SciEn

Proceedings of 8th International Conference on Mechanical Engineering and Renewable Energy 2025
10-12 December, 2025, Chattogram, BANGLADESH

SciEn Conference Series: Engineering Vol. 4, 2026, pp 618-623 https://doi.org/10.38032/scse.2026.4.283

A Comparative Machine Learning Approach for Energy Consumption Prediction in Hybrid Vehicles

Bishawajit Chakraborty[1], Arpita Paul[2], Bhubon Thiotonius Costa[3], MD. Shajratul Alam Towhid[3]*

[1]Department of Mechanical Engineering, Rajshahi University of Engineering and Technology, Rajshahi-6204, Bangladesh
[2]Department of Civil Engineering, Sylhet Engineering College, Sylhet, Bangladesh
[3]Department of Naval Architecture and Marine Engineering, Bangladesh University of Engineering and Technology Dhaka, Bangladesh

ABSTRACT

With the rise in environmental concerns and global shift toward sustainable mobility, accurately predicting energy consumption in hybrid vehicles (HVs) is essential for reliable transportation planning and optimizing efficient energy management. This study employs the National Household Travel Survey (NHTS) dataset to develop predictive models for estimating energy usage in HVs. In this paper, four machine learning models (XGBoost, Linear Regression, Random Forest, and Neural Network) are implemented to analyze and compare their effectiveness in forecasting energy consumption. The dataset is preprocessed and standardized, followed by model training and 5-fold cross-validation. The performance of each model is determined using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Among them, the Random Forest model achieved the best effectiveness with lower RMSE and MAE along with the highest R^2 score for predicting energy consumption in hybrid vehicles. The findings show the prospective of machine learning techniques, particularly the Random Forest model, which delivers the highest accuracy in predicting hybrid vehicle energy consumption. This result supports more informed decision-making for developing eco-friendly transport solutions, reducing fuel consumption and emissions.

Keywords: Hybrid Vehicle, Fuel Consumption, NHTS, Machine Learning

1. Introduction

The transportation sector consumes a significant share of global energy usage and is a major contributor to air pollution. Globally, the transport sector is responsible for consuming over 30% of total energy consumption and 60% of oil demand, where road vehicles account for about 80% of the total transport energy usage [1]. Governments worldwide are implementing comprehensive strategies to reduce fuel consumption and air pollution in the transportation sector, emphasizing a transition from fossil fuels to achieve a sustainable, low-carbon economy. Growing global focus on sustainable transport has accelerated research into hybrid vehicle energy optimization, blending combustion engines with electric propulsion systems to cut fuel use and emissions [2]. Accurate prediction of hybrid vehicle energy use is vital for improving efficiency, supporting intelligent transportation planning, and enhancing performance. Traditional physics-based models fail to capture real-world driving complexities and vehicle conditions, whereas machine learning (ML) methods model and predict complex energy consumption patterns [3]. Several studies show that ML methods such as regression models, decision trees, and deep learning architectures perform well in forecasting fuel economy, emissions, and energy management systems [4], [5]. The growing need to reduce greenhouse gas emissions and enhance transportation efficiency has increased attention on vehicle energy prediction, and recent ML advances have greatly improved prediction accuracy.

A recent study applied a deep learning approach using auto-encoders to forecast energy consumption in real-world scenarios [6]. Zhang et al. [3] analyzed electric vehicle energy demand with real driving data, showing that ML algorithms can model nonlinear patterns without explicit physical models. Hybrid modeling approaches have also gained traction. In contrast to previous studies that focused on EVs [7], this study extends the analysis to HVs with a large scale NHTS dataset. This result shows that the Random Forest model surpasses the advanced LightGBM and XGBoost model achieving higher accuracy and stronger generalization in predicting vehicle energy consumption. A customized energy control strategy for small and medium-sized enterprises (SMEs) combined technical and managerial strategies through a PDCA framework, achieving a 35% energy reduction in a paper mill case study [8]. Ekici et al. [9] developed a backpropagation artificial neural network to forecast energy demand based on orientation, insulation thickness, and transparency ratio, achieving a prediction deviation of 3.43% and accuracy between 94.8% and 98.5%. Mądziel et al. [10] created an electric vehicle energy demand model incorporating dynamic vehicle and environmental data, validating it with R^2 of 0.84 and MAE between 0.75 and 1.23 under diverse driving conditions. A Linear Regression model for short-term load forecasting using Alexa AS data reached 3.339% accuracy over one year by emphasizing feature engineering [11]. Collectively, these studies highlight the importance of advanced ML models, ensemble learning, and feature engineering in improving energy prediction for hybrid and electric vehicles, forming the foundation for the present study's model comparison.

This study employs data extracted from the National Household Travel Survey (NHTS), a comprehensive and widely used dataset that includes detailed trip-level and vehicle-level characteristics for households across the United States. This survey captures a wide range of variables, including vehicle type, fuel economy, trip distance, and social-demographic factors, which makes it ideal for developing a predictive model for hybrid vehicle energy consumption.

This research presents a comparative evaluation of several machine learning models, Linear Regression, Random Forest, XGBoost, and Neural Networks, for modeling energy consumption patterns in hybrid vehicles utilizing a real-world household travel survey dataset in the United States. By evaluating and comparing model performance using statistical metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2), this paper aims to develop the most reliable and accurate machine learning model for predicting energy consumption. This work not only contributes to the development of data-driven transportation models but also supports the broader goal of enabling energy-aware and environmentally conscious vehicular technologies.

The subsequent sections of the paper are organized as follows: Section 2 defines the suggested methodology, incorporating the data cleaning and variable selection, ML model implementation, and estimation model evaluations. After that, Section 3 explains the results and discussion, while Section 4 provides the concluding remarks.

2. Methodology

The 2022 National Household Travel Survey (NHTS) provides the most recent dataset in a long-standing series of travel surveys conducted in the United States every 5 to 8 years since 1969. It captures detailed information on individual and household travel behavior, vehicle characteristics, and trip purposes, offering a valuable foundation for transportation research and policy analysis.

The main goal of this study is to develop and implement a machine learning model for energy consumption forecasting in hybrid vehicles. The proposed methodology, illustrated in Fig. 1, comprises three primary phases: Data preprocessing and variable selection, Machine learning model implementation, and estimation model evaluation.

2.1. Data Preprocessing and Variable Selection

To enhance ML model prediction performance, proper data preprocessing is essential. The NHTS combines vehicle-specific and household-level data on national travel behavior across all trip purposes. In this study, data pre-processing was conducted by data cleaning to improve the performance of predictive ML models. Only binary responses (0 = No and 1=Yes) were included to ensure consistency for analysis. Records containing irrelevant and missing data were excluded. After that, vehicle level and household-level dataset were merged to create a clean and uniform dataset for machine learning model analysis.

In regression analysis, it is crucial to check for multicollinearity among the independent variables, as strong correlations between them can distort the results of the model. High multicollinearity can result in unstable regression coefficient estimates, making it challenging to evaluate how each predictor contributes to the dependent variable separately. Therefore, assessing multicollinearity is an important step to guarantee the precision and reliability of the model's prediction. The presence of multicollinearity among the independent variables is evaluated based on the variance inflation factor (VIF). VIF indicates how much the variation in an projected regression coefficient raises due to multicollinearity among independent variables [12] . It is defined mathematically as (1):

$$VIF_i = \frac{1}{1 - R_i^2} \qquad (1)$$

Here, $R_i^2$is the coefficient of determination which is calculated by running a regression of i-th independent variable on all other variables withing the model.

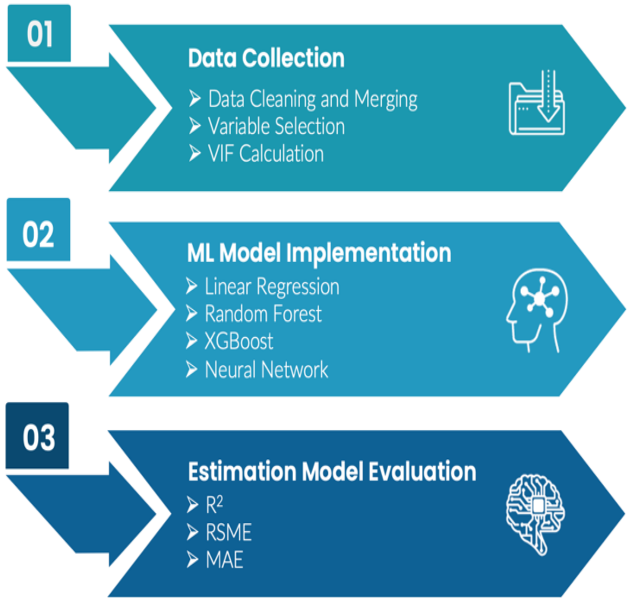


Fig.1 Workflow Diagram for the HVs Energy Consumption Prediction

Table 1 Variation Inflation Factor (VIF) Results

No.	Variable	VIF
1	VEHFUEL	1.018038
2	VEHTYPE	1.009025
3	MAKE	1.022355
4	VEHCOMMERCIAL	1.148309
5	COMMERCIALFREQ	1.167549
6	URBAN	1.766619
7	URBANSIZE	1.694581
8	ANNMILES	1.016312
9	HHVEHCNT	1.576777
10	DRVRCNT	4.449907
11	HHSIZE	2.142429
12	NUMADLT	4.086478
13	WRKCOUNT	1.415691

Through the calculation of the VIF in Table 1, we can identify the problematic variables which are greater than 5 and take corrective measures to ensure the reliable regression estimation coefficients. Additionally, Table 2 displays the HVs energy consumption variables with descriptions to ensure clarity and analytical consistency of the dataset.

Table 2 Variable Definitions

Variable	Type	Description
VEHFUEL	Categorical	Type of fuel vehicle runs on
VEHTYPE	Categorical	Vehicle type
MAKE	Categorical	Vehicle make ID
VEHCOMMERCIAL	Categorical	Vehicle used for business purposes
COMMERCIAL FREQ	Numerical	Over past 30 days, how many days was vehicle used for business purposes?
URBAN	Categorical	Household urban area category
URBANSIZE	Categorical	Urban area size of home address
ANNMILES	Numerical	Self-reported annualized mile estimate
HHVEHCNT	Numerical	Total number of vehicles in household
DRVRCNT	Numerical	Number of drivers in the household
HHSIZE	Numerical	Total number of people in household
NUMADLT	Numerical	Count of adult household members aged 18 and over
WRKCOUNT	Numerical	Count of workers in household

2.2 Machine Learning Model Implementation

This study employs a comparative machine learning framework to forecast energy consumption in hybrid vehicles. Four ML models are performed based on their effectiveness in modeling linearity, non-linearity, and complex feature interactions: Linear Regression, Random Forest, XGBoost, and Artificial Neural Network.

2.2.1 Linear Regression

A linear regression model is introduced as a reference model because of its straightforward implementation and ease of interpretation. It posits a linear relationship between the input variables and the output variable, which allows for determining whether more complex models are necessary [13]. The linear regression model is represented in equation (2).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \varepsilon \qquad (2)$$

Here, Y denotes the target variable; X_1, X_2, X_n, are predictor variables; β_n is the regression coefficient, ε is the random error.

2.2.2 Random Forest

Random forest was chosen for it strength to capture complex, non-linear relationships without relying on any specific assumptions about the data distribution [14]. It operates by aggregating the outputs of numerous decision trees, each developed on random samples of the data. It is denoted by the given formula (3).

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \qquad (3)$$

In this equation, $\hat{y}$ refers to the overall prediction obtained by the averaging results from all trees, N denotes the total count of trees in the forest model, and $\hat{y}_i$ is the predicted value from the i-th tree.

2.2.3 XGBoost

XGBoost is a gradient boosting decision-tree technique that is quick and precise, is employed to capture complex patterns in data [15]. It builds model in a sequential manner, improving performance by correcting the errors of weaker learners. The equation is presented in the following formula (4).

$$\mathcal{L}(\Phi)\sum_{i=1}^{n} l(\hat{y}_i , y_i) + \sum_{m=1}^{M} \Omega(f_m) \qquad (4)$$

here, $\hat{y}_i$ represents the predicted output for sample i, and y_i is the corresponding actual value for sample i. The loss function $l(\hat{y}_i, y_i)$ quantifies the difference between predicted and actual outcomes. M is the total number of trees, f_m is the output of m-th tree, and $\Omega(f_m)$ is the model complexity constraint to prevent overfittings

2.2.4 Artificial Neural Network

Artificial neural network was originally introduced by McCulloch and Pitts [16], is also referred to as a multilayer perception (MLP) [17]. ANNs are used to capture highly complex and non-linear patterns in the data that traditional algorithms may miss. These models simulate human brain function through multiple layers of interconnected neurons. This equation is shown as (5):

$$a^l = \sigma(w^l a^{l-1} + b^l) \qquad (5)$$

here, a^l: output of layer l, w^l: weights matrix of layer l, b^l: bias vector for layer l, σ: activation function, a^{l-1}: output from the previous layer.

2.3 Estimation Model Evaluation

To evaluate the appropriateness of the predictive model, performance metrics were employed. Following the validation of the model's key assumptions, it is crucial to evaluate both the accuracy and the predictive strength of the regression approach. To conduct a numerical comparison of the model's effectiveness, three statistical measures were calculated: R^2, MAE, and RMSE. The corresponding mathematical formulations are presented below in equations (6), (7), (8) respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{i,m} - Y_{i,e})^2}{\sum_{i=1}^{n}(Y_{i,m} - \bar{Y}_{i,m})^2} \qquad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_{i,m} - Y_{i,e}\right| \qquad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{i,m} - Y_{i,e})^2} \qquad (8)$$

Where $Y_{i,m}$ represents actual energy usage for HVs, $Y_{i,e}$ denotes the predicted energy consumption of HVs, $\bar{Y}_{i,m}$ is the average of the measured values, and n stands for the total observations number. A lesser value of MAE and RMSE indicates higher model competence, while a higher R^2 value (closer to 1) implies an closer match of the regression line to the data, signifying improved model performance [18]

3. Results and Discussion

This section demonstrates the comparative evaluation of the machine learning models applied to forecast the hybrid vehicle energy demand using the NHTS 2022 dataset in the United States. Four different regression machine learning models are employed: Linear Regression, Random Forest, XGBoost, and Neural Network. The performance of the models is determined by means of three quantitative metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), along with 5-fold cross-validation for reliability.

3.1. Model Performance Comparison

Based on the evaluation of regression performance using the RMSE, MAE, and R^2, the Random Forest regressor is identified as the optimal model. It shows the lowest RMSE and MAE, indicating superior predictive accuracy along the highest R^2 score reflecting the statistically significant improvements over alternative approaches. Table 3 provides a systematic comparison of these metrics across the evaluated algorithms, confirming the statistically significant advantage of the Random Forest approach.

Table 3 Comparison of Performance Evaluation Metrics

Estimation Models	Performance Indicators of Regression Models		
	R^2	*MAE*	*RSME*
Random Forest	99.93	0.0308	0.1323
XGBoost	99.91	0.0472	0.1509
Linear Regression	99.85	0.0662	0.1910
Neural Network	-849.37	7.8841	15.1151

3.2. Cross Validation Reliability

To determine the reliability of the models, 5-fold cross-validation is conducted for each model (Fig. 2). The Random Forest Model shows the highest median R^2 and narrowest interquartile range, indicating both strong and consistent performance. XGBoost also performs well with slightly lower median and moderate variability. Linear Regression and Neural Network show lower medians and wider spread, especially the Neural Network, suggesting instability and potential sensitivity to training data.

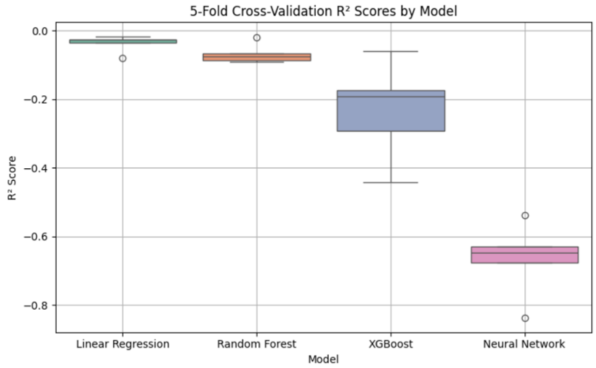


Fig. 2 5-fold Cross-validation of ML Models

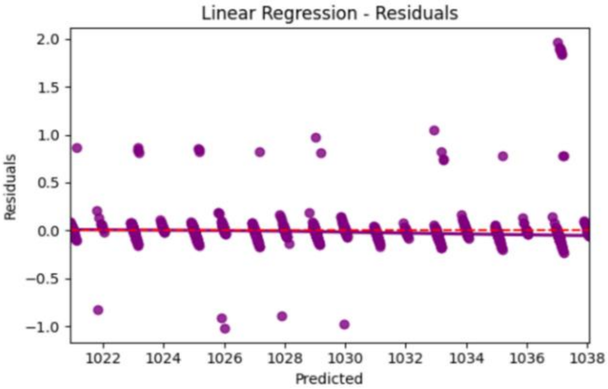
Although all models show low absolute R^2 values, Random Forest performs most consistently and relatively better, demonstrating its ability to capture non-linear relationships in energy consumption without high variance across data splits. The residual patterns indicate that Random Forest provides the most stable predictions with the most balanced residual distribution among all tested models.

3.3. Residual Analysis

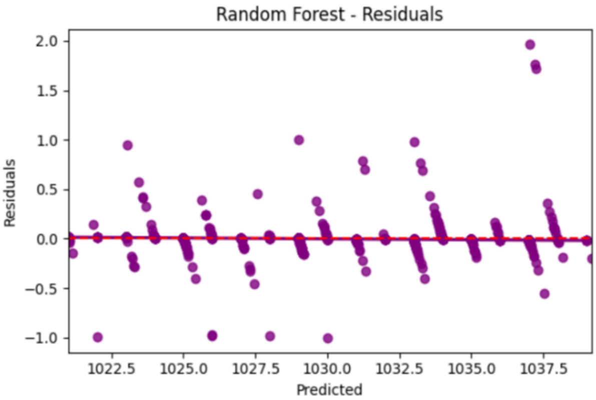
Residual plots evaluate the accuracy and consistency of model predictions across the observed value range. In the Linear Regression model, residuals are uniform around zero, as shown in Fig. 3-a, indicating a reasonably good fit. However, a subtle heteroscedastic pattern emerges, with errors increasing at higher predicted values, suggesting the model struggles with nonlinear variations. In the Random Forest model, residuals cluster closely around zero without any pattern (Fig. 3-b), indicating strong predictive stability and effective capture of underlying data relationships. The XGBoost model shows a nonlinear residual structure with wide variance (Fig. 3-c), suggesting challenges such as overfitting or sensitivity to complex interactions. The Neural Network model displays a pronounced negative residual trend (Fig. 3-d), indicating systematic bias, consistently underpredicting higher values and overpredicting lower ones, resulting in poor generalization. Overall, Random Forest exhibits the most stable residual behavior with minimal variance, confirming its suitability for predicting heavy vehicles' energy consumption in this study.

3.4. Prediction Accuracy

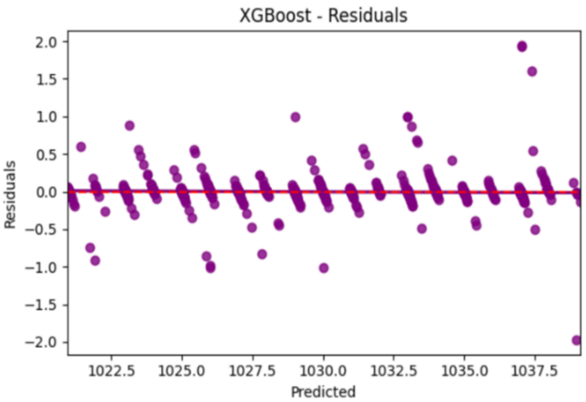
To assess prediction accuracy, actual versus predicted energy consumption values are plotted for each model. In Fig. 4-a, Linear Regression predictions moderately follow the actual values, but some dispersion from the diagonal indicates limited accuracy. Fig. 4-b shows Random Forest predictions tightly clustered along the diagonal, reflecting strong accuracy and minimal deviation. XGBoost predictions in Fig. 4-c exhibit more scatter and deviate from the ideal line, indicating inconsistency. The Neural Network in Fig. 4-d shows substantial scatter and deviation, suggesting weak accuracy and potential overfitting. Overall, the scatter plots confirm that most data points align closely with the diagonal in the Random Forest model, demonstrating its superior performance in estimating energy consumption.



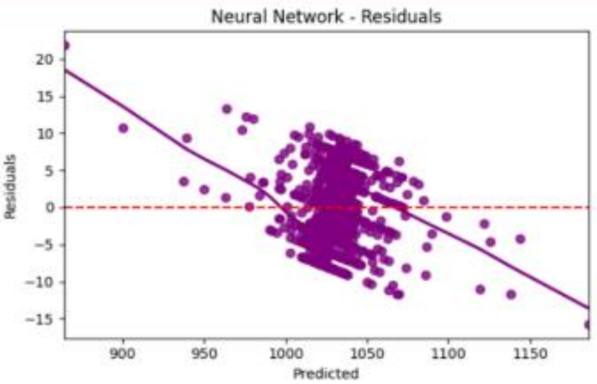
(a)



(b)

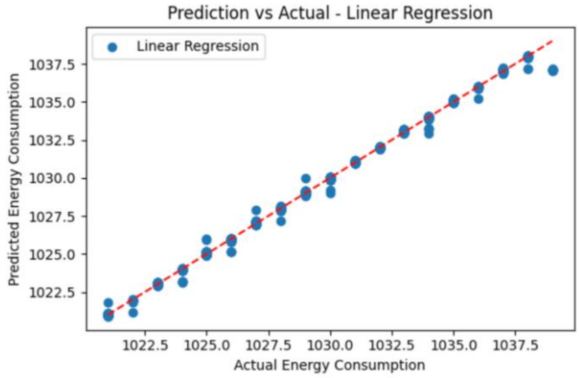


(c)

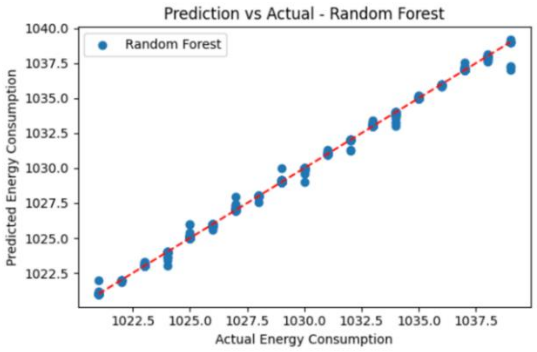


(d)

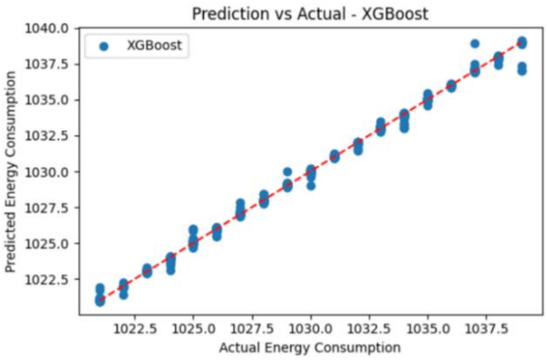
Fig. 3 Residual Analysis of ML Models.



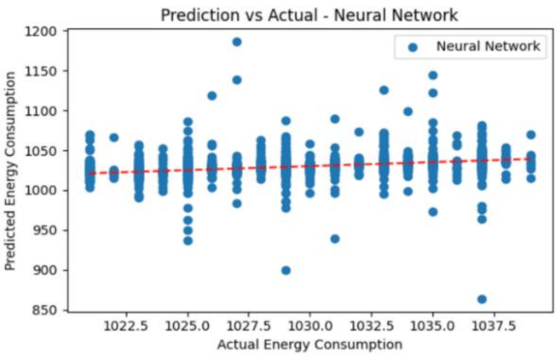
(a)



(b)



(c)



(d)

Fig. 4 Prediction vs Actual Energy Consumption

3.5. Error Visualization

heatmap in Fig. 5 shows RMSE and MAE values, which further clarify the error margins among all models, highlighting the effectiveness of the Random Forest model. This visual comparison supports the choice of tree-based methods for high-dimensional, non-linear NHTS survey data in 2022 in the United States.

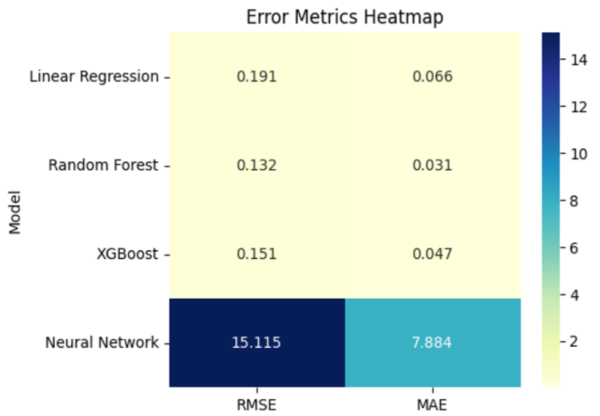


Fig. 5 Error Evaluation Based on RSME and MAE

4. Conclusion

This study evaluates the usefulness of machine learning models in foreseeing hybrid vehicle energy consumption using the US National Household Travel Survey dataset (NHTS). Among the evaluated models, the Random Forest model exhibits the best predictive performance based on RMSE, MAE, and R^2 metrics. The application of 5-fold cross-validation and residual analysis further validates the model's reliability and accuracy. These findings suggest that machine learning, particularly advanced individual models such as Random Forest and XGBoost can play a pivotal role in modelling complex transportation behavior and promoting data-driven decisions in sustainable vehicle usage. Future works could explore feature importance rankings from the Random Forest model to identify key behavioral or household predictors of energy consumption. Additionally, integrating more granular vehicle-level or regional data may enhance model specificity.

4. Conclusion

[1] H.-J. Althaus, “Modern individual mobility,” Int J Life Cycle Assess, vol. 17, no. 3, pp. 267–269, 2012.

[2] A. Emadi, Y. J. Lee, and K. Rajashekara, “Power electronics and motor drives in electric, hybrid electric, and plug-in hybrid electric vehicles,” IEEE Transactions on industrial electronics, vol. 55, no. 6, pp. 2237–2245, 2008.

[3] J. Zhang, Z. Wang, P. Liu, and Z. Zhang, “Energy consumption analysis and prediction of electric vehicles based on real-world driving data,” Appl Energy, vol. 275, p. 115408, 2020.

[4] M. A. Nobi, “A Review: Machine Learning with Electric Vehicles Applications,” Available at SSRN 5085327, 2024.

[5] Y. Yao et al., “Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones,” J Adv Transp, vol. 2020, no. 1, p. 9263605, 2020.

[6] J.-Y. Kim and S.-B. Cho, “Electric energy consumption prediction by deep learning with state explainable autoencoder,” Energies (Basel), vol. 12, no. 4, p. 739, 2019.

[7] I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal, “A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability,” Energy & Environment, vol. 33, no. 8, pp. 1583–1612, Dec. 2022, doi: 10.1177/0958305X211044998.

[8] A. Prashar, “Adopting PDCA (Plan-Do-Check-Act) cycle for energy optimization in energy-intensive SMEs,” J Clean Prod, vol. 145, pp. 277–293, 2017.

[9] B. B. Ekici and U. T. Aksoy, “Prediction of building energy consumption by using artificial neural networks,” Advances in Engineering Software, vol. 40, no. 5, pp. 356–362, 2009.

[10] M. Mądziel, “Impact of Weather Conditions on Energy Consumption Modeling for Electric Vehicles,” Energies (Basel), vol. 18, no. 8, p. 1994, 2025.

[11] M. Spichakova, J. Belikov, K. Nõu, and E. Petlenkov, “Feature engineering for short-term forecast of energy consumption,” in 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), IEEE, 2019, pp. 1–5.

[12] N. Shrestha, “Detecting multicollinearity in regression analysis,” Am J Appl Math Stat, vol. 8, no. 2, pp. 39–42, 2020.

[13] A. F. Schmidt and C. Finan, “Linear regression and the normality assumption,” J Clin Epidemiol, vol. 98, pp. 146–151, 2018.

[14] L. Breiman, “Random forests,” Mach Learn, vol. 45, no. 1, pp. 5–32, 2001.

[15] A. H. Emtiaj, R. H. Rafi, M. A. Nayeem, and M. S. Rahman, “CRISMER: A transformer-based Interpretable Deep Learning Approach for Genome-wide CRISPR Cas-9 Off-Target Prediction and Optimization,” bioRxiv, p. 2025, 2025.

[16] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” Bull Math Biophys, vol. 5, no. 4, pp. 115–133, 1943.

[17] J. Ziółkowski, M. Oszczypała, J. Małachowski, and J. Szkutnik-Rogoż, “Use of artificial neural networks to predict fuel consumption on the basis of technical parameters of vehicles,” Energies (Basel), vol. 14, no. 9, p. 2639, 2021.

[18] J. Fan et al., “Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China,” Energy Convers Manag, vol. 164, pp. 102–111, 2018.