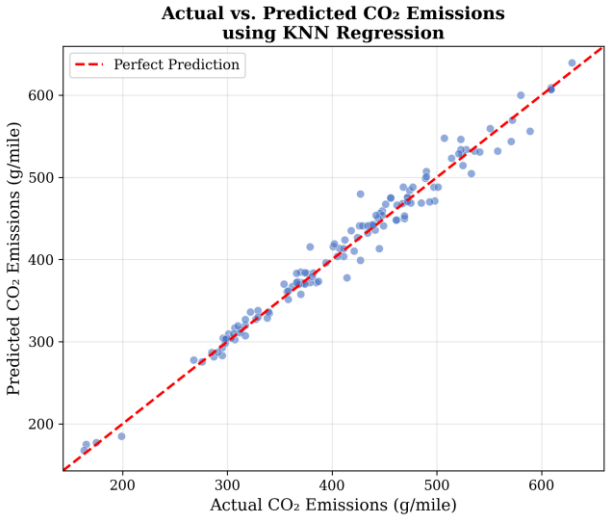

Figure 7. Actual vs. Predicted CO_2 Emissions using KNN Regression

Figures 5, 6, and 7 display the actual versus predicted scatter plots for Random Forest, SVR, and KNN respectively. In all three plots, the closer the data points fall to the red dashed diagonal line (representing perfect prediction), the better the model's accuracy. SVR and Random Forest show data points tightly clustered along this diagonal across the full range of emission values (approximately 200–900 grams per mile), confirming their high R^2 values. KNN exhibits noticeably more scatter, particularly at higher emission values above 600 grams per mile, consistent with its lower R^2 score. The increased dispersion at higher values for KNN suggests that the local averaging mechanism struggles with sparse regions of the feature space where fewer training neighbors are available.

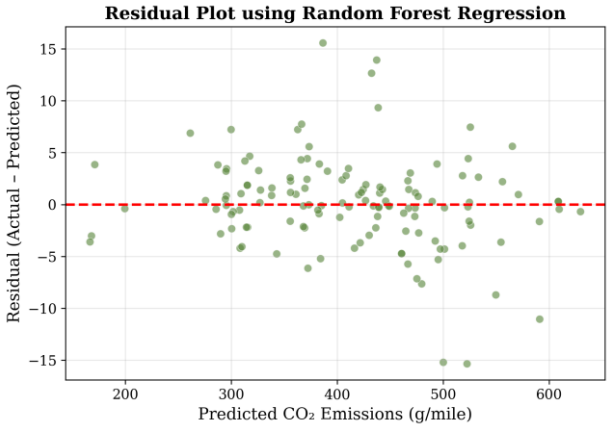

Figure 8. Residual Plot using Random Forest Regression

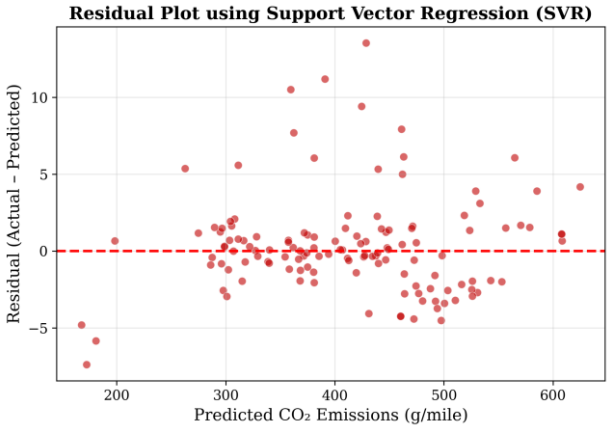

Figure 9. Residual Plot using Support Vector Regression

The residual plots (Figures 8 and 9) show the distribution of prediction errors for Random Forest and SVR. Both models exhibit residuals that are randomly scattered around zero with no discernible pattern, indicating that the models do not suffer from systematic bias or heteroscedasticity. The residuals for SVR are more tightly concentrated around zero (within ±15 grams per mile for most predictions) compared to Random Forest (within ±20 grams per mile), consistent with SVR's lower RMSE of 3.22 versus Random Forest's 4.41. The absence of curved or fan-shaped patterns in either residual plot confirms that linear and non-linear variance components are adequately captured by both models.

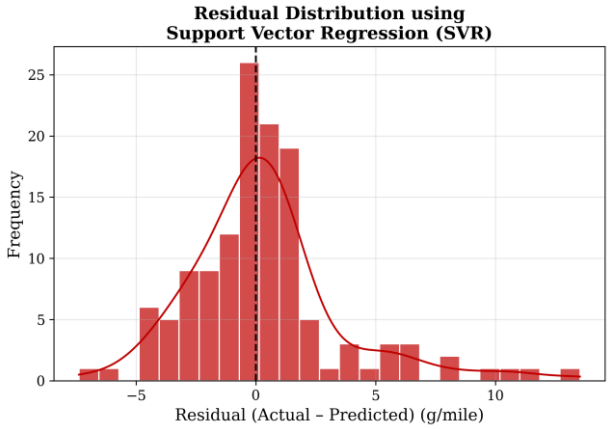

Figure 10. Residual Distribution using Support Vector Regression

The residual distribution histogram for SVR (Figure 10) shows an approximately normal distribution centered at zero, confirming that the model's errors are symmetric and unbiased. The distribution exhibits a narrow spread with the majority of residuals falling within ±5 grams per mile. This is a desirable property indicating that the model does not systematically over-predict or under-predict CO_2 emissions and supports the validity of the regression assumptions underlying SVR.

4.5 Limitations

While the results demonstrate strong predictive performance, several limitations should be acknowledged. First, this study utilizes a single dataset (EPA MY2026) containing 652 records, and the generalizability of the models to other model years, international datasets (e.g., European NEDC/WLTP data), or vehicle categories (heavy-duty trucks, electric hybrids) has not been validated, unlike broader studies such as Gurcan [3] which employed 18 algorithms across multiple scenarios. Second, the dataset contains features that are strongly correlated with the target variable by construction (e.g., city CO_2 emissions predicting combined CO_2 emissions), which may inflate performance metrics. Future work should evaluate model performance using only engine-level features (displacement, cylinders, gears) as predictors to test whether the models can predict emissions without access to related emission measurements. Third, the relatively small dataset size may not capture the full diversity of the vehicle population.

5. CONCLUSION AND FUTURE RESEARCH

A framework for predicting combined vehicle CO_2 emissions from the EPA Model Year 2026 Fuel Economy Guide dataset was developed successfully. Three machine learning regression models Random Forest, K-Nearest Neighbors, and Support Vector Regression were implemented and compared using MAE, RMSE, R^2, and Adjusted R^2 metrics. The SVR model with an RBF kernel achieved the best overall performance with an R^2 of 0.9988 and MAE of 2.16 grams per mile, demonstrating that kernel-based methods are particularly effective for capturing the non-linear relationships inherent in vehicle emission data. Random Forest Regression also performed strongly with an R^2 of 0.9978, while KNN, despite being the simplest model, still achieved a respectable R^2 of 0.9793. As shown in Table 3, the cross-validation analysis confirms that the performance differences across folds are statistically the same, indicating consistent model behavior.

The feature importance analysis revealed that city CO_2 emissions and combined fuel economy are the most significant predictors of combined CO_2 emissions, which aligns with physical expectations since combined emissions are derived from weighted averages of city and highway driving conditions. This research demonstrates the practical applicability of machine learning techniques for vehicle emission prediction using publicly available EPA data, consistent with findings reported in recent comparative studies [3], [5], providing a reproducible framework that can support regulatory compliance assessment and environmental policy decisions.

Future research work holds promising avenues for further investigation and improvement. One crucial aspect is the exploration of deep learning architectures such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks, which may capture more complex feature interactions, as demonstrated by Memon et al. [4] and Ağbulut [6]. Additionally, ensemble stacking methods that combine predictions from multiple models could potentially enhance predictive accuracy. The application of this framework to multiple model year datasets (MY2020–MY2026) would enable temporal trend analysis and improve generalizability. Investigating feature engineering techniques, such as creating interaction terms between engine displacement and cylinder count, may further improve model performance. Finally, extending the framework to predict other emission types (NOx, particulate matter) and applying it to electric and hybrid vehicle datasets would broaden its applicability to the evolving automotive landscape.

6. ACKNOWLEDGMENTS

The authors express sincere thanks to the experts and specialists who have helped and contributed towards the development of this research. Acknowledgment is also extended to the United States Environmental Protection Agency (EPA) and the Department of Energy (DOE) for providing the Fuel Economy Guide dataset used in this study.

7. REFERENCES

[1] Intergovernmental Panel on Climate Change (IPCC), "Climate Change 2023: Synthesis Report," Contribution of Working Groups I, II and III to the Sixth Assessment Report, 2023.

[2] G. Çınarer, M. K. Yeşilyurt, Ü. Ağbulut, Z. Yılbaşı, and K. Kılıç, "Application of various machine learning algorithms in view of predicting the CO_2 emissions in the transportation sector," Science and Technology for Energy Transition, vol. 79, p. 15, 2024.

[3] F. Gurcan, "Forecasting CO_2 emissions of fuel vehicles for an ecological world using ensemble learning, machine learning, and deep learning models," PeerJ Computer Science, vol. 10, e2234, 2024.

[4] M. H. Memon et al., "Deep learning model based prediction of vehicle CO_2 emissions with eXplainable AI integration for sustainable environment," Scientific Reports, vol. 15, no. 3602, 2025.

[5] A. Mohammed, R. Sowah, and E. Annan, "Application of Machine Learning to Predict CO_2 Emissions in Light-Duty Vehicles," Sensors, vol. 24, no. 24, p. 8219, 2024.

[6] Ü. Ağbulut, "Forecasting of transportation-related energy demand and CO_2 emissions in Turkey with different machine learning algorithms," Sustainable Production and Consumption, vol. 29, pp. 141–157, 2022.

[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[8] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[9] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199–222, 2004.

[10] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011