scientific reports

Check for updates

OPEN

Machine learning vehicle fuel efficiency prediction

So-rin Yoo[1], Jae-woo Shin[2] & Seoung-Ho Choi[3]✉

To address the challenges associated with fuel consumption in vehicles with low fuel efficiency, several factors must be recognized. Identifying the key factors of fuel efficiency prediction is crucial for making accurate decisions. Therefore, we propose a comprehensive framework that uses machine learning to predict fuel efficiency by integrating various vehicle information. The proposed method comprises a predictive model and analysis framework utilizing key vehicle attributes, such as fuel type, engine displacement, and vehicle grade, to enhance prediction accuracy. We conducted a comparative study using six machine-learning models. To evaluate the machine learning model, MSE (Mean Square Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R-squared (R^2 Score) were used. We experimented with SHAP(Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and odds ratio analysis to evaluate the impact of various factors on fuel efficiency. We confirmed that the proposed method can predict fuel efficiency. Extra Trees Regressor and Random Forest Regressor demonstrated high prediction accuracy, particularly excelling in capturing nonlinear relationships. We also underscore the importance of identifying markers to support decision-making, offering critical insights into the key factors impacting fuel efficiency predictions.

Keywords Vehicle Fuel, Fuel consumption, Machine learning, Fuel Marker, Fuel Framework

The world's primary energy supply has steadily increased along with global economic growth, supplying approximately 14.3 billion TOE (Ton of Oil Equivalent) as of 2018. Korea is the 8th largest energy-consuming country in the world. The government designs and manages efficiency management equipment to achieve economic growth and reduce energy consumption. Among these, a tire energy consumption efficiency rating system assigns energy consumption efficiency ratings to tires, enabling tire buyers to purchase tires with high efficiency, while establishing and managing minimum consumption efficiency standards to prevent the spread of low-efficiency tires. This system was first implemented in Korea in 2012. The tire's energy consumption efficiency rating is graded from 1 to 5 by measuring the rolling resistance coefficient (RRC) according to the "Regulations on energy consumption efficiency measurement and rating standards and marking of automobile tires" and setting the rolling resistance coefficient (RRC) range. In the case of passenger car tires, high-efficiency tires accounted for only 2.8% of the total market at the beginning of the system's implementation (2012), but their share expanded to 8.5% in 2020 due to continuous technological development by manufacturers and improvement in consumer awareness. However, low-efficiency tires still account for 47.3% of the total tire market, securing a market that is more than five times that of high-efficiency tires[1]. As highlighted in recent studies, the efficient use of fuel is not only vital for economic efficiency but also for environmental protection[2,3]. In particular, as energy conservation and greenhouse gas emission reduction have emerged as important tasks worldwide, accurate vehicle fuel efficiency predictions are essential for effective decision-making in environmental conservation and economic management. Fuel efficiency prediction provides individual benefits, such as reduced fuel costs for vehicle operators, and significant advantages for the environment and national economy.

An effective predictive model for fuel efficiency offers multiple benefits to stakeholders: consumers can save fuel costs, dependency on imported energy decreases, and governments and corporations can utilize this data to develop eco-friendly policies or high-efficiency vehicles. By reducing greenhouse gas emissions, such a model can also help mitigate climate change.

Although existing research often addresses factors like engine performance, aerodynamics, or driving habits in isolation, this study aims to integrate these diverse factors within a holistic framework, which is crucial for generating precise and actionable predictions. Through this comprehensive approach, we aim to reduce the prevalence of low-efficiency vehicles and achieve a greater positive environmental impact. If a comprehensive approach is not possible, the vehicle's optimal fuel efficiency cannot be achieved, and inefficient fuel-efficient

[1]Department of AI Application, Hansung University, 116, Samseongyo-ro 16-gil, Seongbuk-gu, Seoul 02876, Republic of Korea. [2]Department of IT Business Administration, Hanshin University, Gyeonggi Province 18101, Republic of Korea. [3]College of Liberal Arts, Faculty of Basic Liberal Art, Hansung University, 116, Samseongyo-ro 16-gil, Seongbuk-gu, Seoul 02876, Republic of Korea. ✉email: jcn99250@naver.com

vehicles will continue to be produced, leading to environmental issues and excessive fuel consumption. Therefore, it is necessary to comprehensively analyze various vehicle information to predict more precise and accurate fuel efficiency. Key variables include manufacturer/importer, engine displacement, vehicle model, fuel type, combined mode CO_2 emissions, tire inches, rolling resistance coefficient (RRC), vehicle type, transmission type, vehicle grade, combined fuel efficiency, etc. Fuel efficiency is predicted by learning a regression model based on this comprehensive data, and we need a system that predicts more accurate fuel efficiency information to users.

In particular, by analyzing the impact of vehicle manufacturers and importer information on fuel efficiency, it should be possible to predict customized fuel efficiency that reflects the characteristics of each manufacturer. In addition, displacement and fuel type are major variables that directly affect fuel efficiency, and it is necessary to improve the accuracy of prediction models by carefully considering these variables.

In summary, our study seeks to develop an accurate and practical predictive model by integrating diverse vehicle information and providing valuable insights into fuel efficiency determinants, ultimately supporting users and policymakers in making informed and eco-friendly choices. Therefore, our primary contributions are summarized as follows.

- First, we propose a comprehensive framework that integrates public data with additional curated features, to support a holistic approach to fuel efficiency prediction.
- Second, we conduct statistical analyses to evaluate the relationships between various vehicle attributes and fuel efficiency, which enhances the reliability of the findings and provides actionable insights.
- Third, we propose a machine learning-based model for vehicle fuel economy prediction, identifying key factors to improve prediction accuracy.
- Fourth, SHAP and LIME are employed to interpret the model's decision-making process, clarifying how key vehicle characteristics contribute to fuel efficiency predictions. This interpretative analysis provides transparency and supports robust decision-making.
- Fifth, we perform an odds ratio analysis to quantify the impact of specific vehicle characteristics on fuel efficiency, identifying markers that significantly influence prediction outcomes.

The structure of this paper is as follows. Section "Related works" describes existing related research. Section "Our proposal" describes the proposed method. Section "Experiment methods" describes the experimental method for the project. Section "Experiment results" describes the experiment results. Section "Discussion" describes the discussion. Section "Conclusions" concludes the paper with conclusions and future research.

Related works

Shim et al.[1] divided the process of predicting power demand on weekends, when the demand patterns are irregular, into variable selection, hyperparameters that optimize the parameters of the prediction model, and error calculation. To select variables that have a large influence on the prediction process and have a low linearity with other variables, the influence of the variables was measured using the SHAP (Shapley Additive Explanations) technique. The similarity between variables was measured using the Pearson Correlation Coefficient technique. Through this, variables to be used were selected and hyperparameter optimization was performed using grid search techniques and the XGBoost model[4].

To analyze changes in automobile fuel efficiency according to tire energy consumption efficiency ratings, Noh et al.[5] selected four tires for passenger cars and conducted the RRC-rolling resistance-automotive fuel efficiency test. The changes in the car fuel efficiency according to tire rolling resistance were confirmed. In addition, the effectiveness of using high-efficiency tires with low rolling resistance was verified, proving the importance of tire RRC value on changes in fuel efficiency. Tire rolling resistance coefficient (RRC) refers to the energy lost due to frictional heat generated as the tire rolls and can be calculated as the ratio of the vehicle's running resistance and the load applied to the tire.

Lee et al.[6] conducted research on powertrain operation and simulation for hybrid vehicle prediction, presenting the function and simulation analysis of powertrain components for predicting vehicle fuel efficiency and performing fuel efficiency prediction simulation by configuring an IONIQ HYBRID vehicle based on ADVISOR (Advanced Vehicle Simulator). First, we analyzed the characteristics of the main elements that make up the powertrain. The engine's fuel efficiency characteristic curve and the motor's efficiency map were presented to reproduce the driving characteristics necessary for predicting fuel efficiency. Next, we explained how to implement a hybrid powertrain using the ADVISOR simulator. The operation of powertrain components to achieve target speed and torque was reproduced using a backward calculation method. Finally, an ADVISOR model was built reflecting the actual specifications of the IONIQ hybrid vehicle, and a fuel efficiency prediction simulation was performed using a standard driving cycle. The fuel efficiency prediction accuracy was reviewed by comparing simulation results and official fuel efficiency data. Through this, a powertrain modeling and simulation methodology for predicting the fuel efficiency of hybrid vehicles was proposed, and an application case for IONIQ hybrid vehicles was presented.

Kim et al. proposed a data correction technique using Generative Adversarial Network (GAN)[7] to predict the fuel efficiency of fuel cell vehicles. We attempted to predict fuel efficiency based on vehicle driving history and detect factors that cause low fuel efficiency. We confirmed an issue where deviations occurred in fuel efficiency data even under the same driving conditions. To solve this problem, a GAN[7]-based data correction technique was proposed. Based on driving data of actual fuel cell vehicles, linear regression[8], GBM (Gradient Boosting Machines)[9], XGBoost[10], and SVM (Support Vector Machines)[11] were applied to compare and analyze fuel efficiency prediction performance. The data correction technique using GAN[7] showed superior fuel efficiency prediction performance compared to other models. Through this, it was confirmed that deviations in actual measurement data could be effectively resolved[12].

Rho et al. presented a causal relationship between driving distance and fuel efficiency. To determine whether the cumulative mileage negatively affects fuel economy, the mileage was determined using the results of the fuel efficiency test of eight test vehicles over four years and the statistical analysis. Fuel efficiency was measured by analyzing the causal relationship between mileage and fuel consumption. The test vehicle was selected considering key specifications and sales volume. A normality test was conducted to determine whether the fuel efficiency data of the four groups had a normal distribution according to the independent variable, accumulated mileage. Additional analysis was performed using a paired t-test. In conclusion, it is difficult to say that the accumulated mileage of a vehicle affects fuel efficiency. It was confirmed that increasing the mileage of a vehicle in a new car condition does not worsen fuel efficiency and has a positive effect on improving fuel efficiency[13].

Kim et al. researched improving fuel efficiency and reducing exhaust gases in automobiles. It was said that improving the fuel efficiency of automobiles includes increasing the engine efficiency, using lightweight materials, improving the aerodynamic design, and introducing hybrid and electric vehicle technology. This paper explains the principles and application methods of each technology and verifies their effectiveness through experiments and data analysis. The use of lightweight materials is effective in reducing fuel consumption by reducing vehicle weight, and the use and application of advanced materials are explained[14].

Kwon et al. compared the emission and fuel efficiency performance of a 2.0-liter LPG hybrid engine and a vehicle. The researchers compared the gasoline system and LPG conversion system using the 2.0-liter Nu engine of the 2021 K5 hybrid vehicle. When LPG fuel was used by installing a hybrid engine in the dynamometer, the highest output and maximum torque were measured at an equivalent level of less than 1% compared to conventional gasoline fuel. In actual tests, fuel efficiency was measured through FTP (Federal Test Procedure) -75 and HWFET driving modes, and the fuel efficiency of LPG hybrid vehicles was 22.7% lower than that of gasoline vehicles. However, considering the average selling price of fuel, LPG fuel is 41.2% cheaper than gasoline, so consumers could achieve a cost savings of about 18.5% when injecting fuel for the same amount. In conclusion, it was found that although LPG hybrid technology is lower than gasoline in terms of fuel efficiency, it has economic advantages and through this, consumers can enjoy cost savings[15].

Kim et al. analyzed the impact of oil pump power loss on the fuel efficiency of mild hybrid vehicles. For this purpose, the study was conducted by modeling a virtual mild hybrid vehicle and assuming hydraulic lines. The oil pump loss was analyzed for fuel consumption reduction in a vehicle model with only the timeless ISG (Idle Stop&Go) function and a vehicle model using a mechanical oil pump and an electric oil pump. As a result of the simulation, compared to the model that did not consider the oil pump loss, the model that considered the mechanical oil pump loss improved fuel efficiency by about 7%, and the model that additionally considered the electric oil pump also showed a similar improvement in fuel efficiency. It was confirmed that the power loss of the oil pump has a significant impact on fuel efficiency and that the power loss of the mechanical oil pump,in particular, accounts for a large portion[16].

Jo et al. studied a method to predict the fuel efficiency of large vehicles using RDE (Real Driving Emissions) data. By predicting fuel efficiency through simulation, researchers have presented a method to reduce costs and obtain highly reproducible results under various driving conditions. In particular, to comply with RDE regulations, we proposed a method to collect actual road driving data using equipment such as PEMS (Portable Emissions Measurement System) and accurately calculate the fuel efficiency of large trucks based on this. CRUISE software was used to simulate the fuel efficiency of large trucks, and the fuel efficiency map was calibrated to reflect actual road data[17].

Katreddi et al. conducted a study focused on predicting fuel consumption in medium-duty vehicles by leveraging Artificial Neural Networks (ANN)[18]. Their model used a minimal set of input variables, including engine load, engine speed, and vehicle speed, which were collected during real-life driving conditions. This approach allows the model to capture realistic fuel consumption patterns without relying on extensive or complex datasets. This study compared the performance of ANN[18] with other regression models, such as linear regression and random forest, demonstrating that ANN[18] achieved superior predictive accuracy, with significantly lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values. The authors concluded that ANN[18]'s ability to model non-linear relationships made it particularly suitable for fuel consumption prediction in dynamic driving environments. This study emphasizes the potential of ANN[18] in handling complex relationships in-vehicle data, reinforcing the effectiveness of machine learning methods for practical applications in fuel economy studies[19].

In their comprehensive review, Katreddi et al. explore the application of artificial intelligence in enhancing fuel consumption prediction, emissions estimation, and predictive maintenance within the heavy-duty trucking industry. They highlight that AI (Artificial Intelligence) techniques, such as machine learning and deep learning, offer promising advancements in addressing fuel efficiency, emissions reduction, and fleet management, which are critical areas for sustainability and cost reduction. This paper emphasizes the importance of analyzing various parameters, such as engine load and speed, to develop accurate predictive models for fuel consumption, as well as exploring methods for maintenance forecasting and route optimization for fuel savings. The authors also identify gaps in existing AI applications, notably the need for improved data availability and the incorporation of diverse environmental and operational factors to enhance prediction accuracy. These findings underscore the value of integrating AI to meet the trucking sector's evolving environmental and economic demands[20].

Thejovathi et al. conducted a study to compare the performance of XGBoost[10] and Gradient Boosting[21] models in FMCG (Fast-Moving Consumer Goods) demand forecasting. The study used two powerful ensemble learning techniques to improve the accuracy of demand forecasting and applied feature extraction techniques to improve the prediction performance of the models. The predictions were made using two ensemble learning algorithms and feature extraction techniques such as PCA (Principal Component Analysis)[22] and RFE (Recursive Feature Elimination)[23]were applied to extract important information from the data. The results showed that both models

showed high accuracy and reliability, but the Gradient Boost model performed particularly well in maximizing the prediction error, recording a better RMSE value[24].

In previous studies, various approaches have been proposed for predicting vehicle fuel efficiency, but many have limitations. Shim et al.[1] proposed a method for predicting power demand using machine learning, but it focused mainly on optimizing prediction models through variable selection and hyperparameter optimization, without providing a direct mechanism for improving the interpretability of fuel efficiency prediction. While the use of SHAP for variable influence is insightful, the method does not fully address the importance of specific vehicle characteristics such as tire efficiency or fuel type in real-world applications. In addition, Lee et al.[6] modeled hybrid vehicle performance to predict fuel efficiency, but their approach was confined to simulations and did not extend to real-world driving conditions or take into account diverse fuel types and tire specifications. This limitation reduces the practical relevance of their findings for everyday vehicle users and policy-makers aiming for generalized solutions. Furthermore, studies such as Kim et al.[12] and Kwon et al.[15] explored specific areas like fuel cell vehicle efficiency and LPG hybrid performance. However, these models either focus on single vehicle types or specific technological solutions, leaving gaps in understanding how various factors interact to affect fuel efficiency in a broader context. Existing studies have significantly advanced our understanding of fuel efficiency prediction but still present limitations. Many approaches either lack a comprehensive view of all the factors influencing fuel economy or focus too narrowly on one or two attributes. These limitations highlight the need for an approach that considers a broader spectrum of vehicle features and integrates machine-learning techniques for improved accuracy and decision-making. Our study addresses these gaps by combining diverse vehicle attributes in a predictive framework and employing advanced machine learning models, including SHAP and LIME, to identify key biomarkers for more accurate fuel efficiency predictions. By doing so, we offer a more holistic, data-driven solution that is applicable across a wide range of vehicles and driving conditions, contributing to more informed policy-making and better consumer choices.

Our proposal

As illustrated in Fig. 1, we propose a comprehensive methodology for predicting vehicle fuel efficiency, structured around four key components: data analysis, a customized dataset, a fuel efficiency prediction pipeline, and marker analysis for identifying critical factors in fuel efficiency prediction.

First, data analysis is conducted to explore relationships and trends within the dataset. This involves Correlation Coefficient Analysis (CCA) and correlation matrix generation, which help identify statistically significant associations between variables, providing insights into potential predictors of fuel efficiency.

Second, we constructed a customized dataset by integrating data downloaded from a public data portal with additional custom data, such as tire size and rolling resistance coefficient (RRC). During the data preparation stage, rows with missing values were removed to maintain data integrity. This dataset combines standardized public data with manually collected engineering variables, offering in-depth insights into vehicle efficiency.

The dataset we used is a combination of characteristics taken from the Korea Energy Agency's fuel economy labeling system (version 2023-12-31)[25] and our research. The KEA's dataset contains the required vehicle

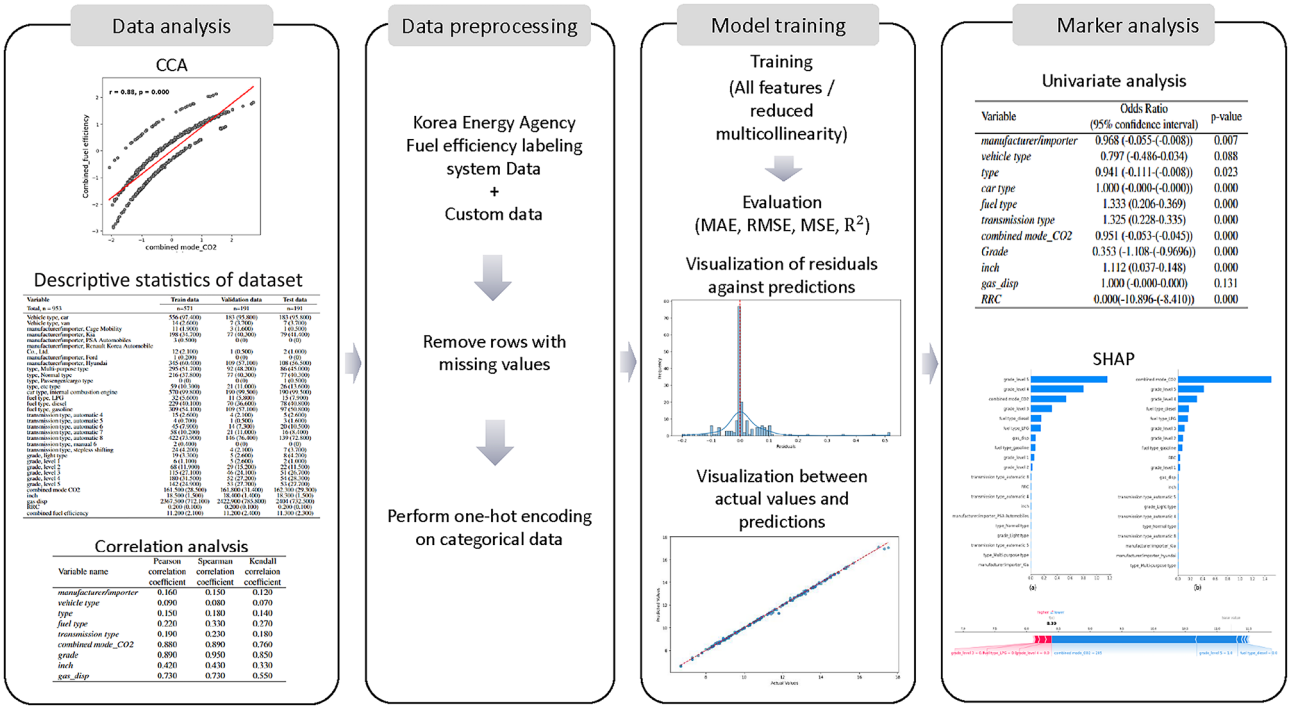

Fig. 1. Our Proposal.

characteristics such as model name, manufacturer/importer, vehicle type, model year, fuel type, transmission type, combined mode CO_2, rating, displacement, and combined fuel economy, and the total number of data is 3358. We also incorporated tire size and RRC, which we researched and calculated manually. The number of data in the public data portal was 3358, but the amount of data collected directly (RRC, tire size) was small, so a total of 953 data were completed in the integration process.

The dataset was divided into three: training set (60%, n = 571), validation set (20%, n = 191), and test set (20%, n = 191), as shown in Fig. 2. For categorical data, we used OneHotEncoder to perform data preprocessing and finally converted the transformed data into pandas DataFrame.

$$RRC = \frac{F_{\mathrm{rr}}}{F_n} \tag{1}$$

Equation (1) was used to calculate the rolling resistance coefficient for each vehicle. F_rr is Rolling Resistance Force, which refers to the resistance generated by friction with the road as the tire rotates, and F_n is Normal Force, which refers to the load acting perpendicular to the tire. In general, the rolling resistance of a tire is determined by the tire structure, tire air pressure, friction coefficient of the road surface, and speed.

Third, the fuel efficiency prediction pipeline involves additional data preprocessing and model training. Categorical variables (such as manufacturer and vehicle type) are transformed into numerical features using one-hot encoding to prepare the data for model input. The model training stage applies six regression models-Linear Regression[8], Extra Trees Regressor[26], Random Forest Regressor[27], Gradient Boosting Regressor[21], Hist Gradient Boosting Regressor[28], and AdaBoost Regressor[29]-implemented with the Scikit-learn library. These models are evaluated based on metrics such as MSE, RMSE, MAE, and R^2 to identify the model with optimal predictive performance for fuel efficiency.

Lastly, Marker analysis identifies key variables that significantly impact fuel efficiency. Using SHAP, LIME, and Odds Ratio analysis, we assess the statistical importance and predictive power of each feature. This analysis helps prioritize influential variables, which can serve as essential markers in predicting fuel efficiency.

This methodology is designed to improve both the accuracy and interpretability of vehicle fuel efficiency predictions. By leveraging a customized dataset and carefully selected predictive variables, our approach aims to enhance predictive performance and provide meaningful insights into vehicle efficiency optimization.

Table 1 is an explanation to help you understand each variable. RRC (Rolling Resistance Coefficient) is a coefficient that represents the rolling resistance that occurs when a tire comes in contact with the road surface and is an important indicator for evaluating the efficiency of a tire.

Experiment Methods

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

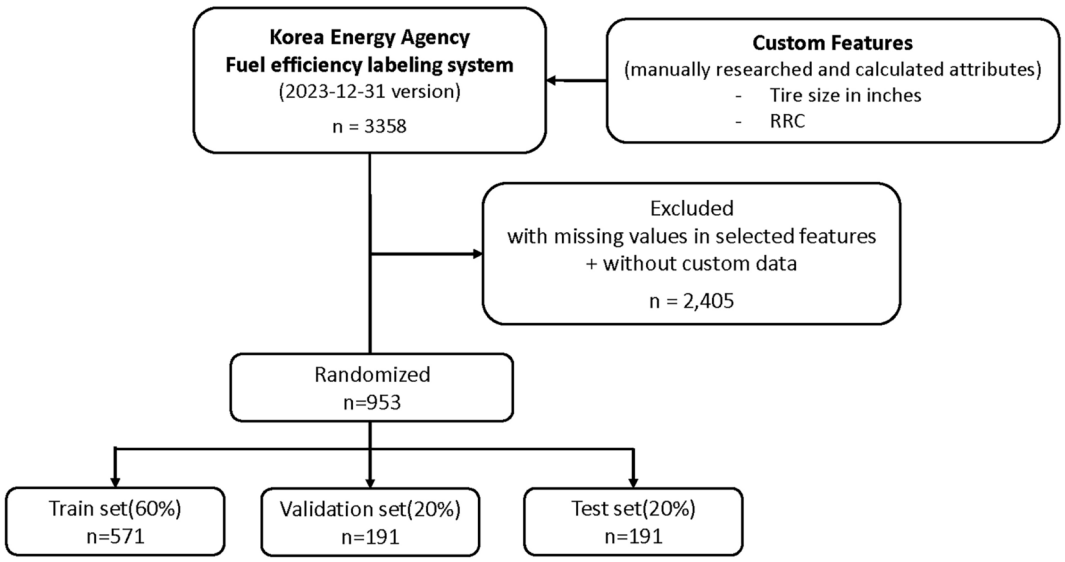

Fig. 2. Flow chart of dataset composition and split.

Variable	Description
manufacturer/importer	Manufacturer or importer
vehicle type	Type of vehicle (e.g., passenger car)
type	Type of vehicle (e.g., standard type)
car type	Type of car (e.g., internal combustion engine)
fuel type	Type of fuel (e.g., LPG, gasoline)
transmission type	Transmission type (e.g. automatic 8-speed)
combined mode CO2	CO2 emissions in combined mode (unit: g/km)
grade	Fuel economy grade
inch	Tire size (unit: inch)
gas disp	Gas (engine) displacement (unit: cc)
RRC	RRC value

Table 1. Description of each variable.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

Baseline characteristics of the discovery and validation cohorts are an important part of describing the basic characteristics of the dataset used in the study. This part helps you clearly understand the entire dataset and the composition of the training, validation, and testing sets. Through this, you can check the number and distribution of samples included in each dataset, and the average and standard deviation of key variables. Descriptions of these basic characteristics help to identify the components of the dataset, which enables analysis that takes into account the characteristics of the data during the model learning and evaluation process. For example, analyzing the distribution of various variables included in the dataset provides useful information to evaluate the representativeness of the data and verify the performance of the model. Therefore, it plays an important role in increasing the reliability of research and verifying the validity of the results.

Correlation coefficient analysis and CCA (Canonical Correlation Analysis) were used for data analysis. Correlation coefficient analysis is used to measure the relationship between two variables and quantify the degree of correlation between them. The correlation coefficient indicates the relationship between variables and can be calculated in a variety of ways. There are three correlation coefficient methods used this time: Pearson, Spearman, and Kendall. The Pearson correlation coefficient measures the linear relationship between continuous variables. It evaluates the linear correlation between two variables and has a value between − 1 and 1. The closer the value is to 1, the stronger the positive linear correlation, and the closer it is to − 1, the stronger the negative linear correlation. The Spearman rank correlation coefficient measures rank-based correlation between variables. Unlike the Pearson correlation coefficient, it can evaluate non-linear relationships between variables and analyze the relationship between the ranks of two variables. Kendall's Tau correlation coefficient is a method of measuring rank correlation between two variables. It evaluates the consistency of ranking changes between variables and provides more stable results, especially in small samples. These correlation coefficient analysis methods can help you deeply understand the relationships between variables and evaluate patterns or relationships in data. CCA (Canonical Correlation Analysis) is a statistical method that finds a linear relationship between two different sets of variables. This method helps to understand and interpret patterns in data. r in the graph is the Pearman correlation coefficient, which is an indicator of the strength and direction of the linear relationship between two variables. The range is from − 1 to 1, and a larger absolute value indicates a stronger linear relationship between the two variables. p is the p-value, which is the probability that the observed data in a statistical test will be obtained under the null hypothesis. In general, the smaller the value, the stronger the evidence to reject the null hypothesis as it provides evidence that the observed data did not occur by chance.

We developed a fuel efficiency prediction model using machine learning, using 'manufacturer/importer', 'vehicle type', 'type', 'car type', 'fuel type', 'transmission type', 'grade', 'combined mode_CO2', 'inch', 'gas displacement', 'RRC' as input variables (X) and the fuel efficiency of the vehicle as the target variable (Y). The data was divided into 60% training data, 20% validation data, and 20% test data. Six regression models from the Scikit-learn library (version 1.5.2) were used in our experiments: Extra Trees Regressor[26], Random Forest Regressor[27], Gradient Boosting Regressor[21], Hist Gradient Boosting Regressor[28], AdaBoost Regressor[29] and Linear Regression[8]. These models were selected based on their unique strengths in regression tasks and their ability to capture complex relationships within the dataset, each offering specific advantages that align with our objective of achieving high predictive accuracy for fuel efficiency.

Extra Trees Regressor[26] was chosen due to its ability to mitigate overfitting and enhance prediction accuracy by using randomized splitting at each node. This model is robust to noise and is well-suited for high-dimensional data, making it ideal for managing our dataset's diverse features. Random Forest Regressor[27] was selected for its stability and resistance to overfitting, achieved through ensemble learning with multiple decision trees. This

model effectively captures non-linear relationships and interactions among features, making it reliable for fuel efficiency prediction. Gradient Boosting Regressor[21] was included for its high predictive accuracy, attained through a sequential process that iteratively builds trees to correct errors made by previous ones. This model performs particularly well with structured data and complex feature interactions. We also selected Hist Gradient Boosting Regressor[28], an extension of Gradient Boosting optimized for large datasets, due to its computational efficiency and ability to handle missing values directly. This model is effective with high-dimensional data, fitting well with datasets containing numerous engineered and categorical features, like ours. Lastly, AdaBoost Regressor[29] was chosen for its simplicity and adaptability, focusing on harder-to-predict instances by adjusting the model's emphasis over successive iterations, which is especially useful for improving performance on imbalanced or noisy data. Finally, we included Linear Regression[8] to serve as a baseline model, providing a point of comparison with the more complex, tree-based ensemble methods. As a simpler model, Linear Regression enables us to evaluate the added benefits of using more sophisticated algorithms for fuel efficiency prediction, especially in capturing non-linear relationships and feature interactions that Linear Regression[8] may overlook. This comparison allows us to assess the practical value of complex models and determine whether their increased computational complexity translates to substantial improvements in predictive accuracy.

By leveraging each model's distinct strengths, we aimed to thoroughly evaluate fuel efficiency prediction across various machine learning approaches, selecting the model that ultimately provided the highest predictive accuracy.

A total of four evaluation indicators were used to measure the performance of the machine learning model: MSE, RMSE, MAE, and R^2 Score. MSE is an indicator that represents the mean square error between the actual observed value and the value predicted by the model. RMSE is the square root of MAE and is an indicator of the average size of prediction error. MAE is the average absolute error between the actual observed value and the value predicted by the model, which is the actual observed value, and yi is the value predicted by the model. The smaller the MSE, RMSE, and MAE, the more accurate the model's predictions. R-squared (Coefficient of Determination) is an indicator of the explanatory power of a regression model and indicates the proportion of the total variation in the dependent variable that the model can explain. The maximum value is 1, and the closer it is to 1, the higher the accuracy of the model. Using all four of these to determine accuracy can provide a richer evaluation of the model's predictive performance.

To ensure the generalizability of the machine learning models, we utilized k-fold cross-validation as an experimental approach. In k-fold cross-validation, the dataset is divided into k equal-sized folds, and the model is trained k times, each time using a different fold as the validation set while the remaining k−1 folds are used for training. This method helps to mitigate potential overfitting and ensures that the model's performance is robust across different subsets of data. By averaging the results from each fold, we can obtain a more reliable measure of the model's predictive accuracy and stability.

Additionally, subgroup analysis was conducted to assess differences in fuel efficiency between different vehicle manufacturers. Specifically, we compared Hyundai and Kia vehicles using statistical tests to determine if there are significant differences in fuel efficiency between the two manufacturers. This analysis aimed to provide insights into the variation in fuel efficiency by brand and to explore the potential influence of manufacturer-specific factors.

We also experimented by removing features with high correlation coefficients to evaluate the impact of multicollinearity on the model's performance. By selectively excluding highly correlated variables, we aimed to reduce redundancy in the feature set and improve the robustness of the model. This approach helps to ensure that each feature contributes unique information, thereby enhancing the interpretability and stability of the model.

Additional analysis was performed on the final model using sharp (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). Sharp is an effective tool used to interpret predictions of machine learning models. Sharp provides information about how much each feature contributed to the model's prediction. This allows you to visually express the importance of each variable and understand the model's prediction process. Additionally, LIME was used to interpret individual predictions by approximating the model's behavior locally, enabling a more detailed, instance-specific explanation of the model's outputs. Together, these methods provided comprehensive insights into both overall feature importance and specific prediction interpretations, enhancing the transparency of the model.

Univariate analysis is a technique used in data analysis to understand the distribution and characteristics of a single variable. It mainly analyzes the distribution, median, mean, and variance of variables to understand the basic properties of the data. Univariate analysis identifies the characteristics of variables through several statistical indicators. Univariate analysis evaluates the characteristics of variables through several statistical indicators. What we used here are Odds Ratio (OR), CI 95%, and p-value. The odds ratio is an indicator that represents the ratio between the probability of occurrence of two events. If the OR is greater than 1, it indicates that the characteristic increases the probability of an event occurring, and if it is less than 1, it indicates that the characteristic decreases the probability of the event occurring. A value of 1 means that there is no difference between the characteristic and the occurrence of the event. CI (Confidence Interval) is a way to express the accuracy of an estimate. It is an interval that represents the probability that a statistic estimated from given data will become the actual value. A confidence interval is an interval that is expected to contain a parameter (characteristic of a population) under a certain confidence level.

A narrower confidence interval indicates a higher accuracy of the estimate, and the range of the confidence interval may vary depending on the volatility of the data and the sample size. Each of these serves as an important tool in understanding and interpreting patterns in data. Odds ratio, CI, and p-value obtained through univariate analysis are useful for drawing conclusions and assigning statistical significance based on data.

Experiment results

Table 2 presents the baseline characteristics of the discovery and validation cohorts used in the training, validation, and test datasets. The total number of data points is 953, divided into 571 for the training data, 191 for the validation data, and 191 for the test data. The dataset includes a variety of vehicle attributes, which are essential for the prediction of fuel efficiency. The "Vehicle type" attribute is divided into "car" and "van" categories. For the training data, 556 vehicles are classified as cars (97.4%), and 14 vehicles as vans (2.6%). In the validation and test datasets, the proportions remain similar, with 183 cars (95.8%) and 7 vans (3.7%) in the validation data, and 183 cars (95.8%) and 7 vans (3.7%) in the test data. The "Manufacturer/importer" attribute includes several brands. For the training data, 198 vehicles are from Kia (34.7%), 345 from Hyundai (60.4%), and a few from other manufacturers. In the validation data, 77 vehicles are from Kia (40.3%) and 109 from Hyundai (57.1%), while in the test data, 86 vehicles are from Kia (45.0%) and 108 from Hyundai (56.5%). The "Fuel type" attribute includes LPG, diesel, and gasoline. Most of the vehicles in all datasets use gasoline, with a few using LPG and diesel. The fuel type distribution is consistent across the datasets, with a slight variation in the number of diesel vehicles. The "Transmission type" attribute includes automatic and manual transmissions. The majority of vehicles across all datasets are equipped with automatic transmissions, with automatic types 4, 5, 6, and 8 being the most common, while a smaller proportion of vehicles have manual transmissions. The "Grade" variable is used to categorize vehicles based on their overall performance and fuel efficiency. The "light type" and "normal type" grades are the most common, with other grades showing fewer vehicles. Additionally, vehicle attributes such as engine displacement, combined CO2 emissions, tire inch size, and rolling resistance coefficient (RRC) were recorded for each vehicle. These variables are crucial as they directly impact fuel efficiency predictions. For example, the combined CO2 emissions for the training data are 161.5 g/km (28.5 standard deviation), which

Variable	Train data	Validation data	Test data
Total, n = 953	**n = 571**	**n = 191**	**n = 191**
Vehicle type, car	556 (97.400)	183 (95.800)	183 (95.800)
Vehicle type, van	14 (2.600)	7 (3.700)	7 (3.700)
manufacturer/importer, Cage Mobility	11 (1.900)	3 (1.600)	1 (0.500)
manufacturer/importer, Kia	198 (34.700)	77 (40.300)	79 (41.400)
manufacturer/importer, PSA Automobiles	3 (0.500)	0 (0)	0 (0)
manufacturer/importer, Renault Korea Automobile Co., Ltd.	12 (2.100)	1 (0.500)	2 (1.000)
manufacturer/importer, Ford	1 (0.200)	0 (0)	0 (0)
manufacturer/importer, Hyundai	345 (60.400)	109 (57.100)	108 (56.500)
type, Multi-purpose type	295 (51.700)	92 (48.200)	86 (45.000)
type, Normal type	216 (37.800)	77 (40.300)	77 (40.300)
type, Passenger/cargo type	0 (0)	0 (0)	1 (0.500)
type, etc type	59 (10.300)	21 (11.000)	26 (13.600)
car type, internal combustion engine	570 (99.800)	190 (99.500)	190 (99.500)
fuel type, LPG	32 (5.600)	11 (5.800)	15 (7.900)
fuel type, diesel	229 (40.100)	70 (36.600)	78 (40.800)
fuel type, gasoline	309 (54.100)	109 (57.100)	97 (50.800)
transmission type, automatic 4	15 (2.600)	4 (2.100)	5 (2.600)
transmission type, automatic 5	4 (0.700)	1 (0.500)	3 (1.600)
transmission type, automatic 6	45 (7.900)	14 (7.300)	20 (10.500)
transmission type, automatic 7	58 (10.200)	21 (11.000)	16 (8.400)
transmission type, automatic 8	422 (73.900)	146 (76.400)	139 (72.800)
transmission type, manual 6	2 (0.400)	0 (0)	0 (0)
transmission type, stepless shifting	24 (4.200)	4 (2.100)	7 (3.700)
grade, light type	19 (3.300)	5 (2.600)	8 (4.200)
grade, level 1	6 (1.100)	5 (2.600)	2 (1.000)
grade, level 2	68 (11.900)	29 (15.200)	22 (11.500)
grade, level 3	115 (27.100)	46 (24.100)	51 (26.700)
grade, level 4	180 (31.500)	52 (27.200)	54 (28.300)
grade, level 5	142 (24.900)	53 (27.700)	53 (27.700)
combined mode CO2	161.500 (28.500)	161.800 (31.400)	162.300 (29.300)
inch	18.500 (1.500)	18.400 (1.400)	18.300 (1.500)
gas disp	2367.500 (712.100)	2422.900 (785.800)	2404 (732.500)
RRC	0.200 (0.100)	0.200 (0.100)	0.200 (0.100)
combined fuel efficiency	11.200 (2.100)	11.200 (2.400)	11.300 (2.300)

Table 2. Descriptive statistics of dataset.

provides insight into the average environmental impact of the vehicles in the dataset. Finally, the "Combined fuel efficiency" variable, which measures the vehicle's fuel efficiency in terms of the fuel consumption rate, is also included in the dataset. The average combined fuel efficiency for the training data is 12.00 (2.00), and the values for the validation and test datasets are similar, with slight variations in the standard deviations.

Table 3 and Fig. 3 together illustrate the correlation analysis between combined fuel efficiency and various predictor variables, using Pearson, Spearman, Kendall correlation coefficients, and Canonical Correlation Analysis (CCA). Both analyses reveal that "combined mode CO2" and "grade" have consistently strong correlations with combined fuel efficiency across all coefficients (e.g., r = 0.88 and r = 0.89 in CCA, respectively, both with p = 0.000). This high degree of correlation suggests that these variables are critical in predicting fuel efficiency, as they have a substantial impact on outcomes. However, their strong correlation also raises concerns about potential multicollinearity, which could affect model stability if not carefully managed. In contrast, variables such as "transmission type" and "vehicle type" show weaker correlations with fuel efficiency, with r-values of 0.19 and 0.09, respectively, indicating that these factors may have limited predictive value and a smaller impact on fuel efficiency variation. The results of this combined analysis guide the selection of features in our predictive model: highly correlated variables like "combined mode CO2" and "grade" are prioritized for their predictive power, while lower-priority variables with weaker correlations can be deprioritized or excluded, thereby reducing model complexity and addressing multicollinearity risks.

We conducted a comparative study on the fuel efficiency of vehicles from Hyundai and Kia using statistical tests as summarized in Table 4. The t-test revealed a significant p-value of 0.00000605, indicating a substantial difference in the average fuel efficiency between the two manufacturers. This result suggests that consumers who prioritize fuel economy might find Hyundai vehicles to be a more advantageous choice compared to Kia, highlighting the importance of fuel efficiency as a key decision factor when selecting a vehicle. On the other hand, the Chi-squared test yielded a p-value of 0.572, indicating no significant difference in the distribution of vehicle types between Hyundai and Kia. This finding implies that both manufacturers have a similar range of vehicle types, such as SUVs and sedans, suggesting comparable competitiveness. Therefore, consumers searching for specific vehicle types may find suitable options available from both manufacturers, which could influence their purchasing decisions and marketing strategies. These insights provide valuable information for understanding consumer behavior and the competitive landscape between manufacturers. Automakers can leverage this data to enhance vehicle performance or refine their marketing messages, thereby appealing more effectively to their target audience. Overall, this analysis serves as a foundation for future research on fuel efficiency and vehicle type preferences, contributing to more informed decision-making by consumers and manufacturers alike.

Tables 5 and 6 present the evaluation results of six machine learning regression models, including both tree-based models (Extra Trees Regressor[26], Random Forest Regressor[27], Gradient Boosting Regressor[21], Hist Gradient Boosting Regressor[28], and AdaBoost Regressor[29]) and a linear model (Linear Regression[8]).

Table 5 shows the performance when using the full set of features and the performance when two highly correlated features, 'class' and 'multimodal CO2', were removed from the set to address multicollinearity. This adjustment was intended to assess the robustness of each model in dealing with potential collinearity issues. The results indicate that Extra Trees Regressor[26] and Random Forest Regressor[27] consistently outperform other models in terms of MSE, RMSE, and MAE values, with R^2 scores closest to 1, both in validation and test sets. These models consistently show high performance even when multicollinearity is reduced, showing that they can reliably handle complex interrelationships in the data. On the other hand, the linear model showed significantly lower performance and was not stable.

Additionally, the K-Fold cross-validation metrics in Table 6 reinforce the reliability of Extra Trees Regressor[26] and Random Forest Regressor[27], showing stable and high accuracy across multiple data folds. This consistency underscores the robustness of these models, making them preferable choices for fuel efficiency prediction in this dataset.

Table 7 presents a comparative analysis of the performance metrics of the Extra Trees Regressor[26] and Random Forest Regressor[27] for Hyundai and Kia vehicles. In the validation data analysis, the Extra Trees Regressor[26] achieves a minimum MSE of 0.007 for Hyundai, indicating excellent predictive performance, with a high R^2 value of 0.998. This suggests that the model effectively predicts fuel efficiency for Hyundai vehicles. In contrast, the Random Forest Regressor[27] has slightly higher MSE values for both manufacturers, indicating lower accuracy. Examining the test data, the Extra Trees Regressor[26] records an MSE of 0.011 for Hyundai and 0.218

Variable name	Pearson correlation coefficient	Spearman correlation coefficient	Kendall correlaion coefficient
manufacturer/importer	0.160	0.150	0.120
vehicle type	0.090	0.080	0.070
type	0.150	0.180	0.140
fuel type	0.220	0.330	0.270
transmission type	0.190	0.230	0.180
combined mode_CO2	0.880	0.890	0.760
grade	0.890	0.950	0.850
inch	0.420	0.430	0.330
gas_disp	0.730	0.730	0.550

Table 3. Correlation analysis of between cohort variables.

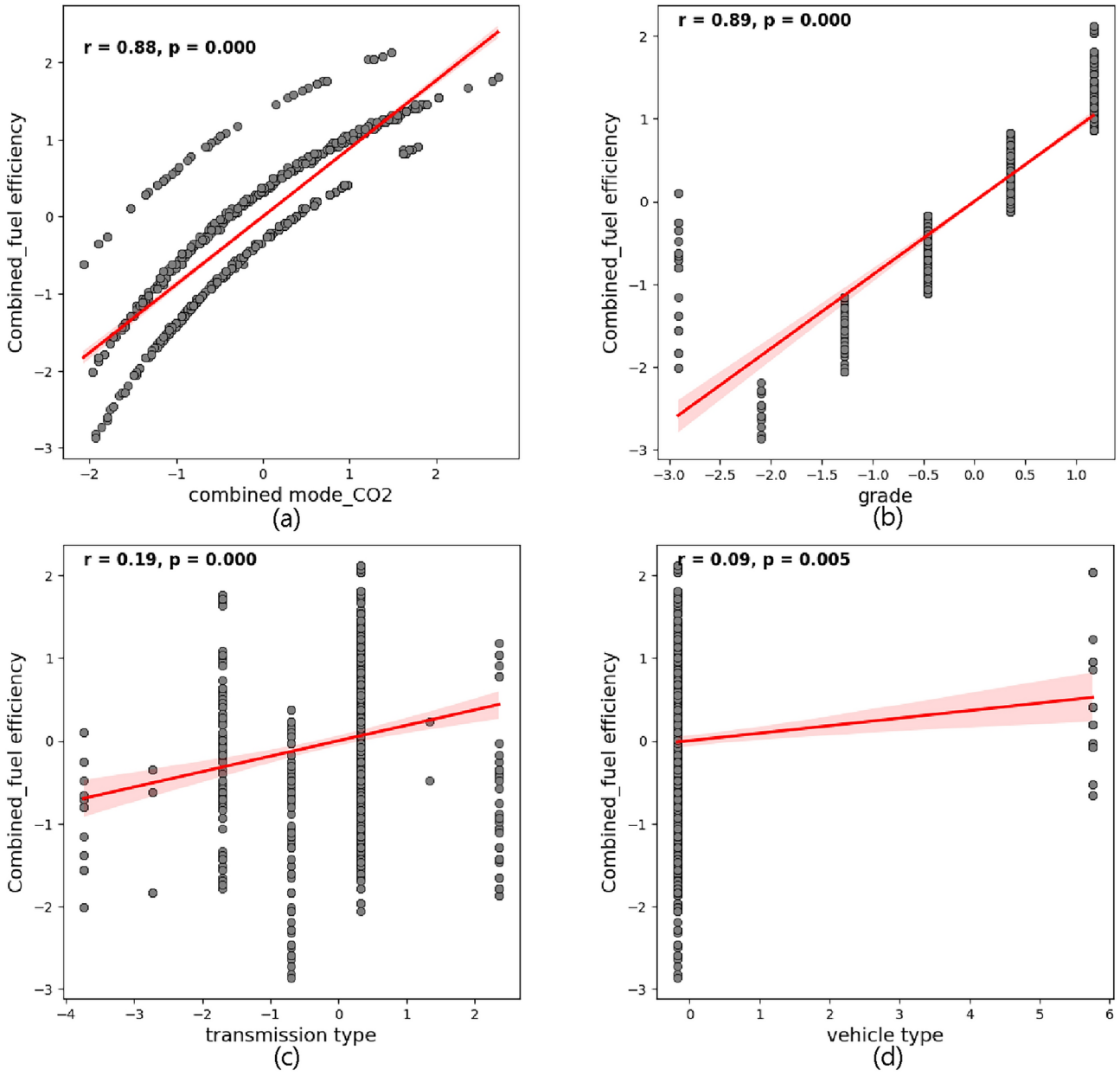

Fig. 3. Analysis of correlation between two variables using CCA. (**a**) combined mode_CO2 and combined fuel efficiency analysis, (**b**) grade and combined fuel efficiency analysis, (**c**) transmission type and combined fuel efficiency analysis, and (**d**) vehicle type and combined fuel efficiency analysis.

Analysis type	Statistic	P-value
T-test	− 4.551	6.048
Chi-squared test	3.840	0.572

Table 4. Statistical analysis results for Hyundai and Kia vehicles.

for Kia. The Random Forest Regressor[27] shows MSE values of 0.017 for Hyundai and 0.315 for Kia. The R^2 values are 0.960 for Hyundai and 0.980 for Kia, emphasizing the models' efficiency. This analysis reveals that the Extra Trees Regressor[26] consistently outperforms the Random Forest Regressor[27] for Hyundai vehicles, suggesting it better captures the feature interactions specific to this manufacturer. However, the higher prediction errors for Kia indicate potential areas for optimization in the predictive model or vehicle design.

Machine learning prediction model	All features				Reduced multicollinearity			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
Validation data								
Linear Regression[8]	0.050	0.224	0.164	0.990	0.674	0.821	0.674	0.856
Extra Trees Regressor[26]	0.006	0.078	0.042	0.998	0.381	0.617	0.406	0.931
Random Forest Regressor[27]	0.013	0.114	0.066	0.997	0.351	0.592	0.399	0.936
Gradient Boosting Regressor[21]	0.017	0.117	0.087	0.997	0.408	0.639	0.483	0.926
Hist Gradient Boosting Regressor[28]	0.060	0.245	0.137	0.989	0.401	0.633	0.444	0.927
AdaBoost Regressor[29]	0.255	0.505	0.433	0.953	0.701	0.837	0.677	0.873
Test data								
Linear Regression[8]	0.043	0.208	0.145	0.991	0.780	0.883	0.719	0.829
Extra Trees Regressor[26]	**0.008**	**0.089**	**0.043**	**0.998**	**0.212**	**0.461**	**0.318**	**0.960**
Random Forest Regressor[27]	**0.013**	**0.115**	**0.063**	**0.997**	**0.167**	**0.409**	**0.308**	**0.968**
Gradient Boosting Regressor[21]	0.016	0.130	0.084	0.996	0.257	0.507	0.408	0.951
Hist Gradient Boosting Regressor[28]	0.077	0.279	0.151	0.985	0.200	0.447	0.328	0.962
AdaBoost Regressor[29]	0.243	0.493	0.410	0.954	0.703	0.838	0.703	0.868

Table 5. Performance comparison analysis between machine learning models for validation and test datasets, including an evaluation of reduced multicollinearity effects.

Model	MAE	RMSE	MSE	R^2
Linear Regression[8]	0.039	0.211	0.044	0.990
Extra Trees Regressor[26]	0.054	0.091	0.011	0.997
Random Forest Regressor[27]	0.074	0.103	0.013	0.997
Gradient Boosting Regressor[21]	0.074	0.104	0.011	0.997
Hist Gradient Boosting Regressor[28]	0.083	0.133	0.018	0.996
AdaBoost Regressor[29]	0.431	0.504	0.255	0.947

Table 6. K-fold cross-validation performance metrics comparing machine learning model performance.

Model	Hyundai				Kia			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
Validation data								
Extra Trees Regressor[26]	0.007	0.085	0.040	0.998	0.050	0.225	0.085	0.990
Random Forest Regressor[27]	0.003	0.055	0.034	0.999	0.061	0.248	0.104	0.987
Test data								
Extra Trees Regressor[26]	0.011	0.108	0.061	0.997	0.218	0.467	0.189	0.960
Random Forest Regressor[27]	0.017	0.131	0.064	0.996	0.099	0.315	0.165	0.980

Table 7. Comparative performance analysis of machine learning models for Hyundai and Kia vehicles, assessing prediction accuracy across validation and test datasets.

In addition to the performance metrics presented in Tables 5 and 6, Figs. 4 and 5 offer a comprehensive visual analysis of model performance by comparing the actual versus predicted values and examining the residual distributions for each machine learning model used for fuel efficiency prediction, using all available features.

Figure 4 illustrates the Actual vs Predicted values, which provide insight into each model's accuracy in replicating real-world fuel efficiency data. In this figure, data points that closely align with the red diagonal line indicate high prediction accuracy. The Extra Trees Regressor[26] and Random Forest Regressor[27] models show data points tightly clustered along the diagonal line, suggesting their effectiveness in capturing complex patterns across the dataset. AdaBoost[29], on the other hand, shows a wider spread, suggesting it may be less accurate due to its sensitivity to noisy data and potential issues with overfitting. Figure 5 shows the residual distributions for each model, which help in understanding the bias and variance in predictions. Ideally, residuals should be symmetrically distributed around zero with minimal variance, indicating unbiased and reliable predictions. Both Extra Trees Regressor[26] and Random Forest Regressor[27] have residuals that are tightly centered around zero, confirming their accuracy and low bias as seen in Fig. 4. The Hist Gradient Boosting[28] and Gradient Boosting[21] models also show centralized residuals, but with slightly wider distributions, indicating a modest increase in variance. AdaBoost[29] and Linear Regression[8] exhibit wider and more dispersed residuals, with Linear

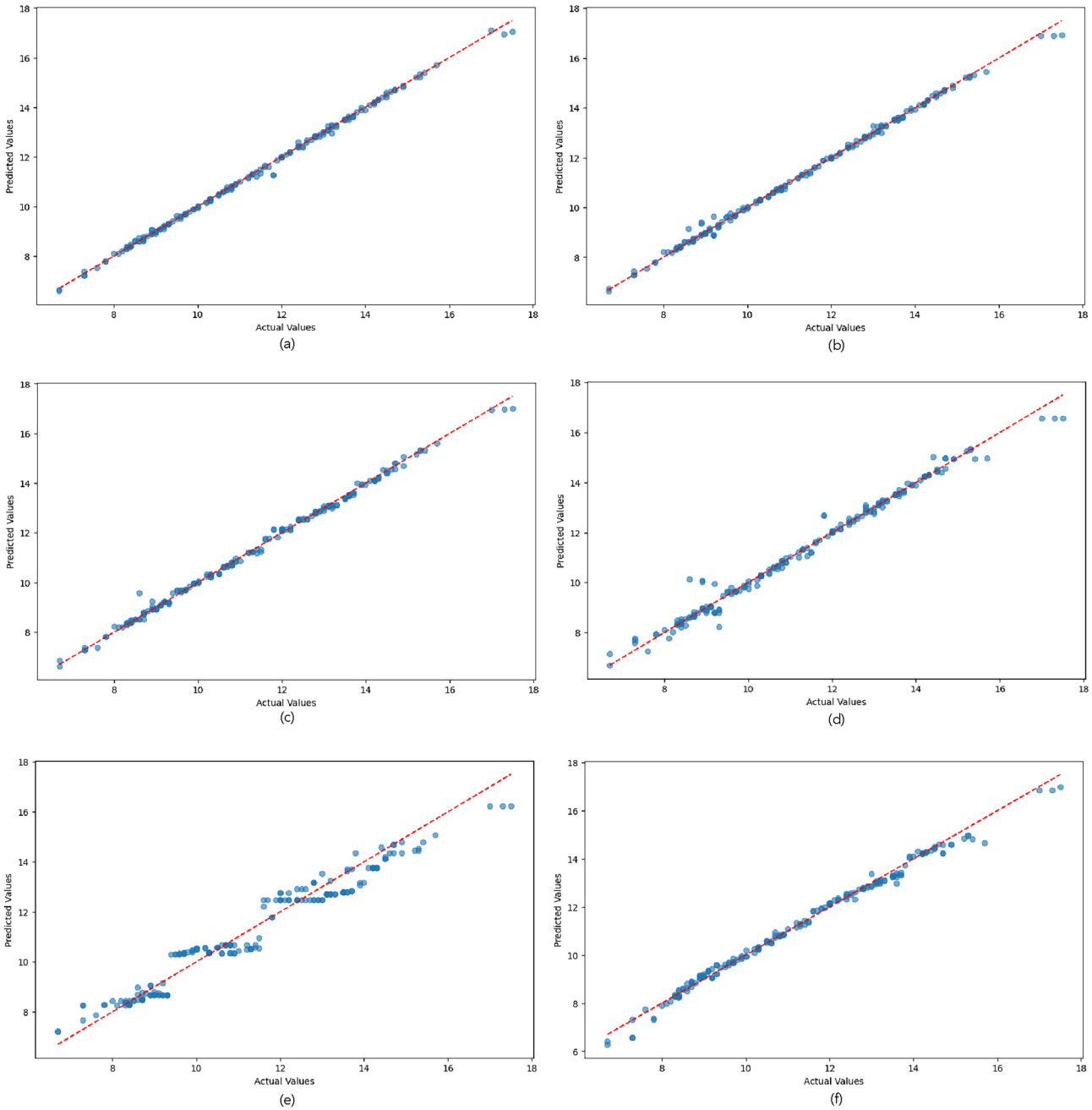

Fig. 4. Actual vs Predicted values for fuel efficiency prediction models. (**a**) Extra Trees Regressor[26], (**b**) Random Forest Regressor[27], (**c**) Gradient Boosting Regressor[21], (**d**) Hist Gradient Boosting Regressor[28], (**e**) AdaBoost Regressor[29] and (**f**) Linear Regression[8].

Regression[8] showing more variability due to its simplicity in capturing complex relationships within the dataset. These visualizations reinforce that tree-based ensemble models, such as Extra Trees[26] and Random Forest[27], are particularly effective for this dataset when all features are included. The close alignment of predictions with actual values and minimal residuals highlights the ability of these models to handle complex relationships. This suggests that ensemble models are well-suited for capturing the nuances of fuel efficiency prediction without the need for feature reduction in this case.

The fuel efficiency prediction system was implemented based on the Extra Trees Regressor model[26], which showed the highest performance. Figure 6 is an example of a fuel efficiency prediction system. As shown in Fig. 6a, by entering vehicle information such as manufacturer/importer, car type, fuel type, and transmission type, the fuel efficiency of the vehicle was predicted. In the example, the fuel efficiency was predicted to be 9.4.

Table 8 presents the results of the univariate analysis, showing the odds ratios and p-values for each variable related to vehicle fuel efficiency. The odds ratio quantifies the strength of the relationship between a given factor and fuel efficiency. Variables with a high odds ratio indicate a stronger impact on fuel efficiency, while those

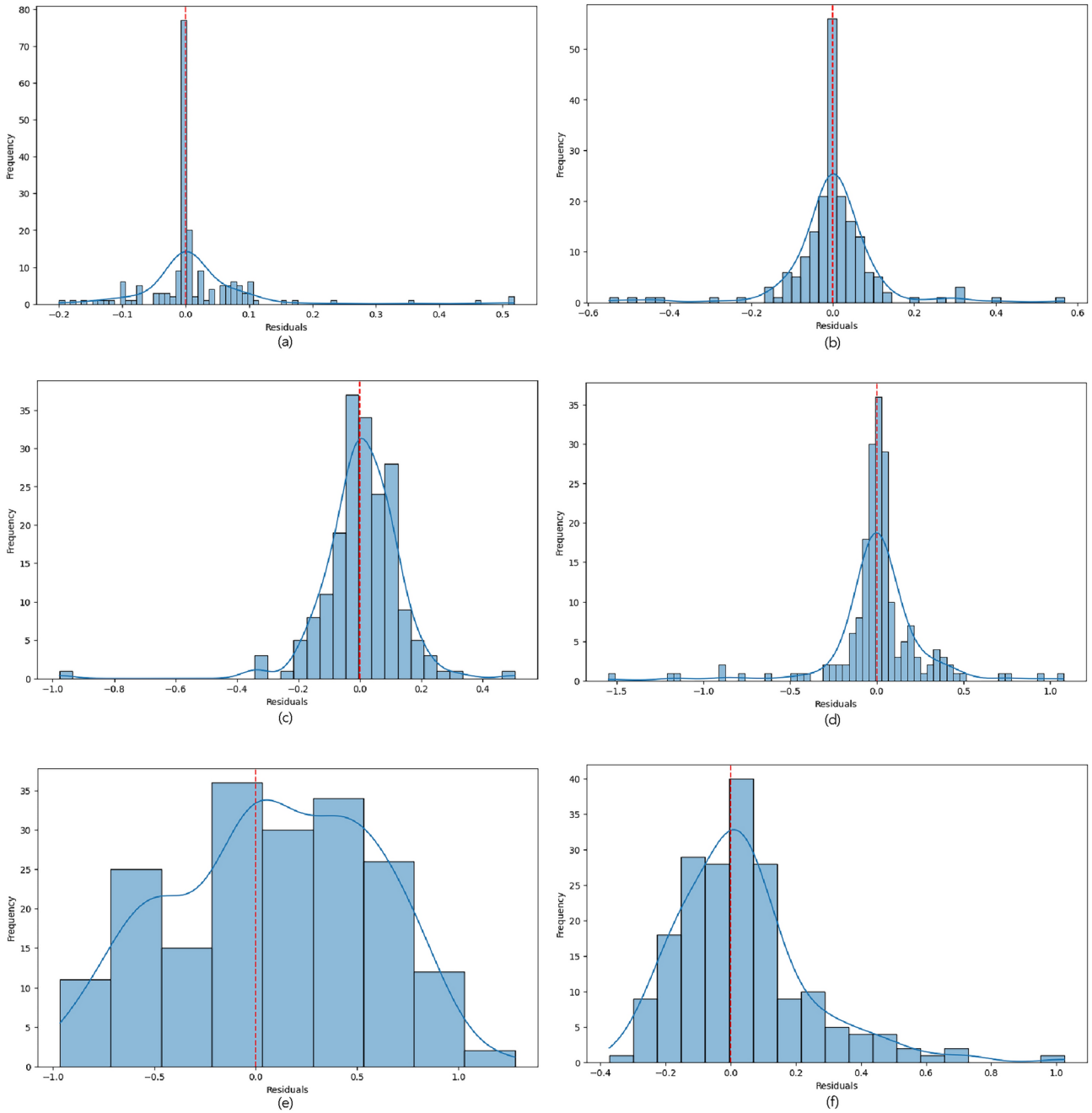

Fig. 5. Residual plots for fuel efficiency prediction models. (**a**) Extra Trees[26], (**b**) Random Forest[27], (**c**) Gradient Boosting[21], (**d**) Hist Gradient Boosting[28], (**e**) AdaBoost[29] and (f) Linear Regression[8].

with a low odds ratio suggest less influence. For instance, fuel type has an odds ratio of 1.333, indicating a positive influence on fuel efficiency, where certain fuel types like diesel or LPG are associated with better fuel economy compared to gasoline. This underscores the relevance of considering fuel type when evaluating vehicle efficiency, as it significantly alters the model's predictive accuracy. On the other hand, vehicle type has an odds ratio of 0.797, implying that specific vehicle types (such as vans) are less efficient than others, possibly due to factors like weight and design, which may increase fuel consumption. Transmission type also plays a significant role, with an odds ratio of 1.325, suggesting that automatic transmissions generally correlate with better fuel efficiency than manual ones. This relationship might be attributed to the optimization of fuel consumption in automatic transmission systems, which are designed to adjust shifting patterns for better fuel use. The significant effect of combined mode CO2 (with an odds ratio of 0.951) confirms that vehicles with lower emissions tend to have higher fuel efficiency. This result highlights the critical importance of reducing CO2 emissions to improve the overall fuel efficiency of vehicles and reduce their environmental impact. The p-values in Table 8 provide further confirmation of these findings, with fuel type, transmission type, combined mode CO2, and grade all showing highly significant relationships with fuel efficiency ($p < 0.05$). This statistical significance implies that

```
Enter manufacturer/importer: hyundai
Enter vehicle type: car
Enter type: Multi-purpose type
Enter car type: internal combustion engine
Enter fuel type: gasoline
Enter transmission type: automatic 8
Enter combined mode CO2 (numeric value): 200
Enter grade: level 4
Enter inch (numeric value): 20
Enter gas disp (numeric value): 3778
Enter RRC (numeric value): 0.155991385
Predicted Fuel Efficiency: 9.405999999999985
```

Fig. 6. Example of fuel efficiency prediction system. (**a**) User input, (**b**) Fuel efficiency prediction value output.

Variable	Odds ratio (95% confidence interval)	p-value
manufacturer/importer	0.968 (− 0.055 to − 0.008))	0.007
vehicle type	0.797 (− 0.486 to 0.034)	0.088
type	0.941 (− 0.111 to − 0.008))	0.023
car type	1.000 (− 0.000 to − 0.000))	<.001
fuel type	1.333 (0.206–0.369)	<.001
transmission type	1.325 (0.228-0.335)	<.001
combined mode_CO2	0.951 (− 0.053 to − 0.045))	<.001
Grade	0.353 (− 1.108 to − 0.9696))	<.001
inch	1.112 (0.037-0.148)	<.001
gas_disp	1.000 (− 0.000 to 0.000)	0.131
RRC	0.000 (− 10.896 to − 8.410))	<.001

Table 8. Univariate analysis using odds ratio.

these variables must be prioritized when developing predictive models for vehicle fuel efficiency, as they directly contribute to improving the accuracy of the predictions.

Based on the model performance comparison in Table 5, Extra Trees Regressor[26] and Random Forest Regressor[27] were selected as the final models due to their superior predictive accuracy and stability across all features and reduced multicollinearity subsets. We employed SHAP analysis to identify and prioritize key markers in vehicle fuel efficiency prediction to enhance interpretability further and validate the importance of specific features.

Figures 7 and 8 present sharp (Shapley Additive Explanations) values, which offer a deeper understanding of how individual features affect fuel efficiency predictions. The sharp values provide an intuitive way to interpret the contributions of each feature in the model's decision-making process. These visualizations demonstrate that certain features consistently exert a high influence on the predicted fuel efficiency across both the Extra Trees Regressor[26] and Random Forest Regressor[27] models. Notably, combined mode CO2 emerges as the most important variable in both models, underscoring its critical role in predicting fuel efficiency. This result reflects the well-established inverse relationship between CO2 emissions and fuel efficiency, where lower CO2 emissions are typically associated with higher fuel economy. The strong influence of combined mode CO2 suggests that optimizing CO2 emissions can significantly enhance vehicle fuel efficiency, making it a key marker for improving overall fuel performance. As such, this variable becomes a crucial factor for policymakers and manufacturers aiming to reduce the environmental footprint of vehicles while optimizing their fuel economy. Further analysis reveals that vehicle grade and fuel type are also critical factors in predicting fuel efficiency. Vehicles with higher grades, particularly grade 4 and grade 5, consistently show better fuel performance. This can be attributed to the superior materials and engineering standards associated with higher-grade vehicles, which likely enhance their efficiency and performance. In contrast, lower-grade vehicles often suffer from poorer fuel efficiency due to less efficient design and construction. Fuel type, particularly diesel and LPG, demonstrates a strong correlation with better fuel efficiency, aligning with real-world knowledge that these fuels generally offer higher energy efficiency compared to gasoline. This reinforces the importance of considering fuel type as a key determinant in fuel efficiency prediction models, as it significantly impacts the overall energy consumption of a vehicle. The sharp dependence plots in Fig. 9 provide a deeper dive into the relationships between these critical features and their effect on fuel efficiency. For example, the plot for grade level clearly shows a positive correlation between higher grade levels (e.g., grade 5) and increased fuel efficiency. This supports the conclusion that higher-grade vehicles are engineered for better fuel performance, likely due to superior components such

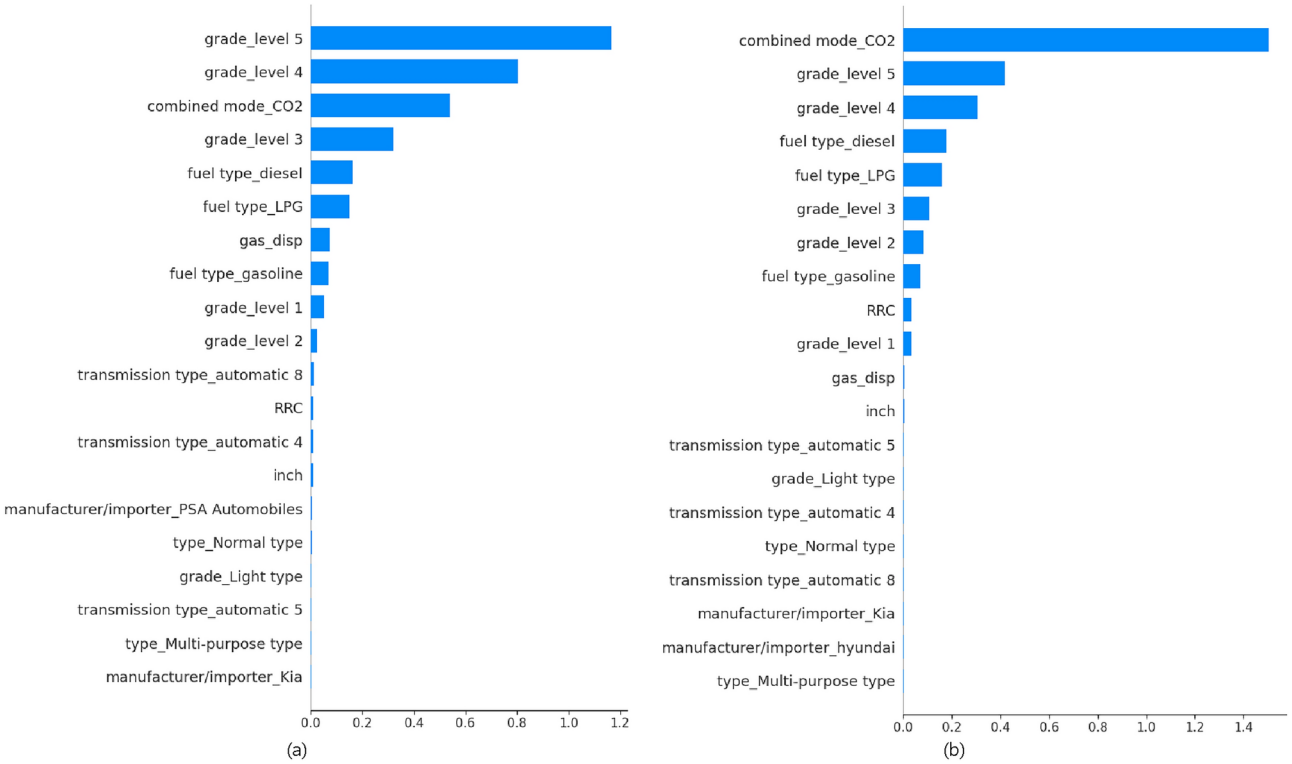

Fig. 7. Visualizing the order of important factors in a machine learning model using sharp summary plots. (**a**) Extra Trees Regressor[26], (**b**) Random Forest Regressor[27].

Fig. 8. Visualizing the order of important factors in a machine learning model using SHAP force plots. (**a**) Extra Trees Regressor[26], (**b**) Random Forest Regressor[27].

as lightweight materials and advanced aerodynamics, which contribute to their fuel efficiency. Similarly, the dependence plot for combined mode CO2 confirms that vehicles emitting higher levels of CO2 tend to exhibit lower fuel efficiency. This reinforces the need for manufacturers to focus on reducing CO2 emissions to optimize fuel economy. The plot for fuel type shows a clear distinction in the relationship between different fuel types (gasoline, diesel, and LPG) and combined mode CO2. Diesel vehicles, in particular, show lower CO2 emissions, suggesting that diesel engines are more efficient in terms of fuel consumption, thereby improving fuel economy. The LIME analysis presented in Fig. 10 further validates these insights by providing detailed, instance-level explanations of the model's predictions. In the visualizations for Hyundai (a) and Kia (b), LIME explains how individual vehicle features, such as fuel type and grade level, contribute to the final predictions of fuel efficiency. For instance, grade level 5 and diesel fuel types are shown to have a significantly higher positive contribution to the predicted fuel efficiency, while lower-grade levels or gasoline fuel types reduce efficiency predictions. These

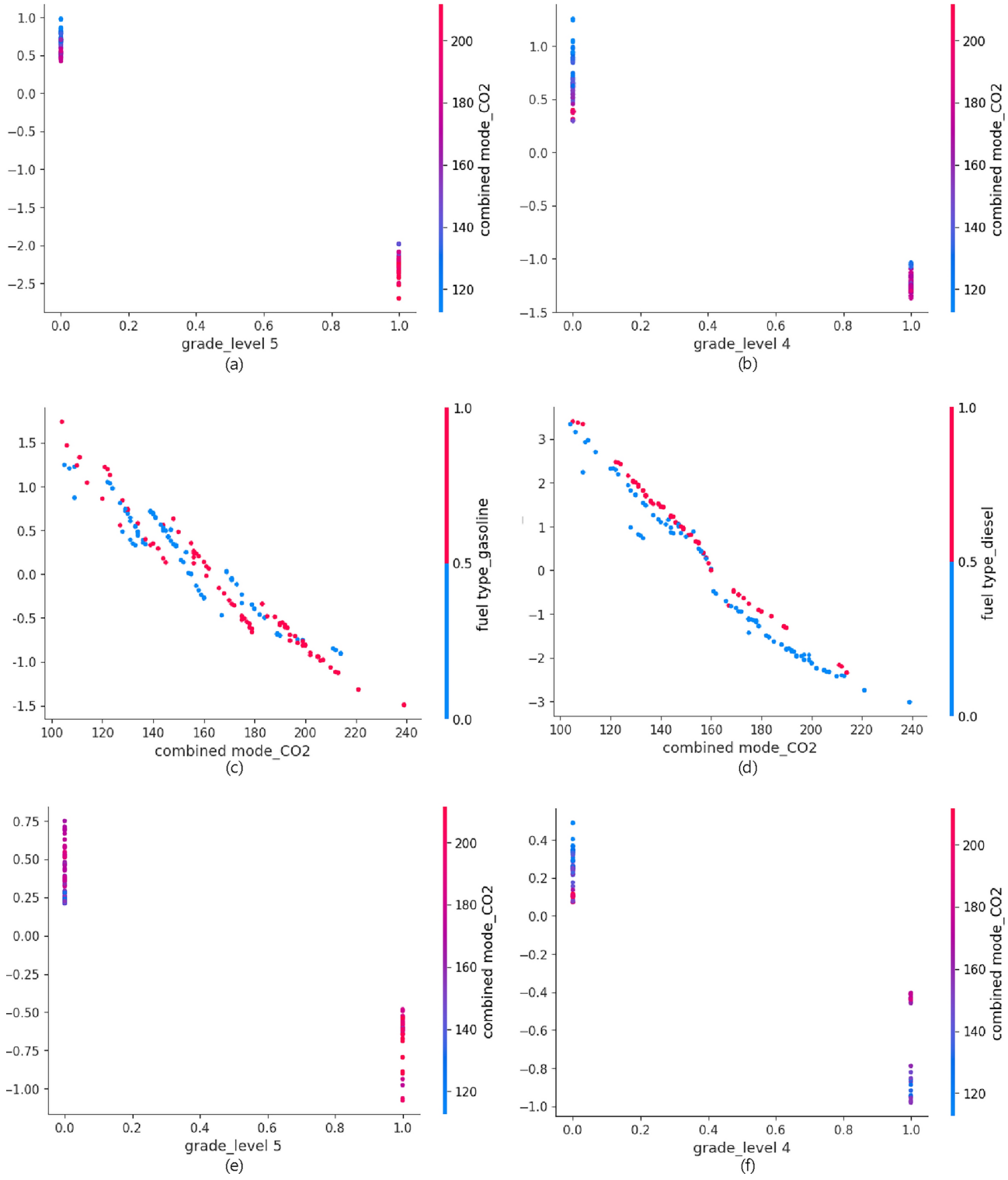

Fig. 9. Sharp dependence plots for Selected features. (**a**) grade_level 5 (Extra Trees Regressor)[26], (**b**) grade_level 4 (Extra Trees Regressor)[26], (**c**) combined_mode_CO2 (Extra Trees Regressor)[26], (**d**) combined_mode_CO2 (Random Forest Regressor)[27], (**e**) grade_level 5 (Random Forest Regressor)[27] and (f) grade_level 4 (Random Forest Regressor)[27].

granular explanations not only confirm the influence of key features but also provide transparency into how the model processes different types of vehicles from specific manufacturers. By understanding how the model reaches its predictions, these findings highlight the importance of incorporating specific vehicle characteristics, such as combined mode CO2, vehicle grade, and fuel type, into predictive models. Optimizing these key factors

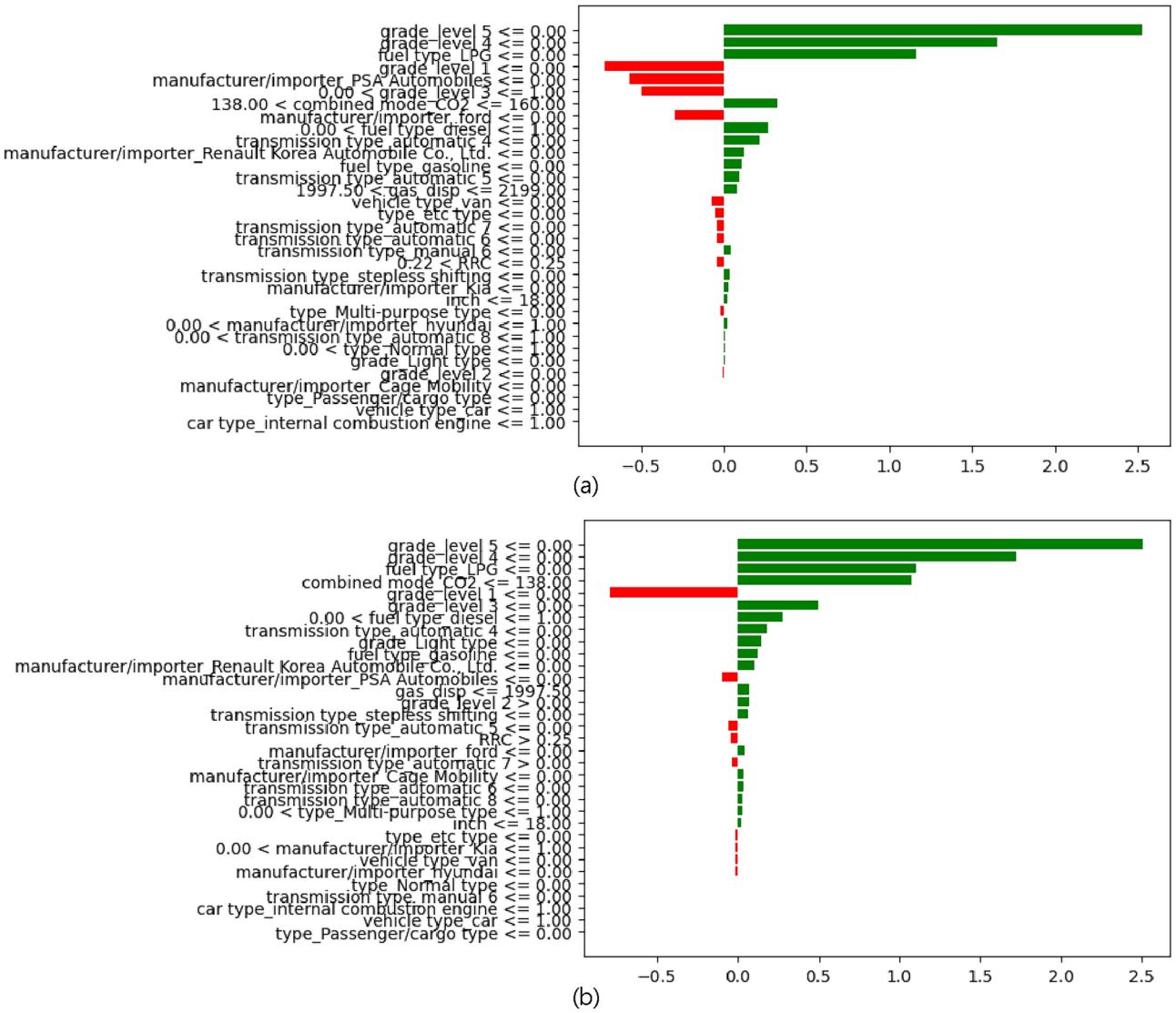

Fig. 10. LIME explanations for selected instances. (**a**) Hyundai, (**b**) Kia.

Model	All features				Top 5 features			
	MAE	RMSE	MSE	R^2	MAE	RMSE	MSE	R^2
Validation data								
Extra Trees Regressor[26]	0.006	0.078	0.042	0.998	0.006	0.077	0.037	0.998
Random Forest Regressor[27]	0.013	0.114	0.066	0.997	0.008	0.093	0.054	0.998
Test data								
Extra Trees Regressor[26]	0.008	0.089	0.043	0.998	0.002	0.052	0.028	0.999
Random Forest Regressor[27]	0.013	0.115	0.063	0.997	0.004	0.063	0.040	0.999

Table 9. Comparison of model performance using All features vs. Top 5 features.

can lead to more accurate predictions of vehicle fuel efficiency, and they should be prioritized in future vehicle designs to promote greater fuel economy.

Table 9 demonstrates the impact of utilizing the top five markers-combined mode CO2, vehicle grade, fuel type, gas displacement, and transmission type-on the performance of fuel efficiency prediction models. When comparing the model performance using all features versus only the top five features, it is evident that these selected markers retain or even enhance the predictive accuracy. The models, Extra Trees Regressor[26] and Random Forest Regressor[27] exhibit marginal improvements in metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score when limited to these core variables.

This enhancement underlines the significance of the proposed markers as essential predictors for fuel efficiency. By focusing on these five key variables, we achieve a streamlined model that not only simplifies data requirements but also upholds or improves predictive performance. This finding suggests that our proposed markers can serve as a reliable foundation for efficient fuel prediction, supporting the potential to reduce model complexity without compromising accuracy.

Discussion

We represent one of the first comprehensive efforts to predict vehicle fuel efficiency by integrating a diverse array of vehicle attributes. Our methodology is structured around three key components: the proposed decision-making pipeline for fuel efficiency prediction, the development of a customized dataset, and the identification of crucial markers for assessing fuel efficiency.

Firstly, we established a decision-making system aimed at predicting vehicle fuel efficiency. This pipeline incorporates not only fundamental specifications such as engine displacement and vehicle type but also nuanced variables like the rolling resistance coefficient (RRC) and combined mode CO2 emissions. By employing machine learning techniques, particularly the Extra Trees Regressor[26] and Random Forest Regressor[27], our analysis successfully captures complex, nonlinear relationships among these variables. Previous studies often concentrated on isolated factors affecting fuel efficiency-such as engine performance or driving habits. However, our research emphasizes the significance of a multifaceted approach that considers various interrelated factors, thus enhancing the accuracy and reliability of fuel efficiency predictions.

Secondly, the dataset we constructed plays a crucial role in this research. We developed a customized dataset using data obtained from a public data portal, supplemented with additional custom data such as tire size and RRC. In the data preprocessing phase, we removed rows with missing values to ensure data integrity. This dataset combines standardized public data with manually collected engineering variables, allowing for a comprehensive analysis of vehicle efficiency. Integrating these features is vital for a deeper understanding of the factors influencing fuel economy.

Finally, our research highlights the importance of specific markers in predicting fuel efficiency. Through univariate analysis and SHAP feature importance insights, we identified critical variables such as vehicle grade, combined mode CO2, fuel type, gas displacement, and transmission type significantly influencing fuel economy. These findings align with a growing body of literature that underscores these factors' critical roles in determining fuel efficiency. For instance, research has shown that combined mode CO2 levels directly impact fuel economy, with lower emissions correlating with improved efficiency[30]. Similarly, vehicle grade is associated with enhanced fuel performance, as higher-grade vehicles often utilize more advanced technologies that optimize fuel usage[30]. The choice of fuel type also significantly influences efficiency, with diesel vehicles consistently demonstrating better fuel economy compared to their gasoline counterparts, reflecting differences in energy content and combustion characteristics[31].

The targeted elimination of multicollinear variables not only improved the model's predictive accuracy but also provided clarity regarding the impact of each variable. This methodological approach is crucial for understanding the intricate relationships among various factors influencing fuel efficiency. The practical implications of our findings are significant for manufacturers aiming to optimize vehicle design and for policymakers seeking data-driven standards that promote energy conservation. Furthermore, we recommend expanding future research to incorporate real-world driving conditions and explore dynamic models capable of real-time predictions. This will ensure the adaptability of our approach across diverse operational settings, ultimately contributing to more efficient and environmentally friendly vehicles.

Conclusions

We proposed a comprehensive framework for predicting vehicle fuel efficiency, centering on three critical components: the proposed decision-making pipeline, the development of a customized vehicle dataset, and the identification of key markers essential for fuel efficiency prediction.

First, we created a customized dataset by aggregating data from the Korea Energy Agency's Fuel Efficiency Labeling System[25] and supplementing it with unique variables, such as tire size and RRC. This tailored dataset not only reflects contemporary vehicle specifications but also addresses limitations in existing studies that often rely on outdated datasets.

Second, we established a decision-making system designed specifically for vehicle fuel efficiency prediction. This pipeline integrates both fundamental specifications, such as engine displacement and vehicle type, and nuanced variables, including the rolling resistance coefficient (RRC) and combined mode CO2 emissions. The incorporation of these diverse factors allows for a more robust prediction model, significantly enhancing the accuracy and reliability of fuel efficiency forecasts.

Lastly, our analysis identified critical markers impacting fuel efficiency through advanced techniques like SHAP, LIME and odds ratio analysis. These markers, which include combined mode CO2, vehicle grade, fuel type, gas displacement, and transmission type, offer valuable insights for manufacturers and policymakers seeking to enhance vehicle design and promote energy conservation.

Looking ahead, future research should focus on validating these findings in real-world settings, particularly through collaborations with vehicle manufacturers and industry stakeholders to implement these models in practical applications. Engaging with manufacturers will facilitate the collection of additional data and feedback that can enhance model accuracy and relevance. This collaborative approach can lead to the development of customized fuel efficiency solutions tailored to specific vehicle models and market segments, thereby maximizing the utility of the proposed framework.

Furthermore, expanding the dataset to include diverse driving conditions, such as urban versus rural settings, highway driving, and varying weather conditions, will provide a more comprehensive understanding of fuel efficiency under real-world scenarios. By capturing a broader range of variables that affect vehicle performance, the predictive models can become more robust and reliable. This diversification will also allow for a nuanced analysis of how different factors interact in various driving contexts.

Additionally, exploring dynamic models capable of real-time predictions will significantly enhance the adaptability of our approach. By integrating telematics and in-vehicle data, future research can investigate how instant feedback on driving behavior and environmental conditions influences fuel efficiency. Implementing machine learning algorithms that can learn from ongoing vehicle operations could yield insights into improving driver habits and vehicle performance in real-time.

Data availability

The data supporting the findings of this study are available from the public data portal. https://www.data.go.kr/data/15083023/fileData.do?recommendDataYn=Y.

Received: 18 September 2024; Accepted: 1 April 2025
Published online: 28 April 2025

References

1. Shim, S., Lee, D., Roh, J. & Park, J. A machine learning based algorithm for short-term weekends load forecasting. *Trans. Korean Inst. Electr. Eng.* **71**, 1578–1584. https://doi.org/10.5370/KIEE.2022.71.11.1578 (2022).
2. Zhang, J., Wang, Z., Liu, P. & Zhang, Z. Energy consumption analysis and prediction of electric vehicles based on real-world driving data. *Appl. Energy* **275**, 115408. https://doi.org/10.1016/j.apenergy.2020.115408 (2020).
3. Wu, T. et al. Impact factors of the real-world fuel consumption rate of light duty vehicles in China. *Energy* **190**, 116388. https://doi.org/10.1016/j.energy.2019.116388 (2020).
4. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794. https://doi.org/10.1145/2939672.2939785 (2016).
5. Noh, M., Kim, J. & Rho, K. A study on economic impact analysis and effect of vehicle fuel economy as rolling resistance on tire energy consumption efficiency rating. In *Proceedings of the Korean Society of Automotive Engineers Fall Conference and Exhibition* 98–104 (Korean Society of Automotive Engineers, 2021).
6. Lee, J., Ko, S. & Lee, B. Research on powertrain system behavior and simulation for prediction of hev fuel consumption. In *Proceedings of the Korean Society of Automotive Engineers Spring Conference* 1296–1302 (Korean Society of Automotive Engineers, 2019).
7. Goodfellow, I. et al. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* 2672–2680 (MIT Press, 2014).
8. Pedregosa, F. et al. Scikit-learn: Machine learning in python - linear regression. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
9. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. https://doi.org/10.1214/aos/1013203451 (2001).
10. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).
11. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. https://doi.org/10.1007/BF00994018 (1995).
12. Kim, J., Song, S. & Baek, J. Fuel cell vehicle fuel efficiency prediction through data correction based on adversarial generation methodology. In *Proceedings of the Spring Conference of the Korean Institute of Industrial Engineers* (Korean Institute of Industrial Engineers, 2021).
13. Rho, K., Jung, H. & Sim, C. A study on causality between vehicle cumulative mileage and fuel economy using statistical analysis method. *Trans. Korean Soc. Autom. Eng.* **29**, 589–595. https://doi.org/10.7467/KSAE.2021.29.6.589 (2021).
14. Kim, B., Kim, J. & Kim, H. Basic research of measurement fuel economy and reducing harmful emissions for passenger car. *J. Korean Soc. Manuf. Technol. Eng.* https://doi.org/10.17958/ksmt.22.6.202012.1143 (2020).
15. Kwon, S. et al. A study on the comparison of emissions and fuel efficiency performance of 2.0 liter lpg hybrid engine and vehicle. *J. ILASS-Korea* **28**, 191–197 (2023).
16. Kim, S., Choi, W., Kim, M., Kim, H. & Lim, W. Analysis of fuel economy of mild hybrid vehicle by the backward simulation with considering power loss of oil pump. *Trans. Korean Soc. Autom. Eng.* **26**, 533–539. https://doi.org/10.7467/KSAE.2018.26.4.533 (2018).
17. Jo, S. et al. Study on the application of rde data for the prediction of heavy-duty vehicle fuel economy. In *Proceedings of the Korean Society of Automotive Engineers Fall Conference and Exhibition* (Korean Society of Automotive Engineers, 2019).
18. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Learning Representations by Back-Propagating Errors* vol. 323 (1986).
19. Katreddi, S. & Thiruvengadam, A. Trip based modeling of fuel consumption in modern heavy-duty vehicles using artificial intelligence. *Energies* https://doi.org/10.3390/en14248592 (2021).
20. Katreddi, S., Kasani, S. & Thiruvengadam, A. A review of applications of artificial intelligence in heavy duty trucks. *Energies* https://doi.org/10.3390/en15207457 (2022).
21. Pedregosa, F. et al. Scikit-learn: Machine learning in python - gradient boosting regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (2011). Accessed: 2024-10-27.
22. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
23. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
24. Thejovathi, M. & Rao, M. C. S. Evaluating the performance of xgboost and gradient boost models with feature extraction in fmcg demand forecasting: A feature-enriched comparative study. *J. Theor. Appl. Inf. Technol.* **102** (2024).
25. Agency, K. E. Fuel efficiency labeling system data (2019). Accessed: 2024-11-04.
26. Pedregosa, F. et al. Scikit-learn: Machine learning in python - extra trees regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html (2011). Accessed: 2024-10-27.
27. Pedregosa, F. et al. Scikit-learn: Machine learning in python - random forest regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (2011). Accessed: 2024-10-27.
28. Pedregosa, F. et al. Scikit-learn: Machine learning in python - hist gradient boosting regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html (2011). Accessed: 2024-10-27.
29. Pedregosa, F. et al. Scikit-learn: Machine learning in python - adaboost regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html (2011). Accessed: 2024-10-27.
30. Yin, X., Li, Z., Shah, S. L., Zhang, L. & Wang, C. Fuel efficiency modeling and prediction for automotive vehicles: A data-driven approach. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* 2527–2532. https://doi.org/10.1109/SMC.2015.442 (IEEE, 2015).

31. Zhao, D. et al. A review of the data-driven prediction method of vehicle fuel consumption. *Energies* https://doi.org/10.3390/en16145258 (2023).

Acknowledgements
This research was financially supported by Hansung University.

Author contributions
S.Y. and S.C. designed the study and performed the experiments. S.Y. analyzed the data. S.Y., S.C. and J.S. wrote the manuscript. All authors contributed to the manuscript.

Declarations

Competing interests
All authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.-H.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.