JSCAI

Journal of Soft Computing and Artificial Intelligence

ISSN (Online) 2717-8226

https://dergipark.org.tr/en/pub/jscai

Research Article

Fuel Efficiency and Emission Prediction Using Auto MPG and EPA Data: A Machine Learning and Deep Learning Approach

Zeynep ÇEKEN[1] **and Fahriye GEMCİ**[1*]

[1*] *Department of Computer Engineering, Faculty of Engineering and Architecture, University of Kahramanmaras Sutcu Imam, Kahramanmaras, Turkey*

[*] *Corresponding author*

Article Info

Keywords:
1. Fuel efficiency of vehicle
2. Carbon emissions of vehicle
3. Convolutional Neural Network
4. Random Forest
5. Xgboost

Received: *29.08.2025*
Accepted: *14.10.2025*
Available online: *15.12.2025*

Abstract

This study consists of two stages. In the first stage, the Auto MPG dataset, and in the second stage, the Fuel Economy dataset published by the United States Environmental Protection Agency (EPA) are utilized to classify vehicle fuel efficiency and predict carbon emissions. In this context, multi-layered and multi-functional artificial intelligence and machine learning-based models capable of solving both classification and regression problems were applied. Comparative analyses were conducted by testing different models such as Random Forest Classifier, Support Vector Classifier (SVC), Logistic Regression, Artificial Neural Network (ANN), and Convolutional Neural Network (CNN).

In the first stage of this project, it was aimed to model the MPG (Miles Per Gallon) value, which represents fuel consumption efficiency, using both regression and classification methods based on vehicles' technical parameters. In the second stage, the classification model focuses on predicting vehicle categories such as A+, A, B, C, and D based on the "Fuel Economy Score," while the regression model aims to accurately predict the "Tailpipe CO2 in Grams/Mile (FT1)" value. The performance of the models was evaluated using various metrics such as accuracy, ROC curves, and mean error [15].

It was observed that CNN achieved highly successful results in classification tasks, while XGBoost produced strong results in regression tasks. This indicates that CNN models can be effectively applied not only in image processing tasks but also in diverse datasets, and XGBoost provides reliable predictions for numerical targets.

The results reveal that both deep learning and traditional models provide high accuracy in vehicle efficiency prediction and environmental impact analysis. In this respect, the study offers significant contributions to intelligent transportation systems and sustainable environmental policies.

Email address: fahriyegemci@ksu.edu.tr
Cite as: Z. Çeken, F. Gemci, *Fuel Efficiency and Emission Prediction Using Auto MPG and EPA Data: A Machine Learning and Deep Learning Approach*, Journal of Soft Computing and Artificial Intelligence, 6(2) (2025), 30-38.

1. Introduction

Today, environmental sustainability and the fight against climate change have become top global priorities. In particular, carbon emissions from land vehicles are recognized as one of the largest sources of greenhouse gas accumulation in the atmosphere, directly contributing to the global warming process. CO_2 (carbon dioxide) emissions from vehicles not only reflect the amount of energy consumption but also serve as a concrete indicator of environmental damage. For this reason, many countries around the world are tightening emission standards and developing environmentally friendly transportation policies to enhance vehicle efficiency and reduce harmful gas emissions.

In this context, it is essential to evaluate the environmental impacts of vehicles not only during the production process but also throughout their entire lifecycle. Predicting vehicle fuel consumption and CO_2 emission levels in advance is particularly important to guide consumers toward environmentally friendly vehicles and to ensure the effectiveness of government-supported incentives. However, traditional statistical methods are often inadequate for such predictions, as they face difficulties in effectively modeling complex and multivariate data.

At this point, Artificial Intelligence (AI) and Machine Learning (ML) techniques have the potential to provide advanced predictions for both manufacturers and users by extracting meaningful insights from vehicle data. In particular, deep learning algorithms [10] are capable of modeling complex structures within data more effectively, automatically learning relationships among numerous variables, and yielding more accurate predictions.

In this study, machine learning and artificial intelligence models [11] are employed to predict vehicle fuel efficiency (Miles Per Gallon – MPG) and carbon emission levels (Tailpipe CO_2 in Grams/Mile). The project was carried out in two main phases. In the first phase, the Auto MPG dataset was used to analyze vehicle efficiency values using both regression and classification methods. In the second phase, a more comprehensive and real-world-oriented analysis was performed with the Fuel Economy dataset published by the EPA.

Across both stages, modeling was conducted using a range of algorithms, including Random Forest, SVM, Decision Tree, Logistic Regression, MLP, ANN, and CNN [9]. The resulting models were evaluated with multiple performance metrics such as accuracy, F1 score, mean error, and ROC AUC [13], assessing both classification and regression success.

This study makes significant contributions to the existing literature on vehicle fuel efficiency and carbon emission prediction. While most previous studies typically adopt a single machine learning algorithm focusing Either solely on regression or solely on classification, research that applies multi-stage modeling with datasets of varying sizes remains limited. By contrast, this study integrates both classification and regression tasks into a two-stage modeling framework. In the first stage, the historical and small-scale Auto MPG dataset was used to build baseline models, applying classical machine learning algorithms such as Random Forest, SVM, KNN, and Logistic Regression. In the second stage, a larger and more up-to-date EPA Fuel Economy dataset was employed, where a 1D-CNN [18] architecture was applied for multi-class fuel efficiency classification, and XGBoost was used for carbon emission prediction. This framework combines the complex pattern recognition capabilities of deep learning with the high predictive performance of ensemble methods, distinguishing it from many studies in the literature.Moreover, model performance was evaluated in a detailed and interpretable manner using ROC-AUC, MAE, and feature importance analyses. Thus, the proposed framework not only provides academic contributions but also offers valuable practical implications for sustainable transportation systems, environmental impact assessment, and decision support mechanisms.

Numerous studies in the literature focus on vehicle efficiency and emission prediction [16, 17]. For example, Akhatkulovet al. [16] similarly compared various machine learning algorithms for fuel consumption prediction and confirmed the superior performance of tree-based methods such as Random Forest and XGBoost.Likewise, Ji et al. [17] highlighted the power of deep learning models in CO2 emission predic-

tion.However, a significant portion of the existing literature tends to focus on either a single dataset (either historical and small or contemporary and large) or a single type of modeling task (either only regression or only classification). This study utilized multiple datasets and, in conjunction, conducted both regression and classification studies for vehicle efficiency and emissions estimation.

2. Material and Methods

In this study, different classical machine learning and deep learning methods were applied to address both classification and regression problems. The Auto MPG dataset [2], obtained from the UCI Machine Learning Repository, contains 398 vehicle records with technical specifications and fuel consumption values of vehicles produced in the 1970s and 1980s, while the Fuel Economy dataset [1], published by the United States Environmental Protection Agency (EPA) and available on Kaggle, provides more than 38,000 records of vehicles sold between 1984 and 2024, including detailed information on fuel consumption, engine characteristics, emission values, and energy efficiency. Among the models used, Random Forest [3] is an ensemble method that combines multiple decision trees trained on random subsets to improve generalization and reduce overfitting; KNN [4] classifies new data by considering the majority class among the k closest samples based on distance metrics such as Euclidean distance; SVM [5] seeks the optimal hyperplane to separate data into classes, using kernel functions to handle non-linear problems; Logistic Regression [7] is a linear method that estimates probabilities via a sigmoid function, particularly effective for binary classification; and Multilayer Perceptron (MLP) [8] is a feed-forward artificial neural network that transforms input data through hidden layers using activation functions and updates weights via backpropagation. In addition, CNN [9], although primarily designed for visual data, were employed in this study to classify vehicle efficiency classes (A+, A, B, C, D), leveraging their ability to capture complex patterns and spatial relationships in structured datasets. For the regression task, Extreme Gradient Boosting (XGBoost) [6] was used to predict carbon emissions, benefiting from its gradient boosting framework that sequentially trains weak learners to achieve robust and accurate results. By applying these approaches independently, the study evaluates the advantages of both deep learning (CNN, MLP) in efficiency classification and traditional algorithms such as Random Forest, SVM, Logistic Regression, and XGBoost in emission prediction, demonstrating their effectiveness on different types of tasks.

2.1. Data Preprocessing

Before training the models, a comprehensive preprocessing procedure was applied to the EPA Fuel Economy dataset in order to improve data quality and prepare the input for both machine learning and deep learning models. In the first step, unnecessary and redundant columns such as Vehicle ID, Engine Index, Engine Descriptor, Unrounded City MPG (FT1), and Unrounded Highway MPG (FT1) were removed from the dataset. Afterwards, the Year column was converted into numeric format, and rows that could not be converted were discarded. To ensure consistency with the Auto MPG dataset, vehicles manufactured before 1980 were also excluded.

Efficiency labels were then created using the Tailpipe CO_2 in Grams/Mile (FT1) attribute. Based on this variable, vehicles were categorized into five classes: A+(¡250 g/mile), A (¡350 g/mile), B (¡450 g/mile), C (¡550 g/mile) , and D ($\geq$ 550 g/mile) . These categorical labels were further processed using Label Encoding. In addition, six technical attributes—Engine Displacement, Engine Cylinders, City MPG (FT1), Highway MPG (FT1), Combined MPG (FT1), and Year—were selected as features for model training, and rows with missing values in these attributes were removed. The dataset was then split into training and test sets with an 80-20 ratio using stratified sampling to preserve the class distribution, ensuring separate preparation for both classification and regression tasks.

To address the imbalance among classes, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. Following this, all features were normalized to the [0,1] range with MinMaxScaler, which was fitted on the training data and applied to the test data to prevent data leakage. Finally, for deep learning applications, particularly the 1D-CNN model, the scaled data was reshaped by adding an additional channel dimension, making it suitable for convolutional

layers.

2.2. Model Architectures and Hyperparameters

For regression tasks, the Random Forest Regressor was implemented with fixed hyperparameters, including n_estimators=500, max_depth=6, colsample_bytree=0.8, subsample=0.8, and learning_rate=0.05. Similarly, XGBoost Regressor was applied to predict carbon emissions using the same parameter configuration, benefiting from its gradient boosting framework and strong generalization capabilities. For the classification of vehicle efficiency classes, a 1D Convolutional Neural Network (1D-CNN) was designed. The architecture consisted of a convolutional layer with 64 filters (kernel size=3, activation=ReLU, padding="same"), followed by batch normalization, a global max pooling layer, and a dense hidden layer with 128 neurons and ReLU activation. To mitigate overfitting, a dropout layer with a rate of 0.3 was included, and the final dense output layer used Softmax activation to predict efficiency classes. The model was trained with the Adam optimizer (learning_rate=0.001) and the sparse categorical cross-entropy loss function. To further enhance training, EarlyStopping (patience=10, restore_best_weights=True) and ReduceLROnPlateau (factor=0.5, patience=5) callbacks were employed. In addition to CNN, a MLP was also developed for classification. The MLP was structured as a fully connected feed-forward neural network with an input layer matching the feature dimensions, two hidden layers with ReLU activation functions, dropout regularization to prevent overfitting, and a Softmax output layer for classification. By clearly defining the preprocessing pipeline and model architectures, this study ensures reproducibility and highlights the complementary strengths of classical machine learning methods (Random Forest, XGBoost) and deep learning architectures (CNN, MLP) in solving both classification and regression problems.

2.3. System Architecture and Workflow

The overall system architecture and workflow of the proposed two-stage AI framework is illustrated in Figure 2.1. The process begins with data loading and preprocessing, where the EPA Fuel Economy dataset is cleaned, encoded, and prepared for model training. Feature engineering includes categorization of vehicles into efficiency classes (A+, A, B, C, D) based on tailpipe CO_2 emissions, followed by label encoding for numerical processing. The preprocessing pipeline incorporates SMOTE for handling class imbalance and MinMaxScaler for feature normalization. The modeling phase encompasses both classification and regression tasks, employing a diverse set of algorithms including Random Forest, SVM, KNN, XGBoost, and 1D-CNN architectures. Model performance is evaluated using comprehensive metrics, with results visualized through confusion matrices, ROC curves, and comparative performance charts. The framework culminates in an interactive Tkinter-based GUI application that enables real-time vehicle efficiency classification and CO_2 emission predictions, providing an end-to-end solution for environmental vehicle analytics.

3. Experimental Results

In this study, a multi-stage and multi-model system that performs both classification and regression tasks on two separate datasets is designed. The modelling process is divided into two main stages. In the first stage, classical machine learning models [12] were applied using the Auto MPG dataset; in the second stage, deep learning architectures suitable for more complex and real-world data were developed with the EPA Fuel Economy dataset.

3.1. Regression Model for First Stage Modelling on Auto MPG Dataset: Random Forest Regressor

Using the technical attributes (e.g. weight, horsepower, number of cylinders) in the Auto MPG dataset, the MPG (Miles Per Gallon) value, which expresses the fuel consumption of vehicles, was numerically estimated. For this purpose, the Random Forest Regressor model, an ensemble learning method based on decision trees, was preferred. The hyperparameters of the model such as n_estimators, max_depth, max_features were optimised by GridSearchCV method. The performance of the model was evaluated both with the Mean Square Error (MSE)

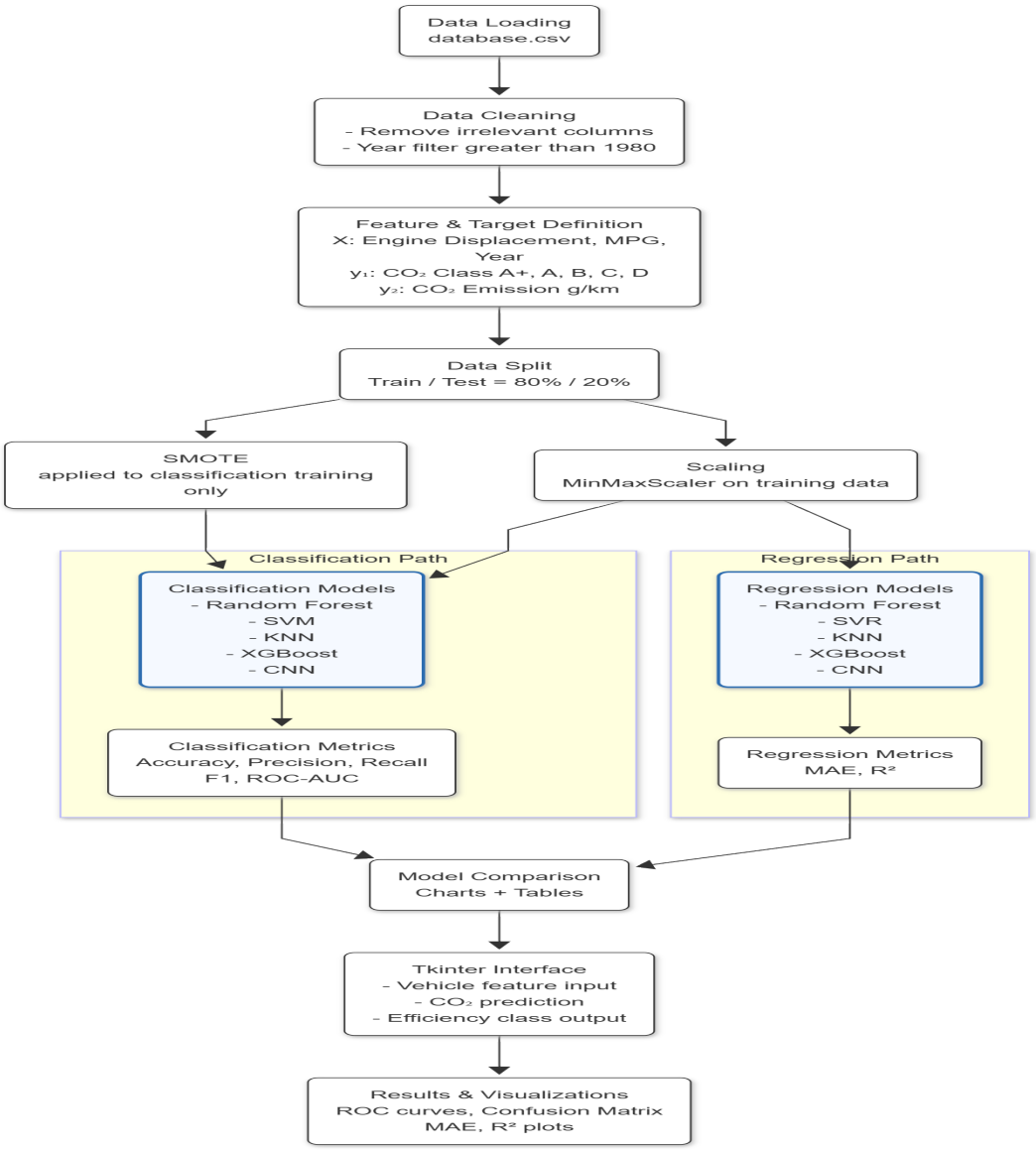

Figure 2.1: Workflow diagram

and R² score on the test set and with the 5-fold cross-validation (K-Fold CV) method. The prediction performance is visualised with a scatter plot comparing the actual and predicted MPG values. In addition, the contribution ratios of the features to the model are presented with feature importance graphs.MPG values are divided into two classes as low and high according to a certain threshold (30 mpg), thus defining the classification problem. Random Forest, SVM, KNN, Logistic Regression and Decision Tree algorithms are applied..Since the Random Forest algorithm shows the highest success, the results of this Random Forest algorithm before the SMOTE process in stage 1 are shown in Table 3.1.

Metric	Precision	Recall	F1-Score
Macro Avg	0.66	0.79	0.67
Weighted Avg	0.89	0.78	0.82

Table 3.1: Comparison of activation functions in a neural network for binary classification.

3.2. Second Stage Modelling on EPA Fuel Economy Dataset

The 1D CNN model trained for efficiency class (A+, A, B, C, D) prediction is evaluated on the EPA Fuel Economy test dataset. It shows that the CNN architecture can effectively learn the distinction between classes and the class performances are balanced as given in Fig.

Economy test data. As shown in Table 2.1., the XGBoost model showed strong regression performance with high

accuracy and explanatory capacity on large data. ROC AUC [14], MAE, R2 performence resuts of XGBoost algorithm for carbon emissions on Second Stage Modelling is 0.999, 2.34 and 1.00, respectively.

In the first stage, the Auto MPG dataset provided by UCI is used and regression and classification tasks are performed on this small dataset containing a limited number of samples. Random Forest Regressor is used as the regression model, and classical machine learning algorithms such as Random Forest, SVM, KNN, Logistic Regression and Decision Trees are used in classification tasks. These models are preferred due to their ability to work with small sample data, explainability and fast training features. These analyses performed on the Auto MPG dataset are successful in understanding basic modeling skills and observing the performances of various algorithms. In the second stage, the much larger and more comprehensive EPA Fuel Economy dataset was used. In this stage, deep learning and advanced machine learning techniques are applied on data with a more complex structure. The 1D-CNN model learned the distinction between classes more successfully and provided a noticeable improvement in critical metrics such as ROC-AUC. For the regression task, the XGBoost Regressor model is preferred and is able to predict the carbon emissions of vehicles with high accuracy. XGBoost's decision tree-based structure has been effective in modeling complex relationships; in addition, the model's decision process has been made explanatory through attribute importance levels. In general, this study has provided comparative testing of both basic and advanced algorithms on different data structures; and has presented a successful system for environmental prediction problems by determining the most appropriate models. The findings prove that artificial intelligence technologies offer effective and applicable solutions in areas such as environmental sustainability, environmentally friendly vehicle preference, and decision support systems.

3.3. Comprehensive Model Performance Comparison

To provide a holistic evaluation of all implemented algorithms, Figure 3.1 presents a comprehensive performance comparison across both classification and regression tasks. The visualization clearly demonstrates the superior capability of CNN architectures in classification, achieving exceptional performance with 0.998 accuracy, 0.997 precision, and balanced recall-f1 scores across all efficiency classes (A+, A, B, C, D). For regression tasks, XGBoost dominates with near-perfect R^2 scores (0.997) and minimal mean absolute error, significantly outperforming traditional algorithms. The comparative analysis reveals that while classical machine learning models (Random Forest, SVM, KNN) maintain competitive performance, deep learning approaches consistently deliver superior results in handling the complex, multidimensional patterns present in the EPA Fuel Economy dataset. This performance advantage is particularly evident in the ROC-AUC metrics and R^2 comparisons, where CNN and XGBoost demonstrate remarkable generalization capability on unseen test data.

4. Conclusion

In this study, a two-stage artificial intelligence–based framework was developed to estimate key environmental indicators such as vehicle fuel efficiency class and tailpipe CO_2 emissions by using both technical and environmental parameters. Classical machine learning algorithms and deep learning architectures were applied at different stages to evaluate their comparative performance and practical applicability. In the first stage, classification models were employed to predict the vehicle's fuel efficiency class. The experimental results demonstrated that even traditional algorithms such as Random Forest maintained stable predictive performance on relatively small and imbalanced datasets. Among all classifiers, the CNN model achieved superior performance across key evaluation metrics such as Accuracy, F1-score, and ROC-AUC, confirming its ability to effectively capture complex feature relationships and spatial patterns in multi-dimensional vehicle data. In the second stage, regression models were applied to estimate tailpipe CO_2 emissions as a continuous environmental parameter. The XGBoost regressor exhibited the lowest Mean Absolute Error (MAE = 1.66) and the highest R^2 score (R^2 = 0.997), outperforming other regression

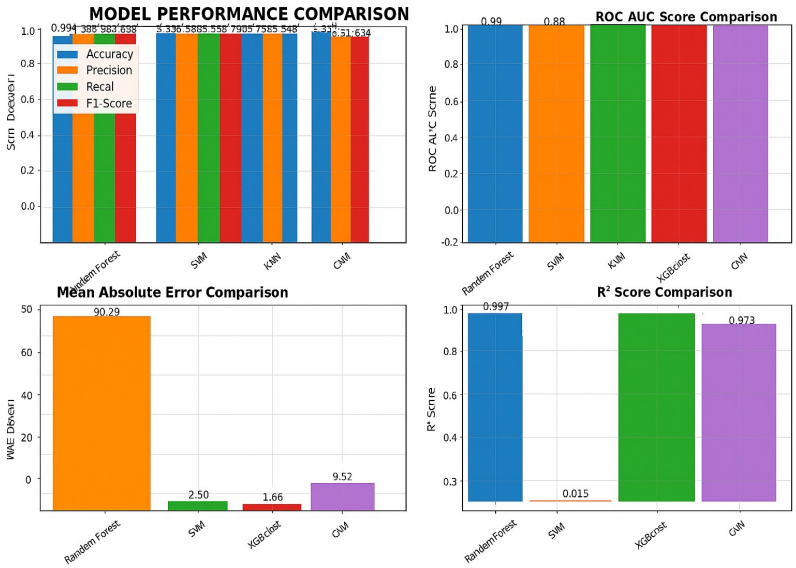

Figure 3.1: Comparative Results of Model Performance

algorithms such as SVM and KNN. This result highlights XGBoost's capability to model non-linear dependencies and to produce robust, high-accuracy emission predictions. Additionally, feature importance analysis provided interpretable insights into which technical variables most strongly influence CO_2 output, supporting the model's transparency and potential integration into real-world emission monitoring systems. Overall, both stages of the proposed system achieved consistent and generalizable results across training and test sets. The successful integration of classification and regression tasks illustrates a comprehensive approach that can contribute to sustainable transportation strategies and intelligent environmental analysis. These findings confirm that artificial intelligence methods can deliver effective, interpretable, and practical solutions for environmental data modeling, vehicle performance assessment, and carbon footprint estimation.

5. Discussion

In this study, a two-stage modeling approach was adopted for predicting vehicle fuel efficiency and carbon emissions using both classical machine learning and deep learning models. The experimental results demonstrated that the proposed framework is highly successful in both classification and regression problems.

The most important contribution of this work is overcoming this dichotomy through an integrated approach. In the first stage, fundamental modeling skills were developed using a small, historical dataset (Auto MPG), while in the second stage, the performance of more complex architectures (1D-CNN, XGBoost) was tested on a large-scale, contemporary dataset (EPA). This multi-stage structure provided a unique opportunity to observe model behavior under different data conditions. Furthermore, the use of 1D-CNN on structured tabular data is relatively novel in the literature.

On the other hand, the feature importance rankings provided by the model offered an explainable insight into the prediction process, contributing to inferences that are not only accurate but also interpretable.

The limitations of the study should also be considered. Firstly, as the models are based on technical specifications, real-world variables such as driving behavior, traffic conditions, or weather were not included. Secondly, although the class distribution in the EPA dataset was balanced using SMOTE, the impact of these synthetic samples on the model's generalization performance could be investigated in more depth. Additionally, this study is limited to North American market data; the extent to which the models would adapt to vehicle data from different geographies could be explored in future work.

In future studies, a multimodal approach could be adopted to overcome these limitations; for example, real-time sensor data or satellite imagery could be integrated alongside technical data. Furthermore, advanced methods such as SHAP or LIME could be used to further enhance model explainability. Finally, the development

of a prototype software for integrating these predictive models into real-life applications, such as intelligent transportation systems and carbon emission tracking platforms, is planned.

In conclusion, this study has shown that a hybrid modeling strategy, optimized for both classification and regression on datasets of different scales, is feasible and highly successful. This approach offers valuable contributions not only to academic research but also to the development of sustainable transportation policies and decision support systems for consumers choosing environmentally friendly vehicles.

Article Information

Acknowledgements: The authors thank the editor and anonymous reviewers for their valuable feedback and constructive suggestions that helped improve the quality of this manuscript.

Author Contributions: In the study, Zeynep Çeken contributed to the formation of the idea, data collection, analysis and obtaining the results; Fahriye GEMCİ contributed to the literature review, selection of the methods used, examination of the results, interpretation, and control of the article in terms of content.

Artificial Intelligence Statement: No AI tools were used in the writing or editing of this manuscript.

Conflict of Interest Disclosure: The authors declare that they have no conflict of interest.

Plagiarism Statement: This manuscript has been checked for plagiarism using appropriate similarity detection tools and complies with the ethical standards of JSCAI.

References

[1]. United States Environmental Protection Agency (EPA), "Fuel economy data (1984–present)," [Online]. Available: https://www.fueleconomy.gov/feg/. Accessed June 16, 2025.

[2]. UCI Machine Learning Repository, "Auto MPG dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/auto+mpg. Accessed June 16, 2025.

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

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

[5]. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[6]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016.

[7]. D. W. Hosmer and S. Lemeshow, Applied Logistic Regression. New York: Wiley, 2000.

[8]. F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psychological Review, vol. 65, no. 6, pp. 386–408, 1958.

[9]. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[10]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[11]. A. M. Foley and A. G. Olabi, "Renewable energy technology and the environment: a review," Renewable and Sustainable Energy Reviews, vol. 79, pp. 1321–1340, 2017.

[12]. X. Wu, V. Kumar, J. R. Quinlan, et al., "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, 2008.

[13]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[14]. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, 1982.

[15]. K. H. Zou, A. J. O'Malley, and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," Circulation, vol. 115, no. 5, pp. 654–657, 2007.

[16]. S Akhatkulov,I Yalgoshev, Z. and Urinboyev, "Vehicle CO_2 Emission Prediction Using Deep Learning and Ensemble Machine Learning Methods," In 2025 International Russian Automation Conference (RusAutoCon) (pp. 819-824). IEEE.

[17.] T Ji, K Li, Q Sun, and Z Duan,"Urban transport emission prediction analysis through machine learning and deep learning techniques" Transportation Research Part D: Transport and Environment, vol. 135, no. 104389, pp. 654–657,2024