Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Machine learning based prediction of automotive fuel economy

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.en.6.5.197-213

***Corresponding author:**
Email address:
nguyenvanhau3912@gmail.com

Received: 06/05/2026
Received in Revised Form: 28/09/2026
Accepted: 20/10/2026

Sushant Kumar Singh[1], Nguyen Van Hau[2,*], Tien Hung Nguyen[3], Hoang-Minh Luu[4], Indra Prakash[5]

[1]Center for Artificial Intelligence and Environmental Sustainability (CAIES) Foundation, Patna, Bihar, India; sushantorama@gmail.com

[2]University of Transport Technology, 54 Trieu Khuc, Thanh Liet, Hanoi, Vietnam; nguyenvanhau3912@gmail.com

[3]Advanced Materials and Intelligent Systems for Infrastructure and High-Speed Rail (AMIS-HSR) research group, University of Transport Technology, Hanoi, Vietnam; nguyentienhung@utt.edu.vn

[4]University of Transport Ho Chi Minh City, 2, Vo Oanh, Thanh My Tay, Ho Chi Minh 700000, Vietnam; minh.luu@ut.edu.vn

[5]Formerly Dy. Director General, Geological Survey of India, Gujarat, India; indra52prakash@gmail.com

Abstract: Accurate prediction of automotive fuel economy is important for improving vehicle design, reducing operating costs, and supporting energy-efficient transportation systems. Conventional empirical and statistical approaches often fail to capture the complex nonlinear relationships among vehicle technical characteristics. This study presents a comparative evaluation of four state-of-the-art ensemble machine learning algorithms, namely Random Forest (RF), Light Gradient Boosting Machine (LGB), XGBoost (XGB), and CatBoost (CGB), for predicting vehicle fuel economy using the publicly available UCI Auto MPG dataset, comprising 398 vehicle records characterized by seven technical attributes (392 valid samples after preprocessing). The models were optimized using Grid search optimization (GSO) and evaluated through an 80:20 training-testing split combined with five-fold cross-validation. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Model interpretability was investigated using SHapley Additive exPlanations (SHAP) to quantify the influence of individual vehicle attributes. Among the evaluated algorithms, CGB achieved the best predictive performance with a testing R^2 of 0.9204, normalized RMSE of 0.0473, normalized MAE of 0.0347, and MAPE of 9.07% on the normalized target scale, indicating promising performance on the held-out test set. SHAP analysis revealed that model year, vehicle weight, horsepower, and engine displacement were the most influential variables affecting fuel economy predictions. The study demonstrates that optimized ensemble learning methods can accurately model complex relationships between vehicle specifications and fuel economy while providing interpretable predictions. Although the findings are based on a benchmark dataset representing vehicles manufactured between 1970 and 1982, the proposed comparative framework provides a transparent comparative framework that can be extended to larger and more contemporary automotive datasets.

Keywords: Fuel economy, automotive engineering, XGBoost, CatBoost, Light Gradient Boosting Machine (LGB), machine learning, SHAP analysis, Random Forest, prediction model.

1. Introduction

In the contemporary landscape where energy security and climate change have emerged as urgent global challenges, the automotive industry is under immense pressure to optimize operational performance and mitigate environmental impacts. Fuel Economy (FE) is not merely a critical technical parameter reflecting vehicle energy efficiency; it also directly influences the operating expenditure (OPEX) for users and the total volume of greenhouse gas emissions [1], [2], [3], [4]. Consequently, the precise determination and forecasting of FE play a pivotal role in enabling manufacturers to optimize vehicle design, while assisting consumers and transport units in planning effective energy management strategies [3], [4].

Technically, vehicle FE is closely related to the complex interactions among physical and operational factors that govern Fuel Consumption (FC) in internal combustion engines. Classical experimental studies, such as the work by Ghani et al. [5] and recent analyses by Tezcan [6], have examined factors related to fuel consumption. Vehicle specifications—such as weight, displacement, horsepower, and acceleration—also contribute to differences in fuel economy. Traditionally, physical calculation methods and static empirical formulas have been the benchmark for estimating FC. Although foundational, these conventional approaches frequently fail to capture the intricate non-linear correlations and multi-collinearity among variables, leading to substantial prediction errors when applied to diverse real-world scenarios [2].

Recently, machine learning (ML) methods have emerged as powerful alternatives for modeling and solving complex automotive engineering problems, including FC and FE prediction [2], [3], [4]. ML models can autonomously learn non-linear relationships directly from historical data and have been successfully applied to deliver predictions with superior accuracy. Several studies have demonstrated the potential of individual ML models and ensemble approaches for estimating vehicle energy efficiency. However, many existing studies rely on limited hyper-parameter tuning or older algorithms, which can significantly constrain model performance and robustness. Moreover, systematic comparisons of optimized modern Gradient Boosting frameworks, combined with robust interpretability methods, remain limited in the literature.

ML is fundamentally based on the principle of learning patterns from data to make predictions about unseen or future data. In predictive tasks within the automotive sector, ML models learn how vehicle specifications and operational variables influence the outcome. More specifically, early studies successfully applied algorithms such as Random Forest (RF) and artificial neural networks (ANN) to predict FC for fleets and heavy-duty vehicles [7], [8]. Recently, the focus has shifted towards highly optimized ensemble models for commercial vehicles. For instance, studies by Canal et al [9], Hamed et al. [10], and Olubusayo et al. [11] investigated fuel consumption prediction using machine learning methods. Furthermore, recent studies by Yoo et al. [3], Keskin et al. [12], and Hussain et al. [13] evaluated various ML models for fuel efficiency, fuel consumption, and vehicle energy-consumption prediction, respectively, demonstrating the potential of modern ML frameworks in automotive energy modeling [14].

In general, while global advancements are evident, specialized research applying advanced

ML algorithms to exploit the operational data of traditional internal combustion engine vehicles remains relatively limited in certain domestic contexts, such as Vietnam, where focus often leans toward hardware optimization or Hybrid strategies [15]. Furthermore, the "black box" nature of ML—specifically the difficulty in explaining why a model generates a particular prediction—remains a significant technical challenge. This study addresses these gaps by developing and rigorously optimizing four state-of-the-art ensemble models: RF, LGB, XGB, and CGB for FE prediction, with MPG as the target variable. A robust dataset comprising detailed vehicle specifications is utilized to train and validate the models. The contribution of this work lies in its systematic comparison of these methods under a common evaluation framework, offering not only improved predictive performance but also addressing the "black box" issue by integrating the SHapley Additive exPlanations (SHAP) technique [16] to quantify physical interpretability. This contributes to the development of efficient, reliable, and transparent tools for modern automotive engineering applications.

Based on these considerations, the objectives of the present study are as follows: (i) to develop and optimize ensemble machine learning models-RF, LGB, XGB, and CGB-for predicting automotive fuel economy (FE), expressed in MPG, using intrinsic technical specifications; (ii) to systematically evaluate and compare the performance of these models using standard statistical metrics; (iii) to interpret model behavior using the SHAP technique to quantify the influence of input variables and examine their consistency with fundamental automotive engineering principles; and (iv) to discuss the applicability and limitations of the optimized models for supporting energy management research.

2. Materials and Methods

2.1. Data Used

This study utilized the UCI Auto MPG dataset [17], a standardized dataset encompassing the technical specifications and fuel economy (FE), expressed in miles per gallon (MPG) of diverse automobile models (https://www.kaggle.com/datasets/uciml/autompg-dataset). Comprising 398 instances, the dataset covers a wide spectrum of vehicle segments—from fuel-efficient compact cars to high-performance sports cars and light-duty trucks—providing a range of vehicle characteristics for comparative analysis. Key technical parameters, such as the number of cylinders, engine displacement, horsepower, vehicle weight, acceleration, and model year, were meticulously recorded for each sample. To guarantee the integrity and accuracy of the subsequent Machine Learning (ML) training phase, a data preprocessing procedure was implemented to remove records containing missing values, resulting in 392 valid samples.

The determination of input variables for FE prediction was rooted in the fundamental principles of internal combustion engines and supported by prominent prior studies [5], [6], [2]. Prior research has investigated relationships between vehicle characteristics, operating factors, and fuel consumption [2], [3]. Notably, vehicle weight (X4) serves as a critical factor directly proportional to rolling resistance and inertia; thus, heavier vehicles demand greater engine work, which inherently increases FC and consequently reduces FE. Likewise, engine displacement (X2) and the number of cylinders (X1) govern the volume of the air-fuel mixture drawn into the combustion chamber, directly influencing heat release and power output [5], [18], [14]. In addition, vehicle attributes like acceleration (X5) and horsepower (X3) were included to adequately reflect engine load characteristics under varying driving conditions [9], [12], [3].

Ultimately, the integration of these specific input features—Cylinders (X1), Displacement (X2), Horsepower (X3), Weight (X4), Acceleration (X5), Model Year (X6), and Origin (X7)—for predicting

the target variable (Y, representing FE in MPG) follows the feature definitions of the Auto MPG dataset [17]. A preliminary statistical analysis of these selected variables, which form the foundation for developing the predictive models, is presented in Table 1.

Table 1. Statistical analysis of the data used

Variable	Mean	Std	Minimum	25%	50% (Median)	75%	Maximum
Y	23.45	7.81	9.00	17.00	22.75	29.00	46.60
X1	5.47	1.71	3.00	4.00	4.00	8.00	8.00
X2	194.41	104.64	68.00	105.00	151.00	275.75	455.00
X3	104.47	38.49	46.00	75.00	93.50	126.00	230.00
X4	2977.58	849.40	1613.00	2225.25	2803.50	3614.75	5140.00
X5	15.54	2.76	8.00	13.78	15.50	17.02	24.80
X6	75.98	3.68	70.00	73.00	76.00	79.00	82.00
X7	1.58	0.81	1.00	1.00	1.00	2.00	3.00

The statistical results presented in the table provide a comprehensive overview of the dataset characteristics used in this research. The target variable Y, representing fuel economy (FE) expressed in miles per gallon (MPG), has a mean value of 23.45 with a standard deviation of 7.81, indicating a moderate level of variation across a range from 9.00 to 46.60. The first independent variable X1, which denotes the number of cylinders, has a mean of 5.47 and a median of 4.00, suggesting that 4-cylinder engines are the most common in the sample within a range of 3.00 to 8.00. Variable X2, representing engine displacement, exhibits a mean of 194.41 and a high standard deviation of 104.64, reflecting significant diversity in engine sizes that span from 68.00 to 455.00. Horsepower, denoted as X3, averages 104.47 with a median of 93.50, and its values show a wide distribution from a minimum of 46.00 to a maximum of 230.00. Vehicle weight, variable X4, is characterized by a mean of 2977.58 and a standard deviation of 849.40, with observations ranging from 1613.00 to 5140.00, which highlights the inclusion of both light and heavy vehicle categories. Variable X5, representing acceleration, maintains a relatively stable distribution with a mean of 15.54 and a standard deviation of 2.76, where the recorded times fall between 8.00 and 24.80 seconds. The model year, X6, shows an average of 75.98 with a standard deviation of 3.68, covering a production period from 70.00 to 82.00. Finally, the origin variable X7 has a mean of 1.58 and a median of 1.00, indicating the geographical distribution of the vehicle manufacturing regions across three distinct categories.

Fig. 1 illustrates the frequency distribution charts of the input variables and the target variable used in the prediction model. A general observation of the charts reveals a diverse variation in technical characteristics among the vehicle samples in the dataset, which reflects the complex reality of the energy problem. Regarding the variables characterizing engine structure, the number of cylinders X1 does not follow a continuous distribution but is concentrated into three discrete value groups of 4, 6, and 8. Among these, the 4-cylinder engine group accounts for the largest proportion, reflecting the common trend of small and medium-sized commercial vehicles aimed at optimizing operating costs. Similar to this trend, engine displacement X2 and engine power X3 both exhibit a distinct right-skewed distribution. Most of the data is concentrated in the low-value range, with X2 under 150 cubic inches and X3 under 100 horsepower; however, there is still a distribution tail

extending toward high values, representing the high-performance sports car or light truck segment.

The vehicle weight variable X4 has a wide distribution range from 1613 to 5140 lbs, with the highest concentration density between 2000 and 2500 lbs. This is an important physical factor that directly affects rolling resistance and movement inertia. The fact that the data range covers both light and heavy vehicles ensures the model can learn the relationship between mass and energy consumption in the most general way. In contrast to the polarization of the structural variables, the acceleration variable X5 shows an approximately bell-shaped distribution, centered around the average value of 15.5 seconds. This pattern describes the distribution of recorded acceleration times within the dataset.

In terms of temporal factors and origin, the model year variable X6 has a fairly even distribution ranging from 1970 to 1982, which covers multiple model years in the historical sample. Finally, the target variable, FE (Y) (MPG) has a slightly right-skewed distribution, with values mainly concentrated between 15 and 30 mpg. Although the value range extends from 9 to 46.6 mpg, the number of super-fuel-efficient vehicle samples with Y greater than 40 is still relatively limited compared to the general level. Overall, the characteristic and diverse distribution of variables X1 to X7 along with the target variable Y indicates that the research dataset provides descriptive variation, offering a basis for building and verifying the effectiveness of machine learning algorithms.

Fig. 2 presents the correlation matrix between the input variables and the target variable used in the prediction model. The correlation values are visualized through a heatmap with a color scale transitioning from light to dark blue, reflecting the degree of linear correlation between pairs of variables according to Pearson coefficients. Through this matrix, clear statistical relationships between the technical characteristics and the target variable Y can be observed. Specifically, vehicle technical characteristics, such as vehicle weight X4, displacement X2, and horsepower X3, exhibit strong negative correlations with fuel economy Y (MPG), with coefficients reaching -0.83, -0.81, and -0.78, respectively. These statistical results are entirely consistent with the mechanical principles discussed by Ghani et al. [5] and Schoen et al. [18], suggesting that heavier vehicles and larger engines generally require more energy, corresponding to lower fuel economy values expressed in MPG. Conversely, the model year variable X6 shows a moderately positive correlation with Y at a coefficient of 0.58, reflecting the trend of technological improvements over time that help newer vehicle models be more fuel-efficient than previous generations [10], [2]. The origin variable X7 has a correlation of 0.57 with Y; since Origin is a categorical code, this Pearson coefficient depends on its numerical encoding and should be interpreted cautiously.

Notably, the correlation matrix also reveals strong multicollinearity among the independent variables. A typical example is the very tight linear relationships between displacement X2 and the number of cylinders X1 and between displacement X2 and vehicle weight X4, with correlation coefficients as high as 0.95 and 0.93, respectively. Such interdependence between these input variables often poses challenges for parameter estimation in traditional linear regression models, reducing prediction stability [2]. This provides an important practical basis for the study to select advanced machine learning algorithms such as Gradient Boosting or RF, which are relatively robust to multicollinearity and capable of exploiting complex non-linear relationships in the data [13], [11], [3].

Overall, the correlation matrix provides deep quantitative insights into the relationship between technical parameters and energy performance, providing descriptive context for interpreting the model results in subsequent sections.

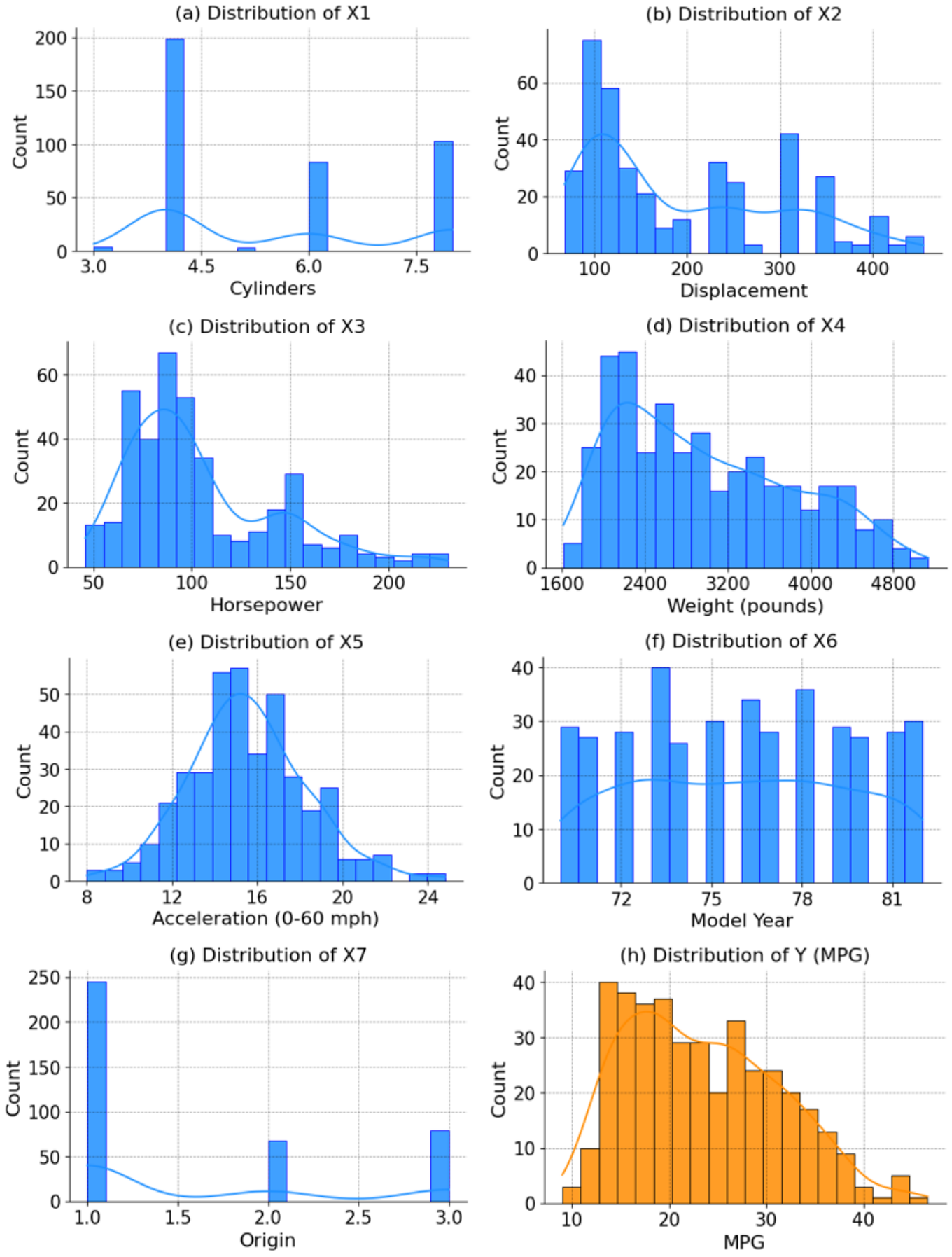


Fig. 1. Frequency distribution of the input variables (X1–X7) and target fuel economy (Y, MPG).

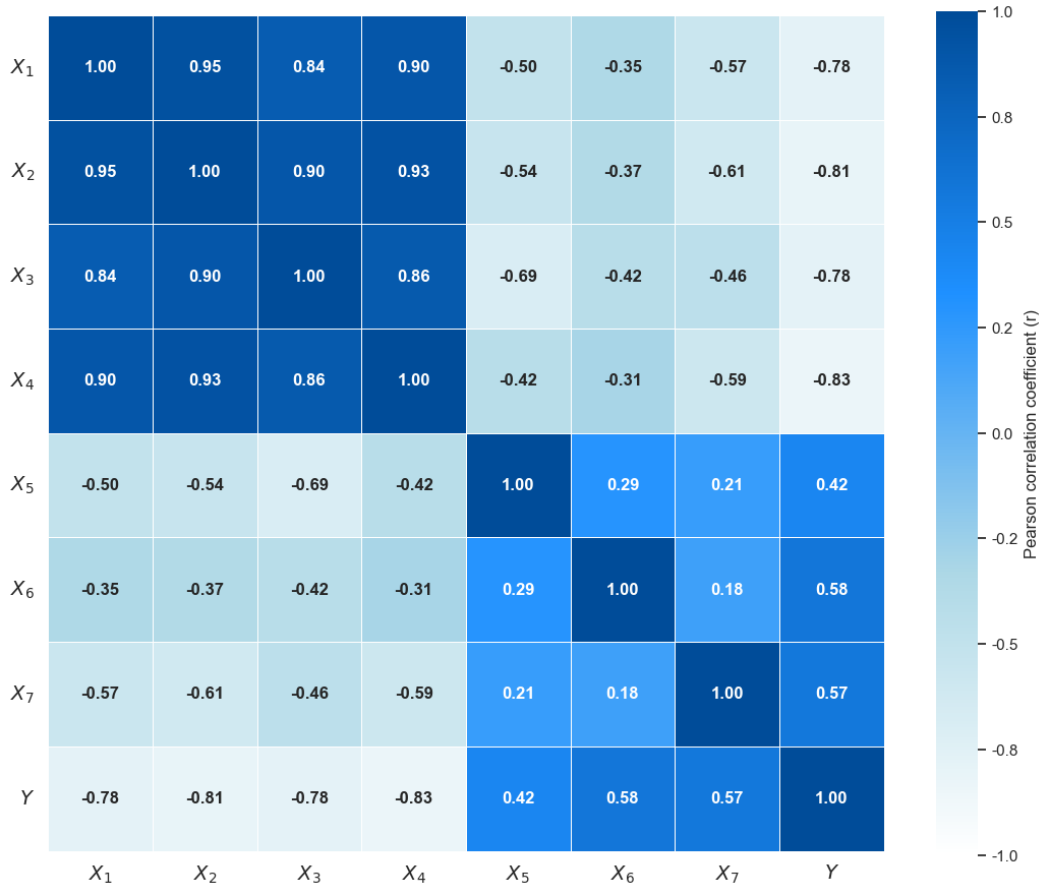


Fig. 2. Pearson correlation matrix between vehicle technical specifications and fuel economy (Y, MPG).

2.2. Methods used

Based on a comprehensive review of previous studies on energy forecasting in transportation [2], fuel economy (FE), expressed in miles per gallon (MPG), is influenced by complex non-linear relationships between engine technical parameters and operating conditions. Traditional regression methods often face limitations in modeling these multi-dimensional interactions. Therefore, this study proposes an approach utilizing Ensemble Learning techniques—specifically combining Bagging and Boosting—to optimize the model's accuracy and generalization capabilities. The specific implementation of the four ensemble algorithms and the optimization strategy are detailed below.

2.2.1. Random Forest (RF)

RF algorithm was selected as the representative model for the Bagging technique [19]. This method operates by simultaneously constructing a large ensemble of independent decision trees on random subsets of the data. The final forecast $\hat{y}$ is synthesized from the average value of all constituent trees:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \quad (1)$$

where B denotes the total number of decision trees in the ensemble, and $f_{b(x)}$ represents the prediction output of the B-th individual tree. This architecture helps minimize variance and significantly mitigates the overfitting phenomenon commonly found in individual decision trees, making it particularly effective for high-noise operational data [7], [8].

2.2.2. Light Gradient Boosting Machine (LGB)

LGB is a gradient boosting algorithm designed for high speed and memory efficiency on large datasets [20]. It builds an ensemble of decision trees sequentially, and can use Gradient-based One-Side Sampling (GOSS) to improve split estimation. The prediction for a sample x_l is expressed as the sum of predictions from k trees

[20]:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{2}$$

where F is the space of regression trees. When GOSS is enabled, LGB focuses on instances with larger gradients to improve training efficiency [20].

2.2.3. Extreme Gradient Boosting - XGBoost (XGB)

XGB is an optimized gradient boosting algorithm renowned for its high performance and scalability [21]. The model builds trees sequentially, where each tree corrects the errors of the previous ensemble. The objective function $\mathcal{L}$ to be minimized combines a loss function l and a regularization term Ω:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3}$$

where y_x is the actual value and $\hat{y}_x$ is the predicted value. The regularization term Ω penalizes model complexity to reduce overfitting [21], [2].

2.2.4. Categorical Boosting (CGB)

CGB is a gradient boosting algorithm designed to handle categorical features efficiently and mitigate prediction shift [22]. Its strong predictive performance and capacity to handle categorical features make it highly suitable for complex regression tasks, such as predicting fuel economy [11]. CGB builds an ensemble of decision trees sequentially, utilizing an ordered boosting mechanism to prevent target leakage and prediction shift. The final prediction $\hat{y}_i$ for a sample x_i after K iterations is expressed as the sum of all additive trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{4}$$

where f_k is the decision tree added at the k-th iteration. When categorical features are configured, CGB can utilize Ordered Target Statistics (OTS) to limit target leakage [22]. For a dataset with a random permutation σ, the expected target statistic for a category k of the i-th instance is calculated strictly using historical data:

where $S_{\{i,k\}}$ is the sum of the target values for preceding instances of the same category, and $C_{\{i,k\}}$ is the count of such instances. They are mathematically defined as follows:

$$\hat{x}_{i,k} = \frac{\sum_{j=1}^{i-1} \left[x_{j,k} = x_{i,k}\right] \cdot y_j + a \cdot p}{\sum_{j=1}^{i-1} \left[x_{j,k} = x_{i,k}\right] + a} \tag{5}$$

In this expression,([x_{j,k}=x_{i,k}]) denotes the indicator function, which equals 1 when(x_{j,k}=x_{i,k}) and 0 otherwise, y_j is the target value (FE, expressed in MPG), p is the prior expected value, and a is the smoothing weight. During the training process, the overall objective function $\mathcal{L}$ *is* minimized by CGB combines the training loss and a regularization penalty:

$$\mathcal{L} = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \Omega(f) \tag{6}$$

where $l(y_i, \hat{y_i})$ represents the loss function measuring the deviation between actual and predicted fuel economy, and Ω (F) the regularization term designed to penalize model complexity, which may reduce overfitting.

2.2.5. K-Fold Cross-Validation

To complement the hold-out evaluation strategy and provide a more robust assessment of model generalization, 5-fold cross-validation (k = 5) was applied to the training subset (80% of the full dataset). In this procedure, the training data was partitioned into five non-overlapping folds of approximately equal size. In each of the five iterations, the model was trained on four folds and evaluated on the remaining fold, ensuring that every sample served as a validation observation exactly once. The mean R^2 across all five folds is reported as the cross-validation score $\bar{R}^2_{CV}$, computed as:

$$\bar{R}^2_{CV} = \frac{1}{k}\sum_{i=1}^{k} R_i^2, \quad k = 5 \tag{7}$$

This approach mitigates the sensitivity of model evaluation to any particular train-test

partition, which is especially important given the moderate size of the dataset (n = 392 complete samples). In this study, cross-validation was integrated into the Grid Search Optimization (GSO) pipeline to guide hyperparameter optimization, allowing candidate configurations to be compared across multiple validation folds.

2.3. Grid search optimization (GSO)

GSO is a systematic approach used to fine-tune hyperparameters to enhance the predictive accuracy of machine learning models [10], [11]. The fundamental principle of this method involves defining a discrete grid of candidate values for each hyperparameter [9], [6]. The algorithm exhaustively searches through every possible combination within this predefined space θ^*, training and evaluating the model performance at each point [7], [14]. To ensure a robust and assessment of performance across training folds, k-fold cross-validation is typically integrated into the grid search process [12], [10]. In this study, grid search is employed to optimize the hyperparameters of the selected algorithms, aiming to minimize a specified loss criterion S; when R^2 is used as the selection score, the corresponding optimization should maximize it. The optimal set of hyperparameters θ^* is identified as follows:

$$\theta^* = \arg\min_{\theta \in GS} S(\theta) \quad (8)$$

where $S(\theta)$ represents the evaluation metric for the model trained with the hyperparameter configuration θ. Although grid search can be computationally intensive, its application in this research is justified by the structured nature of the vehicle fuel economy dataset and the focused hyperparameter space, ensuring an exhaustive and reproducible optimization process [12] , [2]. By identifying the most effective parameter settings, the method supports systematic comparison of candidate hyperparameter settings on the training data.

2.4. Training and Testing Strategy

To evaluate the model's generalization performance and ensure its ability to predict fuel economy (MPG) on unseen data, the collected dataset was partitioned into training and testing subsets using a fixed 80:20 ratio [9], [14]. This specific distribution—comprising 80% of the data for model training and 20% for final validation—is widely recognized in automotive energy modeling for providing a sufficient volume of data to capture complex engine behaviors while maintaining a statistically significant testing set [7], [3].

In this study, a fixed random state (RS=100) was implemented during the splitting process to make the data split reproducible [2]. By utilizing this 80:20 split, the models—specifically XGB, CGB, LGB, and Random Forest—were trained to learn relationships between vehicle parameters and fuel economy [10], [16]. This strategy allows evaluation on the held-out records from the same historical dataset [15], [3]. Min–Max normalization was applied to the input variables and the target variable Y, with the latter scaled to the range [0.1, 1.0]. The scaling parameters were fitted exclusively on the training data and subsequently applied to the testing data. Performance evaluation was conducted using the normalized target values.

2.5. Evaluation of Predictive Performance: Determination Coefficient (R^2), RMSE, MAE, and MAPE

To evaluate the forecasting performance of the models quantitatively and objectively, this research utilizes a set of four standard statistical metrics for regression problems, including the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics not only reflect the degree of fit between the predicted results and experimental data but also demonstrate the models' ability to operate effectively on new datasets, providing complementary measures of prediction performance on the evaluated data [23]. RMSE and MAE were calculated on the normalized target

scale. MAPE was evaluated on the same scale and reported as a decimal fraction in Table 2 and Fig. 3, with equivalent percentage labels in Fig. 4. R^2 was dimensionless. Therefore, the reported MAPE represents relative errors on the normalized target scale rather than the original MPG scale.

Table 2. Comparison of model performance using Grid Search Optimization

Model	Metric	Training Set	Testing Set	Model	Metric	Training Set	Testing Set
XGB	R^2	0.9710	0.9102	LGB	R^2	0.9545	0.9101
	RMSE	0.0324	0.0503		RMSE	0.0406	0.0503
	MAE	0.0243	0.0368		MAE	0.0298	0.0361
	MAPE	0.0606	0.0982		MAPE	0.0724	0.0987
RF	R^2	0.9819	0.9048	CGB	R^2	0.9697	0.9204
	RMSE	0.0256	0.0517		RMSE	0.0331	0.0473
	MAE	0.0175	0.0353		MAE	0.0252	0.0347
	MAPE	0.0412	0.0922		MAPE	0.0628	0.0907

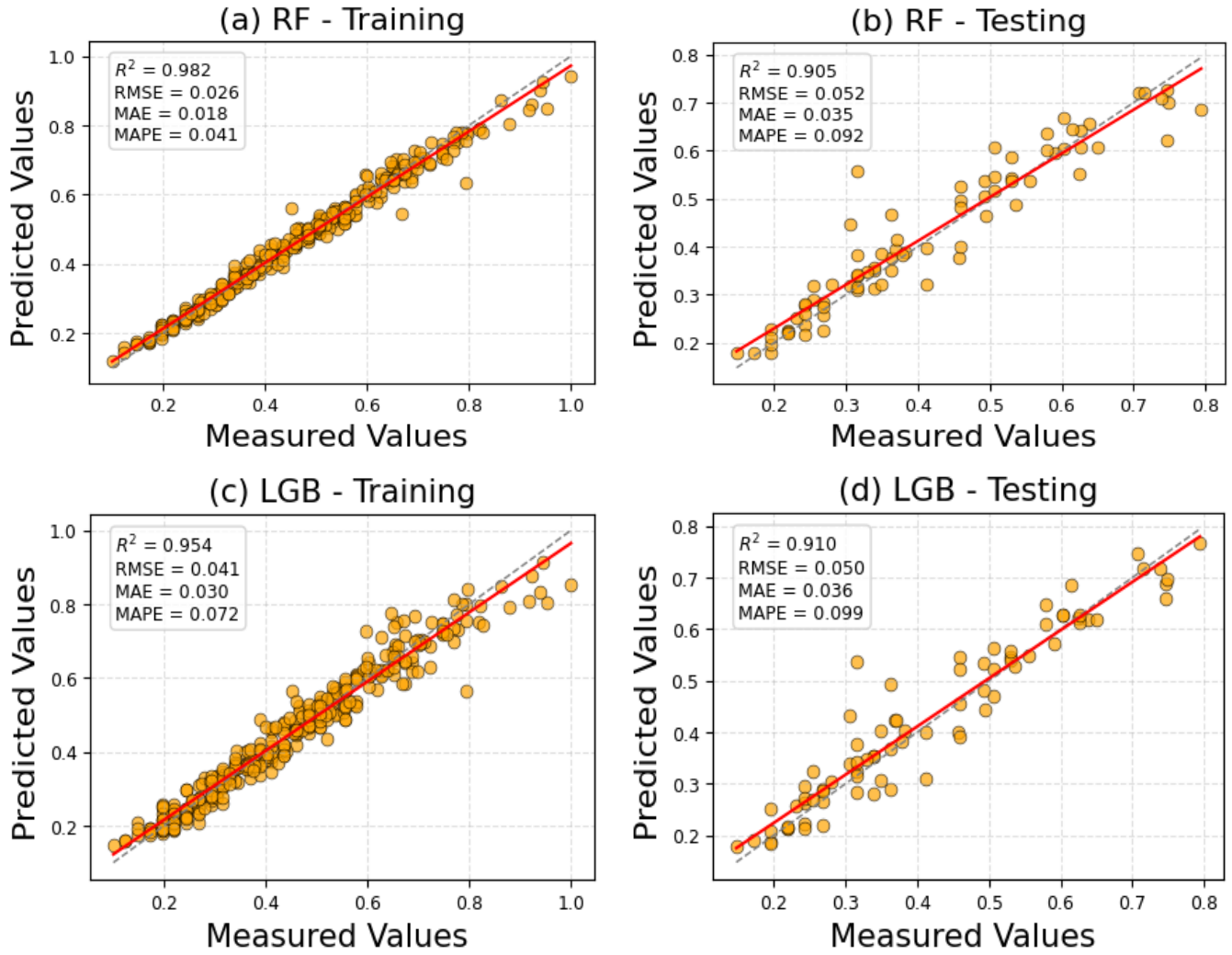


Fig. 3. Scatter plots of actual versus predicted normalized fuel economy: (a) RF training, (b) RF testing, (c) LGBM training, (d) LGBM testing, (e) XGB training, (f) XGB testing, (g) CGB training, and (h) CGB testing datasets.

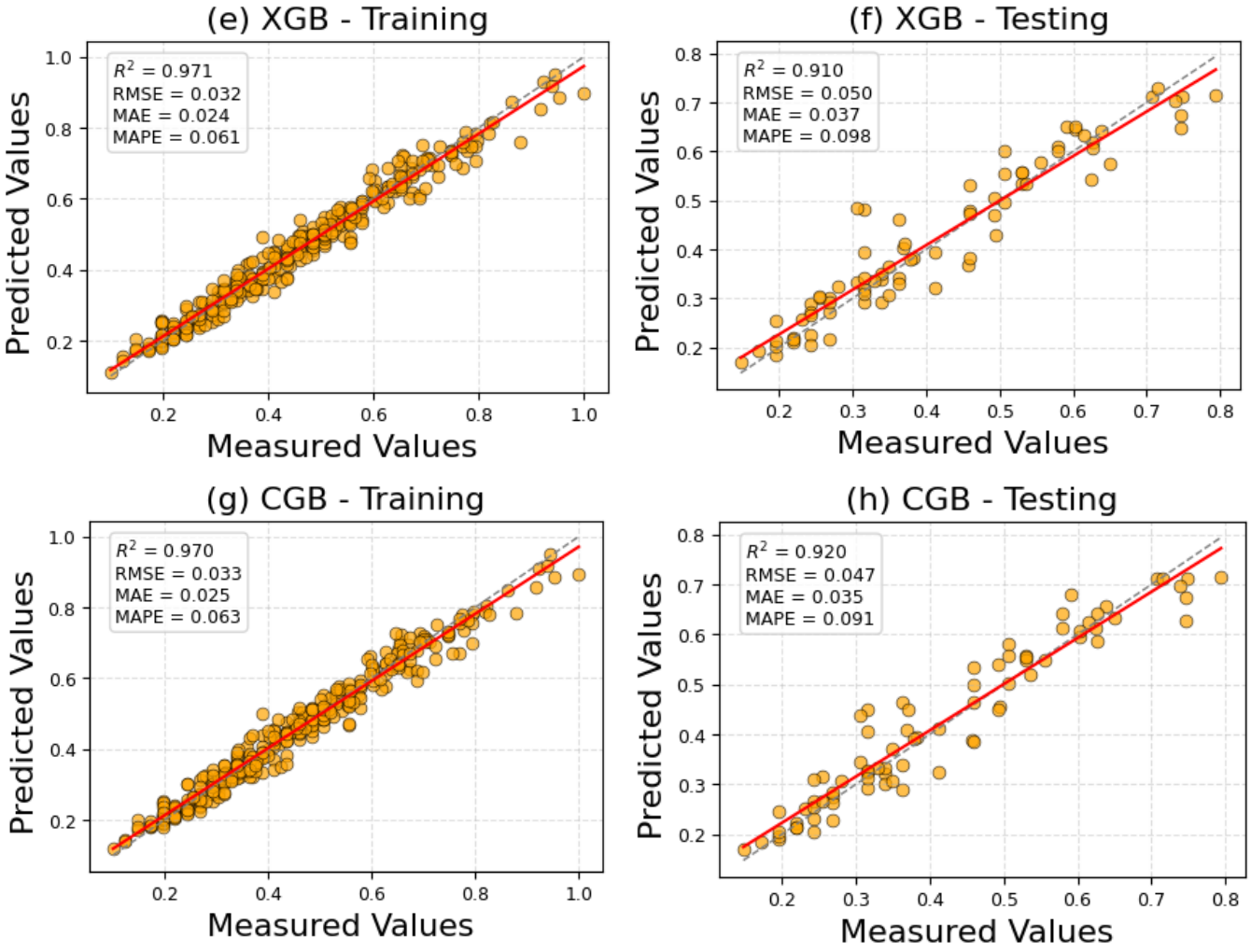


Fig. 3. (continued)

The Coefficient of Determination (R^2) represents the proportion of the variance in the target variable (fuel economy, expressed in MPG) that is explained by the input variables in the model, with a maximum of 1; negative values are also possible. A value closer to 1 indicates that the model fits the actual data better and effectively captures the variation in fuel economy. The formula for calculating the coefficient of determination is expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

In these expressions, y_i denotes the actual observed value and $\hat{y}_i$ represents the predicted value; meanwhile, $\bar{y}$ and $\bar{\hat{y}}$ refer to the mean observed value used in the R^2 denominator.

The RMSE index is employed to quantify the average deviation between predicted and actual values. A distinctive feature of RMSE is its ability to impose a greater penalty on large-scale errors through the squaring operation. In the context of automotive engineering forecasting, monitoring this metric is crucial for identifying instances of severe prediction bias, which could significantly impact transport economic optimization [7]. The formula for calculating RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

A lower RMSE value indicates superior predictive accuracy and a reduced dispersion of errors. The concurrent use of R^2 and RMSE facilitates a more comprehensive evaluation of

model performance, ensuring a balance between overall goodness-of-fit and stability across individual data points [23].

Mean Absolute Error (MAE) provides the average of absolute differences between actual and predicted values. This metric offers an intuitive perspective on error magnitude on the same normalized scale as the target variable and is notably less sensitive to outliers compared to RMSE [23]. The formula for calculating MAE is defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (11)$$

Finally, the Mean Absolute Percentage Error (MAPE) is employed to evaluate the relative accuracy of the forecasts in percentage terms. This metric facilitates a comparison of forecasting performance across different vehicle types with varying fuel economy values, as it expresses prediction errors relative to the observed normalized target values; these percentages are not equivalent to MAPE on the original MPG scale [23]. The formula for calculating MAPE is defined as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (12)$$

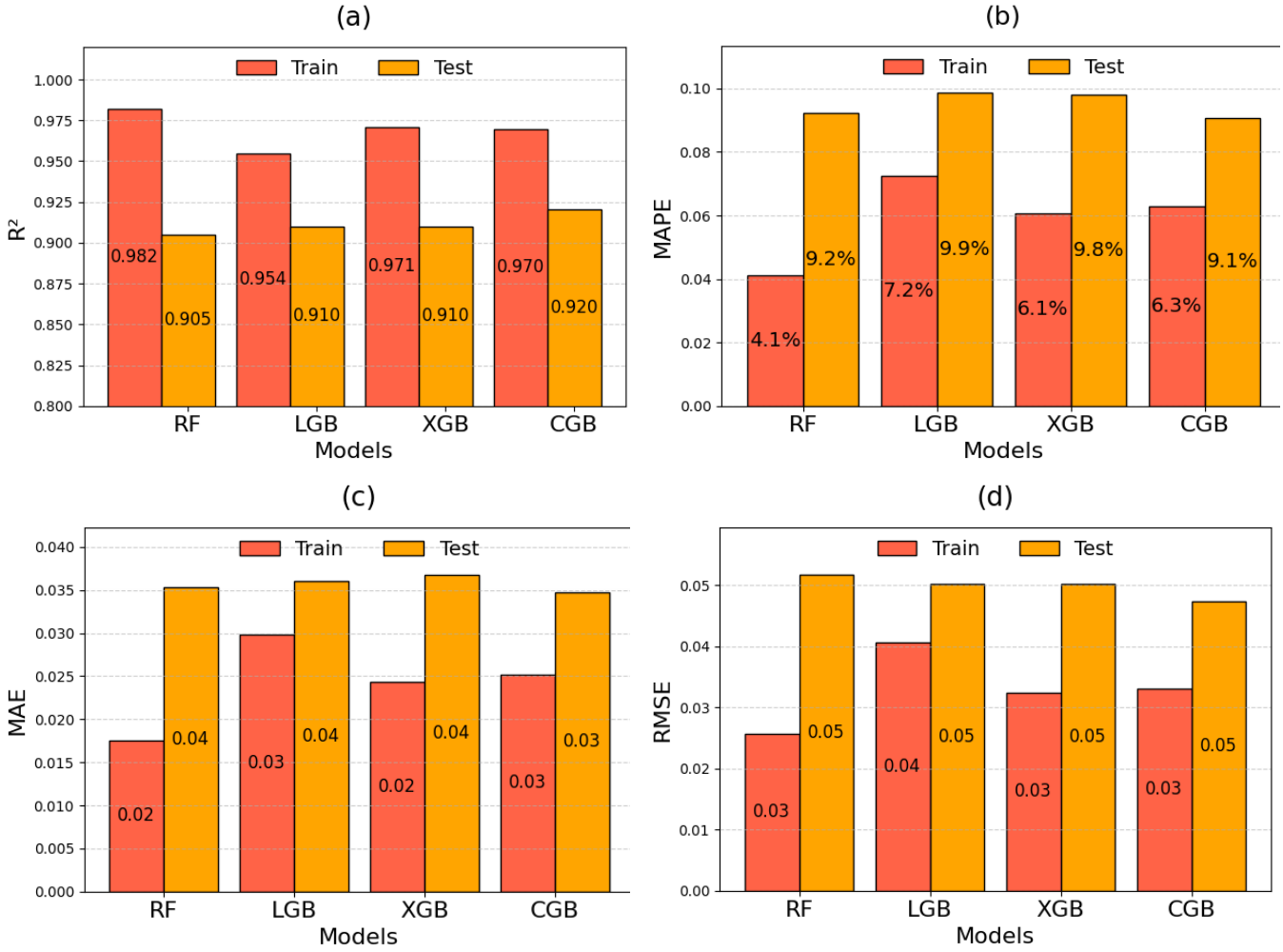


Fig. 4. Comparative bar charts of predictive performance metrics on normalized target values: (a) R^2, (b) MAPE, (c) MAE, and (d) RMSE.

The primary advantage of employing this comprehensive set of metrics is the creation of a multi-dimensional evaluation matrix, which addresses the individual limitations of each index when used in isolation. While R^2 reflects the goodness-of-fit, RMSE effectively controls for outliers and large-scale errors, MAE assesses average practical accuracy, and MAPE provides a

perspective on relative error. Consequently, a simultaneous consideration of these metrics is essential to derive the most accurate conclusions regarding the performance of the machine learning models in this study.

2.6. SHAP (SHapley Additive exPLanations)

While advanced machine learning algorithms such as LGB, XGB, and RF typically yield superior predictive accuracy compared to traditional regression models, they are often perceived as 'black-box' models due to their complex decision-making mechanisms and lack of transparency. This opacity poses challenges in interpreting the underlying causes of prediction results—a critical requirement in automotive engineering, where understanding the specific impact of design parameters on fuel efficiency is essential [5], [2]. To address this limitation, this study adopts the SHAP method, a unified model interpretation technique grounded in cooperative game theory as developed by Lundberg and Lee [16].

The core principle of SHAP involves allocating attribution values for each input variable relative to the prediction output for a specific data instance, ensuring fairness and consistency according to the axioms of Shapley values. Specifically, for a predictive model f and an input sample x, the prediction f(x) is represented as a linear sum of the SHAP values for each feature plus the model's base expected value:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i' \quad (13)$$

In this context, ϕ_0 represents the base expected value (mean prediction) of the model across the entire training dataset, M denotes the number of input features, ϕ_i is the SHAP value of the i-th feature, which signifies the contribution of that specific variable—either positive or negative—to the final prediction relative to the average baseline.

The Shapley value ϕ_i is calculated based on the average change in the predicted output when feature i is added to all possible subsets of other input variables. The general formula is defined as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \quad (14)$$

In this context, F denotes the set of all features, S is a subset of F excluding feature i, and |S| represents the number of elements in subset S. The expression $[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)]$ reflects the marginal contribution of variable i when incorporated into the set S.

In the framework of fuel economy prediction, the application of SHAP offers two fundamental benefits. First, it enables the quantification of the Global Importance of technical parameters—such as vehicle weight, displacement, and engine power—thereby identifying the features most influential in the model predictions [16]. Second, SHAP provides Local Explanations for specific vehicle instances, allowing engineers to understand why a particular model exhibits higher or lower fuel economy than expected based on its unique characteristics. These explanations help interpret model predictions without establishing causal effects [16].

A prominent advantage of SHAP over traditional feature importance methods (such as Gini importance in RF) is its consistency and its ability to indicate the direction of impact (positive or negative) of each variable. However, a limitation of this method is the high computational cost, particularly as the number of input variables increases. To address this, TreeSHAP is an efficient approach for calculating SHAP values in tree-based models such as XGB and LGB [24].

3. Results and analysis

3.1. Comparison of model performance using Grid search optimization

The performance metrics of the four ensemble learning models after the grid search optimization process are summarized in Table 2.

Overall, the results demonstrate that all models achieve a high degree of predictive accuracy, with testing R^2 values exceeding 0.90 (Table 2 and Fig. 3, and Fig. 4), indicating that the models captured substantial variation in the evaluated fuel economy data. Among the implemented algorithms, CGB achieved the best reported performance, achieving the highest testing R^2 of 0.9204 and the lowest testing RMSE of 0.0473. XGB and LGB follow closely, exhibiting nearly identical performance with testing R^2 values of 0.9102 and 0.9101, respectively, and the testing RMSE are 0.0503 and 0.0503, respectively. The training–testing differences should be interpreted in the context of the single hold-out split and five-fold cross-validation.

Table 3. 5-Fold Cross-Validation R^2 scores of the four optimized models on the training set

Model	CV R^2 (k=5)	Test R^2 (hold-out)	R^2 Gap (Train − CV)
CGB	0.8802	0.9204	0.0895
RF	0.8632	0.9048	0.1187
XGB	0.8623	0.9102	0.1087
LGB	0.8548	0.9101	0.0997

In contrast, while the RF model displays the highest training R^2 of 0.9819, it records the lowest R^2 on the testing subset at 0.9048. When examining error magnitudes, the higher testing RMSE value for RF (0.0517) compared to the boosting models indicates a slightly lower precision in capturing extreme variations in mpg, despite it having a marginally lower testing MAPE (0.0922) than XGB and LGB. The more significant decrease from its training to testing performance suggests a tendency toward overfitting, although it remains a highly accurate model. These results support the comparative performance of the boosting-based approaches, particularly CGB and XGB, on the present dataset.

3.2. Cross-Validation Analysis

To further validate the robustness of the optimized models, 5-fold cross-validation was conducted on the training set, with results presented in Table 3. The CV R^2 scores are systematically lower than the hold-out test scores for all models, which may reflect differences among the evaluation partitions. CGB again achieves the highest CV R^2 of 0.880, indicating the highest mean CV score in this comparison. RF records the largest gap between training R^2 (0.982) and CV R^2 (0.863), with a difference of 0.119, providing further quantitative evidence of its tendency to overfit—a finding consistent with the train-test analysis in Section 3.1. In contrast, LGB and XGB demonstrate more stable generalization patterns, with CV R^2 values of 0.855 and 0.862, respectively. Overall, the cross-validation results corroborate the conclusions drawn from the hold-out evaluation, supporting CGB as the best-performing model under the reported evaluation protocol.

3.3. Interpretability Analysis Using SHAP

To demystify the "black box" nature of the high-performing CGB model, SHAP analysis was conducted to quantify the contribution of each technical feature to the predicted fuel economy. Fig. 5a presents the SHAP summary plot, where features are ranked by their global importance, and individual points represent specific samples from the dataset.

Fig. 5b ranks Model Year (X6), Weight (X4), Horsepower (X3), and Displacement (X2) as the four most influential features. The analysis also reveals differences in how features influence the model. Vehicle Weight (X4) and Horsepower (X3) exhibit a predominant negative SHAP impact, where high feature values (indicated by red points)

are densely clustered on the left side of the axis. This suggests a consistent and significant reduction in fuel efficiency as weight and power increase. In contrast, Model Year (X6) displays a strong positive SHAP influence, with newer models (red points) shifting toward the right. This trend underscores the impact of technological advancements on improving miles-per-gallon over time. The density of the points provides further insight into the model's reliability. For variables like Weight (X4), the high concentration of points at extreme SHAP values indicates that the model has learned a pattern within the evaluated dataset. Conversely, for Acceleration (X5), the points are more centrally located and intermixed, signifying a more nuanced or secondary role in predicting fuel economy. In addition, Fig. 5b presents the mean absolute SHAP values, highlighting the relative importance of individual features in predicting fuel economy. Overall, the SHAP results improve the interpretability of the model, without establishing causal relationships.

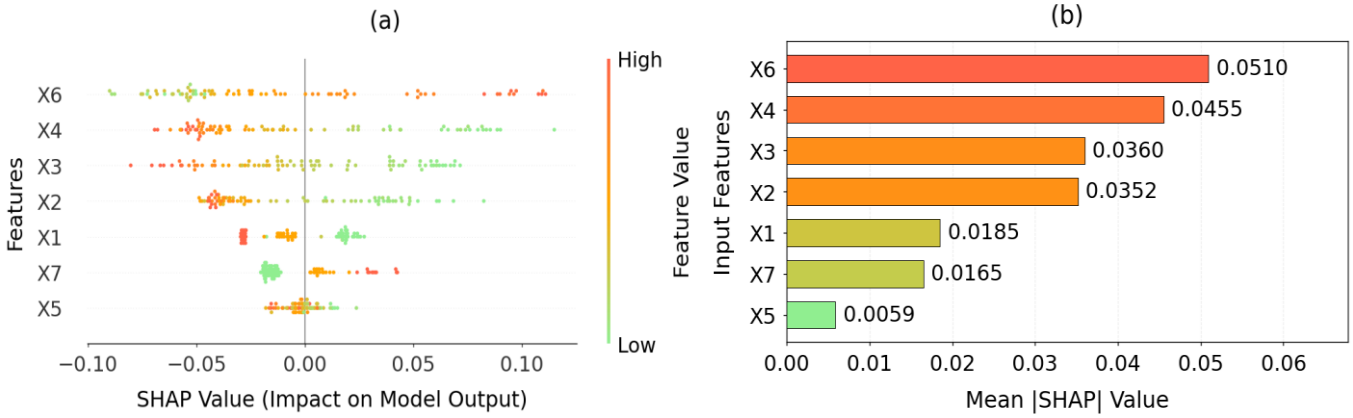


Fig. 5. SHAP-based interpretability analysis of the optimized CGB model: (a) SHAP summary plot of feature impacts and (b) mean absolute SHAP values showing global feature importance.

4. Conclusions

This study presented a systematic comparative evaluation of four ensemble machine learning algorithms— RF, LGB, XGB, and CGB — for predicting automotive fuel economy (FE), expressed in miles per gallon (MPG). A unified modelling framework incorporating GSO -based hyperparameter optimization and five-fold cross-validation was adopted to ensure a fair and robust comparison among the models.

The experimental results demonstrated that all four algorithms achieved high predictive accuracy, confirming the suitability of ensemble learning for modelling the complex nonlinear relationships between vehicle technical specifications and fuel economy. Among the evaluated models, CGB consistently exhibited the best overall performance, achieving the highest testing coefficient of determination (R^2 = 0.9204) together with the lowest prediction errors on the normalized target scale (RMSE = 0.0473, MAE = 0.0347, and MAPE = 9.07%). The results favor CGB in this comparison, although additional evaluation would be necessary to assess generalization and overfitting more fully.

The SHAP-based interpretability analysis further demonstrated that model year, vehicle weight, horsepower, and engine displacement are the dominant factors influencing fuel economy predictions. These findings are consistent with established automotive engineering principles and improve the transparency and interpretability of the developed machine learning models.

The principal contribution of this research lies in providing a rigorous comparative assessment of optimized ensemble learning algorithms under a common experimental framework, together with interpretable model analysis. Rather than proposing a new machine learning algorithm, this study offers a comparative benchmark that may

assist future research in selecting appropriate ensemble models for automotive fuel economy prediction.

This study has certain limitations. The analysis is based on the widely used UCI Auto MPG benchmark dataset representing vehicles manufactured between 1970 and 1982. Consequently, the developed models do not account for modern vehicle technologies such as turbocharging, direct fuel injection, hybrid and electric powertrains, advanced transmission systems, or real-world operating conditions including traffic, road gradient, driving behaviour, payload, ambient temperature, and fuel quality. Future research should validate the proposed framework using larger and more recent automotive datasets while incorporating dynamic operational variables to improve model applicability for contemporary intelligent transportation and energy management systems.

Declaration of Generative AI

During the preparation of this work, the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] F.T. Johora, M.B.A. Topu, M.M. Rahman, K.T. Islam. (2025). Machine Learning-Based CO2 Emission Prediction to Support Sustainable Urban Development. *International Journal of Science and Research Archive*, 17(3), 939-956. doi: 10.30574/ijsra.2025.17.3.3344

[2] D. Zhao, H. Li, J. Hou, P. Gong, Y. Zhong, W. He, Z. Fu. (2023). A Review of the Data-Driven Prediction Method of Vehicle Fuel Consumption. *Energies*, 16(14), 5258. doi: 10.3390/en16145258

[3] S.-r. Yoo, J.-w. Shin, S.-H. Choi. (2025). Machine learning vehicle fuel efficiency prediction. *Scientific Reports*, 15, 14815. doi: 10.1038/s41598-025-96999-0

[4] M.S. Zahid, U. Jamil. (2026). Data-driven machine learning techniques for fuel economy prediction in sustainable transportation systems. *Green Energy and Intelligent Transportation*, 5(1), 100303. doi: 10.1016/j.geits.2025.100303

[5] M.U. Ghani, M.U. Farooq, S.S. Khan, N. Khan, M. Ali. (2018). Factors Effecting Fuel Consumption of SI Engine - A Case Study. *Professional Trends in Industrial and Systems Engineering (PTISE), 2018, pp. 602-610.*

[6] T.A. Tezcan. (2024). Determining the Key Factors and Estimation of Fuel Consumption in Cold Chain Logistics: A Machine Learning Approach. *Politecnico di Torino*, 2024. [Online]. https://webthesis.biblio.polito.it/32000/

[7] F. Perrotta, T. Parry, L.C. Neves. (2017). Application of machine learning for fuel consumption modelling of trucks. *2017 IEEE International Conference on Big* Data, pp. 3810-3815. doi: 10.1109/BigData.2017.8258382

[8] S. Wickramanayake, H.M.N.D. Bandara. (2016). Fuel consumption prediction of fleet vehicles using Machine Learning: A comparative study. *2016 Moratuwa Engineering Research Conference (MERCon), pp. 90-95.* doi: 10.1109/MERCon.2016.7480121

[9] R. Canal, F.K. Riffel, G. Gracioli. (2024). Machine learning for real-time fuel consumption prediction and driving profile classification based on ECU data. *IEEE Access*, 12, 68586-68600. doi: 10.1109/ACCESS.2024.3400933

[10] M.A. Hamed, M.H. Khafagy, R.M. Badry. (2021). Fuel Consumption Prediction Model using Machine Learning. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(11), 406-414. doi: 10.14569/IJACSA.2021.0121146

[11] O. Olubusayo, O. Patrick, I. Musibau, A. Adepeju, O. Olufunke, D. Ariyo, O. Oluwafolake, A. Mutahir. (2025). Fuel

Consumption Prediction of Vehicles Using Machine Learning Algorithm. *Journal of Computing and Social Informatics*, 4(2), 28-36. doi: 10.33736/jcsi.8262.2025

[12] R. Keskin, E. Belge, S.H. Kutoglu. (2025). Optimizing Fuel Consumption Prediction Model Without an On-Board Diagnostic System in Deep Learning Frameworks. *Sensors*, 25(22), 7031. doi: 10.3390/s25227031

[13] I. Hussain, K.B. Ching, C. Uttraphan, K.G. Tay, A. Noor. (2025). Evaluating machine learning algorithms for energy consumption prediction in electric vehicles: A comparative study. *Scientific Reports*, 15, 16124. doi: 10.1038/s41598-025-94946-7

[14] J. Gong, J. Shang, L. Li, C. Zhang, J. He, J. Ma. (2021). A Comparative Study on Fuel Consumption Prediction Methods of Heavy-Duty Diesel Trucks Considering 21 Influencing Factors. *Energies*, 14(23), 8106. doi: 10.3390/en14238106

[15] S.Đ. Nguyen, T.L. Tran, Q.T. La. (2021). Research for optimizing control strategy of a parallel hybrid drivetrain in vehicle. *Journal of Marine Science and Technology*, Special issue, 10/2021, 195-201. (in Vietnamese)

[16] S.M. Lundberg, S.-I. Lee. (2017). A Unified Approach to Interpreting Model Predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768-4777.*

[17] R. Quinlan. (1993). Auto MPG [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5859H

[18] A. Schoen, A. Byerly, B. Hendrix, R.M. Bagwe, E.C. dos Santos, Jr., Z. Ben Miled. (2019). A Machine Learning Model for Average Fuel Consumption in Heavy Vehicles. *IEEE Transactions on Vehicular Technology*, 68, 6343-6351. doi: 10.1109/TVT.2019.2916299

[19] L. Breiman. (2001). Random Forests. *Machine Learning*, 45, 5-32. doi: 10.1023/A:1010933404324

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3149 – 3157.*

[21] T. Chen, C. Guestrin. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794*. doi: 10.1145/2939672.2939785

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin. (2018). CatBoost: Unbiased Boosting with Categorical Features. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 6639 – 6649.*

[23] H. Khoshvaght, R.R. Permala, A. Razmjou, M. Khiadani. (2025). A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction. *Journal of Environmental Chemical Engineering*, 13(6), 119675. doi: 10.1016/j.jece.2025.119675

[24] S.M. Lundberg, et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56-67. doi: 10.1038/s42256-019-0138-9