National College of Ireland

Predicting Energy Consumption in Electric Vehicles: A Machine Learning Approach for Enhanced Efficiency and Sustainability

MSc Research ProjectData Analytics

Krutika Rajesh Gite
Student ID:
23164441

School of Computing
National College of Ireland

Supervisor: Prof. Athanasios Staikopoulos

National College of

Ireland MSc Project

Submission Sheet

School of Computing

Student Name:	Krutika Rajesh Gite
Student ID:	23164441
Programme:	Master's in data Analytics
Year:	2024
Module:	Msc In Research Project
Supervisor:	Prof. Athanasios Staikopoulos
Submission Due Date:	12/12/2024
Project Title:	Predicting Energy Consumption in Electric Vehicles: A Machine Learning Approach for Enhanced Efficiency and Sustainability
Word Count:	
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other thanmy own contribution will be fully referenced and listed in the relevant bibliography sectionat the rear of the project.

ALL internet material must be referenced in the bibliography section. Students arerequired to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Krutika…Gite……
……

Date: ……12/12/2024……
……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placedinto the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	

Penalty Applied (if applicable):	

Predicting Energy Consumption in Electric Vehicles: A Machine Learning Approach for Enhanced Efficiency and Sustainability

Krutika Gite
Msc. Data Analytics
National college of Ireland
Student ID- 23164441

Abstract

With the increasing rate of EV ownership worldwide, there is renewed interest in gaining a better understanding of the battery's performance – with emphasis on energy. The research study relies on data from a four-year electric vehicle usage database which includes detailed operational data. Once preliminary data cleaning and the handling of missing values were done, we employed exploratory and predictive analysis on the energy patterns using different forms of regression models. Feature important in energy consumption was estimated by applying feature engineering and correlation analysis. Several multiple regression models were built and tested for their predictive performance with the MAE, RMSE, and the R^2 being used as performance measures. Accordingly, the outcomes identified the unique indicators and approaches to forecast energy consumption. The conclusion is that this research adds new data-driven approaches to the overall improvement of energy efficiency in the EVs, including battery management and prognostics. The next steps for future work will be the inclusion of other data for the analysis of environmental and operation factors.

1 Introduction

Sustainable transport therefore has become one of the most important focus points for climate change mitigation, greenhouse gas emissions reduction, and more broadly energy security in the world. Among all the potentially replaceable fossil fuel-based vehicles, electric vehicles or the EVs are ideal because they emit considerably low operational pollution and also they meet the intergovernmental climate change goals (Chen et al., 2021). However, the large-scale deployment of EVs depends on the solution of the main problems that are associated with the use of batteries, their consumption and reliability. Energy consumption of EV, as a factor affecting vehicle range, cost, and impact on the environment is, hence, another research area that needs to be understood and optimised.

The energy consumption in an EV varies with the following parameters; the driving behavior, surrounding environment, load and power used by the ancillary systems. Liu, Xu et al, in their findings state that authors have identified factors that cause variation in battery

efficiency depending on the situation on the road including steep inclines and heat that affects the vehicle (2017). The management of these variables is crucial not only for advancing battery science but also for reducing the worries that users continue to have, like range anxiety, which holds the key to global EV diffusion (Lin et al., 2020). Hence, knowledge of the predictors of energy consumption remains very important to improve on the performance of the EV and make it sustainable as the world's leading means of transport.

Specifically, this paper aims at conducting the energy consumption analysis of EV batteries based on the detailed dataset provided by the U.S. Department of Energy's EV Data Collection project (LeCroy & Dobbelaere, 2024). The primary research questions are as follows: Where are batteries consuming energy in electric vehicles and how could model predict help improve knowledge of these factors? The purposes of this study are as follows: to determine the important factors that affect energy consumption to compare the effectiveness of different regression models in predicting energy use; and to offer recommendations for improving Battery Efficiency of EVs. Answering these questions, this work is expected to advance the general goal of improving EV performance and boosting the rate of moving towards their use.

One strength of this research is the data collection of both facility-level data like the energy performance of the solar arrays and storage systems as well as specific vehicle levels of operational data. In contrast with many of the articles, which emphasize the vehicle attributes, this approach provides a comprehensive view of the factors affecting the energy consumption of EVs (LeCroy & Dobbelaere, 2024). It also permits a precise assessment of how infrastructure and vehicle dynamics occur, making the findings useful for policy makers, automotive producers as well as energy supervisors in the EV business.

The potential of this research for the existing body of knowledge is reflected in the methodology used and the database inclusion. First, the study uses more sophisticated regression techniques of estimating models of energy consumption to address issues of predictor identification and measurement. Second, by using a large and long panel data, this study provides not only population-level findings but also findings relevant to a broad range of EV models and locations (Ullah et al., 2021). Furthermore, the study has real-world applications to EV design, the development of superior battery management systems, and the determination of infrastructure needed to accommodate EV usage.

The structure of this paper is as follows. Subsequently, the current literature concerning the energy consumed from EVs is reviewed and synthesised whereby, recent progress in modeling approaches and the existing research deficiencies are discussed. The data preparation, regression modeling, and the evaluation criteria that have being used in the present work are outlined in the methodology section. The final section of the paper, the results and discussion section, brings out the results focusing on the predictors of energy consumption and the implication for EV optimisation. Last, the conclusion generalizes the results of the conducted investigation, reports about existing limitations, and points out further research directions.

Therefore, this research makes an important contribution towards filling the gap in the understanding of energy demand management in EVs to enhance proliferation of sustainable transport and the low-carbon economy. The conclusions are to permanently affect the advancement of higher effective EV systems and the accommodation of renewable power into

the transport networks.

2 Related Work

The Importance of Electric Vehicles in Sustainable Mobility

The quest for the sustainable development of the global economy has seen electric vehicles (EVs) as central to the transport change. Since transportation contributes to approximately 23% of the total GHG emissions, thus substituting fossil-fueled cars with EVs could be perceived as a way of lowering the environmental footprint (Chen et al., 2021). Subsidies, tax credits and emission standards have been major drivers of technology advancement through policies (S. Koengkan et al., 2022). However, adoption on a large scale depends on the factors relating to battery technology, by enhancing the capability of the battery as well as the operational reliability.

Factors Influencing Energy Consumption in EVs

The energy consumption pattern of EVs is not an easy one and depends on many factors. Road gradient, driving speed and temperature greatly influence the efficiency of energy (Liu et al., 2017). For example, the utilization of the car engine as a motor for the wheels, which is usual in uphill drives, leads to the high consumption of batteries, weather conditions, especially low temperatures, are also known to be influential since they cause higher usage of the car heating or cooling system (Wang, 2017). Moreover, vehicle load and usage intensity other than propulsion, including air conditioning, exaggerate variability in energy usage (Lin et al., 2020). Knowledge of these factors is crucial in enhancing batteries power delivery and creating improved vehicle architectures for electric cars.

Predictive Modeling for Energy Consumption

Due to the nature of EV charging and the availability of data, predictive modeling has Thus, become pivotal in analyzing and anticipating EV energy consumption. Linear regression is popular employed because of its simplicity coupled with easy understandability. A verity of linear and multilevel regression methods provide good results regarding the tendencies between energy consumption and impact factors, which gives prospects for optimization (Ullah et al., 2021). Nevertheless, such procedures are limited in applicability when faced with non-linear response or several interactions among predictors.

Advancements in Machine Learning

The introduction of machine learning ML has embraced energy modeling for EVs in the most practical way possible. Neural networks, the ensemble method, and various clustering regression models have improved the predictive performance (Yang et al., 2023). In another study Chen et al. (2022) used density based clustering regression models to model the energy consumption of urban EVs and identified their applicability to nonlinear patterns and various kinds of driving situations. Likewise, integration of weak models has been applied in ensemble

methods like the gradient boosting in order to enhance the prediction models (Liang et al., 2023).

Integration of Real-Time Data

Real-time data integration has enhanced the energy consumption modeling for EVs in a new dimension. The data from the sensors in the EVs results into modifications on the estimated energy usage depending on the circumstances of driving and the conditions of environs (Wongsapai et al., 2023). The present strategy, apart from providing desirable density and accuracy, would also help create better battery management systems that adapt drastically for electric vehicles.

Facility-Level Data and Its Implications

LeCroy & Dobbelaere (2024) have incorporated a systematic database containing the facility's energy measurements including those of the solar arrays and storage systems with the vehicles. This integration offering an comprehensive observation of and between infrastructure and vehicle energy dynamics. In reconciling facility operation and electric vehicle performance, this approach has major implications for the enhancement of energy efficiency of the two entities.

Multilevel and Ensemble Models

Variability in energy consumption has also been well captured by analizing the systems under different driving conditions, through the use of Multilevel mixed-effects models. In their study, Besselink and Nijmeijer (2018) also proved that these models are capable of identifying variations of energy consumption based on individual trip, which involves higher granularity in energy models. On the other hand, Ensemble learning techniques, like Stacked generalization, just exploit the best features of different models to enhance the overall accuracy of final predictability (Ullah et al., 2021).

The Role of Environmental Variables

It has been established that temperature and humidity greatly affect the energy usage of electric vehicles. Wang et al. (2017) established that low temperatures enhance battery resistance thus decreasing the battery energy efficiency. Furthermore, the heating or cooling energies consumed by the battery can consume a considerable amount in the battery in extreme climates (Lin et al., 2020). Therefore, using these variables in the prediction entails is crucial to enhancing the predictive models.

Urban vs. Highway Driving Patterns

Traffic behavior can be considered the primary factor that defines the energy intensity of vehicles. City driving which involves constant starting and stopping uses more energy than highway driving that is mostly a steady speed (Koengkan et al., 2022). Chen et al. (2022) also

pointed out that these variations shall be studied by using real-world driving data for regulating energy dissipation that may exist depending on the environment.

Addressing Gaps in Current Research

However, there are still some limitations in existing knowledge about EV energy consumption. While many publications address the issue, few generalise results across car brands or geographical areas. Further, facility-level data and its link with the vehicle data collection analysis have not been widely addressed in the context of (LeCroy & Dobbelaere, 2024). More work has to be conducted to include climate characteristics which include wind speed and precipitation as input variables, in energy consumption models to have more realistic analyses (Wang et al., 2017).

Contributions of This Study

It extends prior work by using a rich, longitudinally structured dataset, which combines information at the facility-level and the vehicle-level. This work applies superior models of regression analysis and considers a broad range of predictors to support better comprehension of energy consumption by EV. As such, the knowledge generated through the research has relevance to the deployment of battery systems and infrastructure, as well as the establishment of policies to support battery usage in electric transportation systems.

3 Research Methodology

This methodological section offers a broad and clear description of the methods used in clarification of the research, presented in such a way that all aspects of the study are all systematically described without omission leading to an account having methodological depth and accuracy that enables replication.

1. Data Collection

1.1 Key Variables

The dataset includes the following critical variables essential for analyzing energy consumption in electric vehicles (EVs)

- **Total Energy Consumption**: This acts as the main dependent variable, amount of energy consumed per trip, and what the model predicts is to be reduced.
- **State of Charge (SOC)**: Both pre and post SOC values are obtained which gives an understanding about battery usage, discharge, and charge capabilities.
- **Distance Metrics**: Other measurements such as total kilometers on the car odometer facilitate determining a correlation between total usage of cars and energy usage patterns.

- **Driving Pattern**: These are such headings as driving time, idling time, average speed, etc., which characterize operational activity related to energy consumption.
- **Environmental Variables**: Application factors such as temperature and weather conditions as they affect energy consumption giving other features of consumption patterns.

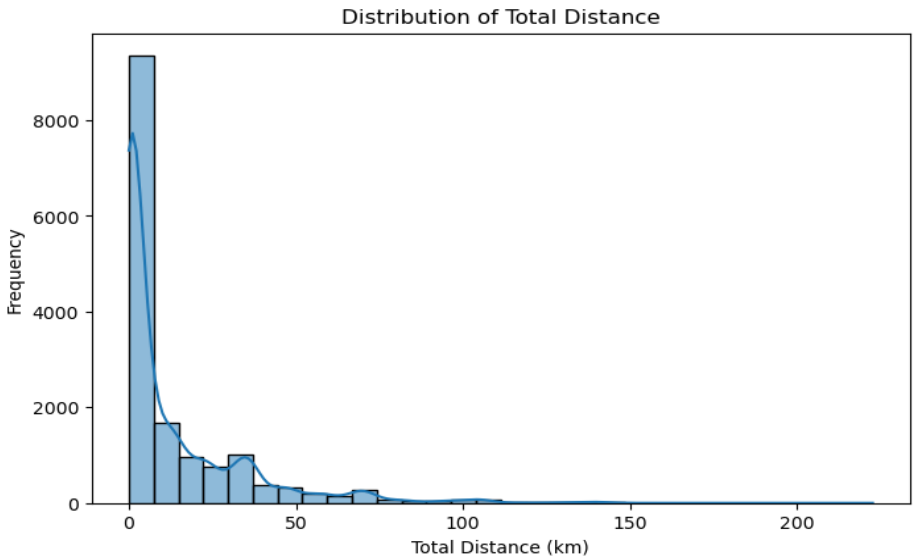

Fig. 1 Distribution of Total Distance

This histogram in fig. 1 represents the frequencies of the total distance traveled, which is also in kilometer, of electric vehicles (EVs). Stated clearly, the majority of trips accomplished are far much less than the 50 kilometer range as evidenced by the graph above. The effects of distance are highly significant with high frequency of short distance transport as evident from the peak at very low distance. Trips exceeding 100 km are especially rare, in proportion to the distance, while the frequency drops sharply as soon as the distance exceeds 50 km. This distribution corresponds to typical use of EVs, which benefit from short distances probably due to range constraints of the technology or work-related trips. The long tail indicates that the distances are occasionally greater than the short- to-mid-range distances indicated by the other entries.

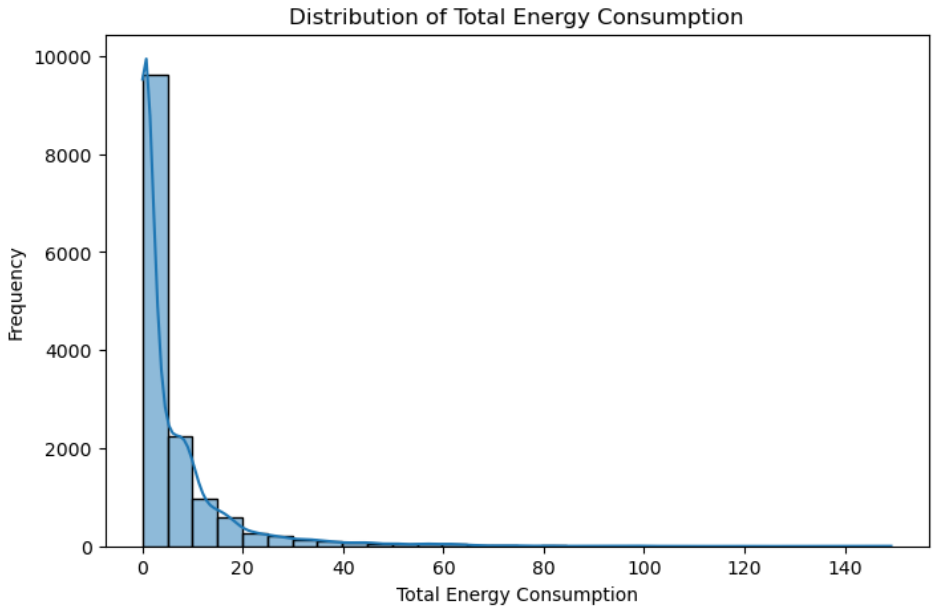

Fig. 2 Total Energy Consumption

The histogram in Fig. 2 demonstrates the total energy consumption throughout EV trips in units that are not specified. To be precise, the data heavily favors low energy consumption range, particularly within the 0 and 25 units range. The sharp rise towards the y-axis below 10 probably corresponds to short or low power travelling. Rare negative values can be explained by the fact of getting wrong data or the process of energy recovery in regenerative brakes. Frequency starts reducing as energy consumption moves above 25 units, and there are very rare incidences where energy consumption moves past 50 units. This pattern is consistent with normal low-energy usage patterns characteristic of EVs where they are mainly used for short-range mobility.

1.2 Data Source and Validation

The dataset was collected from the US Department of Energy's EV Data Collection Project which is a documented database of functional EV data. The validation of this dataset was conducted through several steps to ensure its reliability and robustness:

- **Cross-referencing:** Since key metrics were compared reflecting manufacturer requirements and industrial reports.
- **Integrity Checks:** To this end scripts were written to look for such things as negative SOC values, out of range distances, or unreasonably low energy levels.
- **Consistency Validation:** This process confirmed several hypothesis regarding relations between variables, for example relation between SOC and energy consumption.

- **Metadata Review:** The measurement units provided in the metadata helped establish preprocessing solutions and resolve inconsistencies in the data, while data collection frequencies and known issues enriched collected data.

2. Data Processing

2.1 Handling Missing Values

Missing data in key variables, such as SOC and energy consumption, was addressed using the following techniques:

- **Median Imputation:** Imputing missing values by median still retained the statistical trends of the results while at the same time dealing with the effects of outliers.
- **Temporal Interpolation:** For temporal data, dummy variables, forward & backward filling was applied in order to maintain the temporal logical continuity of various records.

2.2 Feature Scaling and Selection

- **Scaling:** Measures that assume continuous values, including SOC percentages and total distance in kilometres, were scaled to the same range to avoid quantitation and dominant range effects in models.
- **Selection:** Chi square test of independence, odds ratio, sensitivity, specificity and accuracy were used in model validation.

  Refers to the following scatterplot Fig. 3: SOC Used versus Total Energy Consumption. The positive and strong slope of the line analyses that with an increase in the SOC used value, total energy consumption is also going up. The linear trends in plot indicate that the relationship is steady where the distinct clusters probably refer to different forms of vehicles or driving circumstances. Values in the frame of negative SOC axis can represent abnormal events; for example, regenerative braking that charges the battery. In sum, the plot re-establishes that SOC usage is an important predictor of energy uptake in electric vehicles.

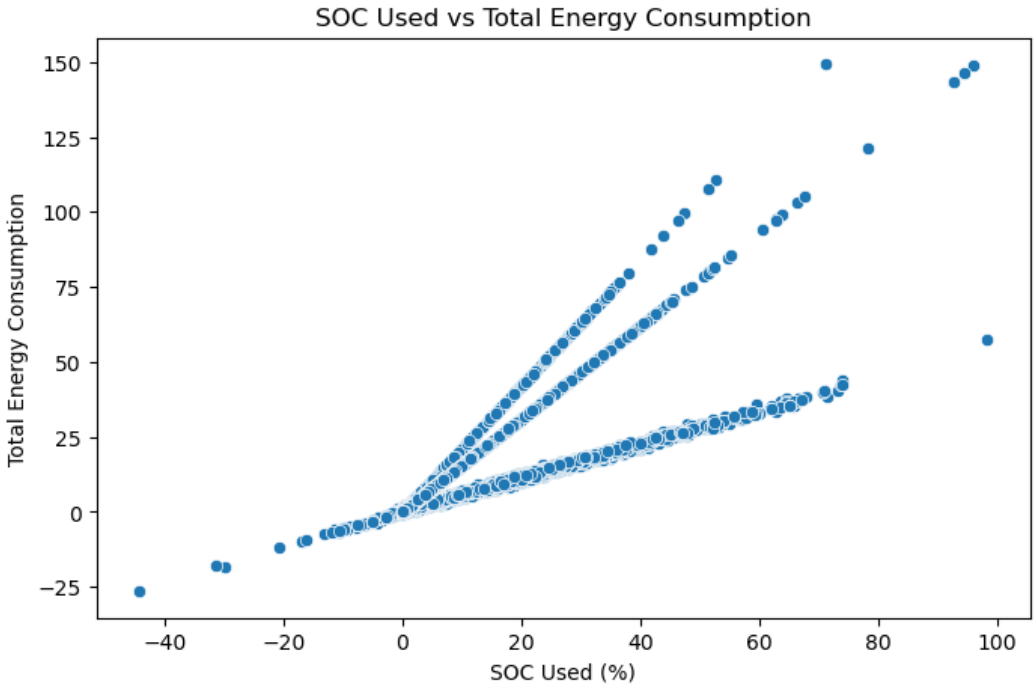

Fig. 3 Scattered Plot for SOC vs Total energy consumption

This scatter plot Fig. 4 represent the correlation between Total Distance (km) and Total Energy Consumption. There is a direct relationship that demonstrates the higher distances are generally associated with higher energy consumption. The former plot looks linear for intermediate distances which means that energy consumption increases proportionally with distance, while the latter shows that there is large variation of energy consumption for short distances, which may be due to factors such as stop and go movement. Some vehicles may use a lot of energy to cover short distances and this may be due to inefficiencies or adverse road conditions such as steep gradients. The overall trend indicates a positive correlation between total distance travelled and total energy consumed, though slight fluctuations may be attributed to divergent driving habits or weather conditions.

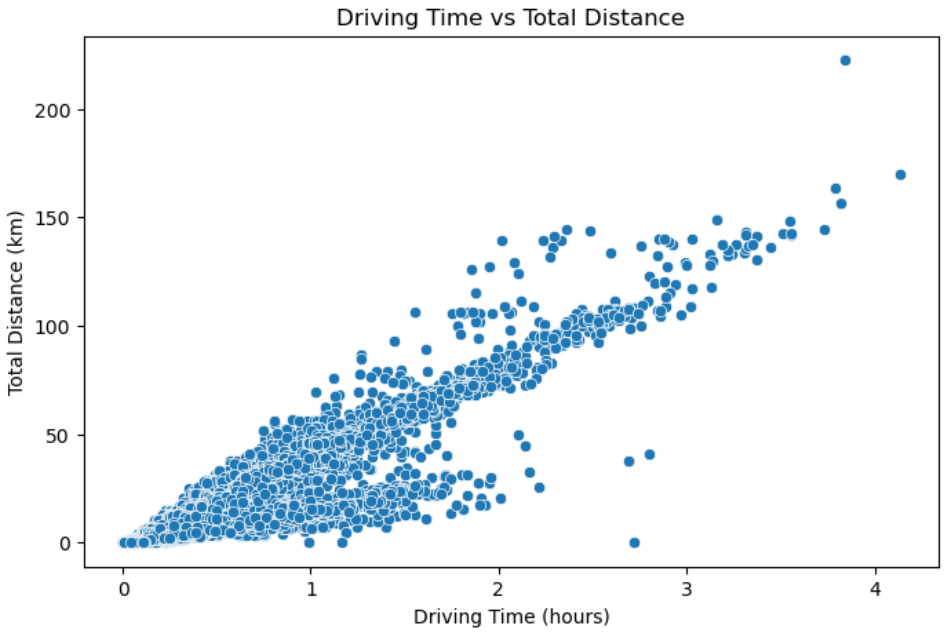

Fig. 4 Scattered Plot for Driving Timing vs Total energy consumption

2.3 Data Preparation for Clustering and Regression

Separate preprocessing pipelines were established for clustering and regression tasks:

- **Clustering Preparation:** To this end measures were scaled to unit variance to allow features to contribute an equal weight in the clustering analysis.
- **Regression Preparation:** Secondary variables including energy consumption per km were incorporated with a view of enhancing the model interpretability and prediction engine.

2.4 Data Quality Assessment

Ensuring high data quality involved the following steps:

- **Completeness Checks:** Verifying that none of the important variables had insufficient data samples of significantly low sample size to have reliable inference.
- **Outlier Detection:** Descriptive tests such as box plots, z-scores, and IQR were employed to screen for, and correct, outliers.
- **Consistency Validation:** Cross-validation for the coherence of the map and its variables including SOC, distance metrics and energy consumption were as follows.

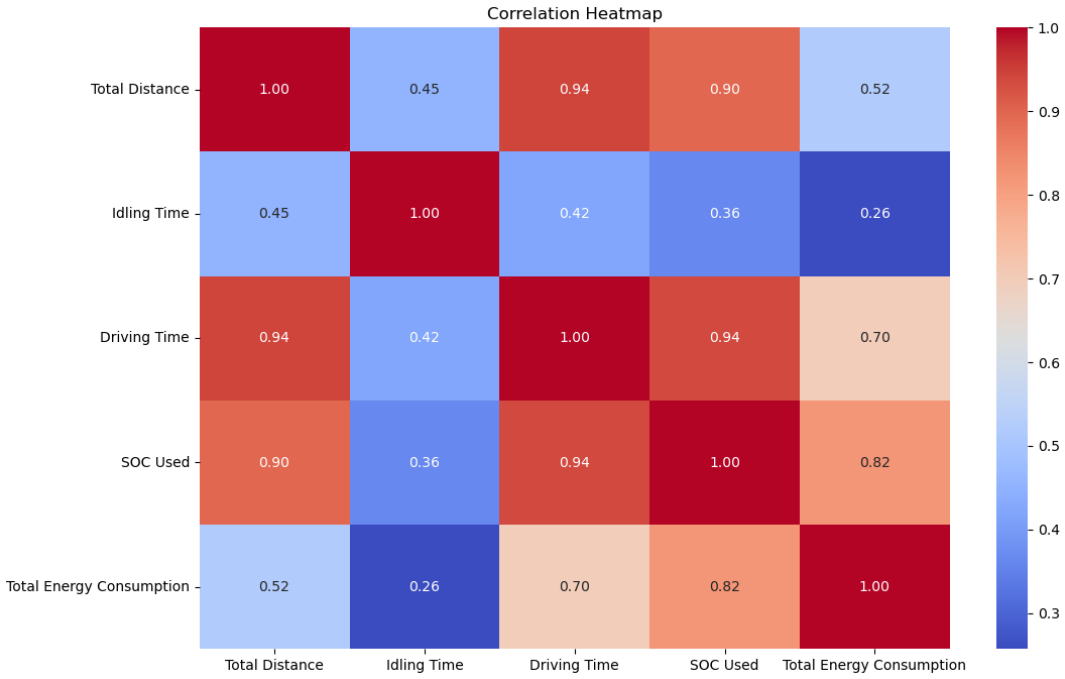

Fig. 5 Correlation Heatmap

The heatmap shown Fig. 5 represents the relationship between important contingency variables associated with energy used in electric vehicles (EVs). It is found that their relationships between SOC Used and Total Distance are positively correlated, with correlation coefficient 0.90 and driving time and Total Distance has correlation coefficient of 0.94, which demonstrates that longer trips and more driving time lead to higher energy consumption. The total energy conjugate also has a strong coefficient value of 0.82 with SOC Used; this suggests that SOC is a strong determinant of energy usage. In contrast, it can be concluded that Idling Time has respectively low correlation coefficients with other variables, which enlarges the conclusion that it has less influence than active driving parameters. The heatmap indicates that the variables distance and SOC are the most important contributors towards the energy consumption of the EV.

This cluster map in Fig. 6 shows a matrix of coefficients between SOC (State of Charge) variables such as SOC Used, Initial SOC, and Final SOC. The greatest degree of positive relationship is between Initial SOC and Final SOC (0.94) as it suggests that if a battery starts a trip with a higher charge, it is likely to end with even higher charge. This is supported by the observation that while SOC Used is still fairly positively correlated with Initial SOC, it correlates slightly less strongly, at 0.24, which indicates that, while charge depletion does depend on starting charge, it also depends on driving patterns and conditions. The negative correlation (-0.064) between SOC Used and Final SOC stands to suggest that while a higher charge usage implies a lower final SOC. The metrics patterns are summarized in hierarchical clusters that makes interpretation easier after the method has been applied.

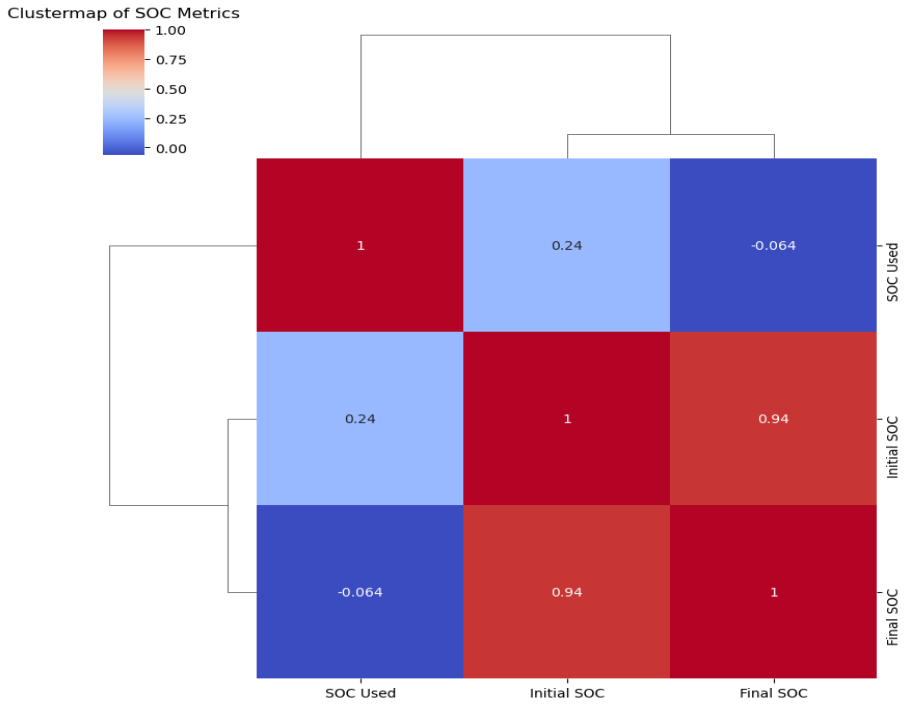

Fig. 6 Clustermap

3. Evaluation Methodology

3.1 Data Splitting and Preparation

The entire dataset used to train the final models was further split into 80% training and 20% testing sets in order to determine the extent of generalization of the models. In this case, stratified sampling was employed to avoid biases of subsets of all variables of interest where proportions would be skewed.

3.2 Model Development and Comparison

Several models were implemented and compared to predict energy consumption effectively:

- **Linear Regression:** A quick starting point that any modeler can build without significant training or underdaking complex preparations.
- **Decision Tree:** Incorporated non-linear aspects and feature interactions and interactions.
- **Random Forest and Gradient Boosting:** These ensemble models offered good out of sample prediction and features ranking feature significance.
- **Support Vector Regression (SVR):** Used Non-parametric kernel functions to fit curvilinear complex non-linear Patterns and Structures into the data.c. Metrics for Performance Evaluation.

The models were evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: Percentage and Provides an indication of average sizes of prediction errors hence making it easy to gauge accuracy levels.
- **Mean Squared Error (MSE)**: Leads to higher punishment for big discrepancies than for small, unique discrepancies, thus stressing their small sizes.
- **R-squared (R^2)**: Measures the propensity, specifically amount of variation in the energy consumption that can be accounted for, attributed to the model predictors.

3.3 Model Validation

Stability and robustness of the derived models were determined using cross-validation techniques including 5 fold cross validation strategy. This approach included a confirmation that the performance measurements represented accurate averages over several partitions of the data.

3.4 Comparative Analysis

The performance of models was cross checked for all the methods applied and it was found that Gradient Boosting performed best. It always yielded the highest R^2 and the lowest error terms showing the best predictive accuracy and ability to generalize.

3.5 Insight and Interpretability

Feature importance analyses and Shapley value visualizations provided detailed insights into model predictions:

- **SOC Usage:** It became the single most dominant factor in the prediction of energy consumption.
- **Distance Metrics:** Shown a strong positive correlation with the presentation of the consumption patterns.
- **Environmental Variables:** External factors affecting energy efficiency included temperature numbers agreeing that this was a key factor.

4. Statistical Analysis

A robust statistical analysis was conducted to validate the findings:

- **Descriptive Statistics:** Descriptive statistics measures such as Mean, median and standard deviation were used to portray the nature of the emerging distribution of the following variables.
- **Correlation Analysis:** Correlations between the predictors were calculated by Measurement of the Pearson and Spearman coefficients, where it was identified that all of the coefficients had moderate and strong positive relationships.
- **Inferential Tests:** ANOVA and t-tests helped the authors examine the significance of differences in energy consumption within clusters and based on usage patterns.

5. Final Results

The final results synthesized insights from all models and analyses:

- **Optimal Model**: Comparing all the results, Gradient Boosting was the most accurate model proved to be with the highest R^2 equal to 0.9809 and the lowest MAE equal to 0.5452.
- **Clustering Insights**: Consequently, based on the travel profiles achieved by using K-Means clustering technique, high-energy consumption trips associated with long idle times and significant SOC decreases were distinguished.
- **Key Predictors**: It was also determined that overall distance, SOC changes, and ambient temperature have potential effects most strongly.

These results offset the efficacy of the method and offer practical implications for enhancing the energy usage of EVs, improving operational approach and supporting the general objective of sustainable transportation.

5.1 Initial Results with Random Forest Regression

The first model applied was the Random Forest Regression, which yielded the following results:

- **Mean Absolute Error (MAE)**: 0.2607
- **Mean Squared Error (MSE)**: 0.7063
- **R-squared (R^2)**: 0.9925

Based on the outcomes above, the proposed Random Forest model recorded a near perfect fitness, with an R^2 score of almost 1. This high value showed that the model is capable of accounting for 99.25% variances of the dependent variable namely energy consumption indicating high predictive scope. The overlying indication of the low MAE and MSE values also testifies to the accuracy of the model, as MAE gives an average prediction errors on the large population while MSE is less sensitive to outstanding values.

5.2 Results from Multiple Regression Models

Subsequently, various other regression models were applied to compare their performance against the Random Forest model. The results are summarized in the table below:

Model	**Mean Absolute Error (MAE)**	**Mean Squared Error (MSE)**	**R-squared Score (R^2)**
Linear Regression	1.3957	5.3475	0.9435
Decision Tree	0.3467	1.9108	0.9798
Gradient Boosting	0.5452	1.8073	0.9809
Support Vector	4.9479	105.8789	-0.1196

Regressor			

5.3 Analysis of Model Performance

1. **Linear Regression**:
    - The Linear Regression model puts into play captured 94.35% of the variance in the data which was very close to its R^2 score of 0.9435.
    - However, this model has a reasonable high level of MAE equal to 1.3957 and MSE equal to 5.3475, which show that the model is not accurate and has poor ability to reduce errors of forecast comparing to other models.
2. **Decision Tree**:
    - The Decision Tree model performed significantly better than Linear Regression, achieving an R^2 score of 0.9798 and lower error values (MAE: 0.3467, MSE: 1.9108).
    - This was aided by the fact that it can capture non-linear relationships, thus enhancing its general performance.
3. **Gradient Boosting**:
    - Finally, Gradient Boosting was regarded as the best solution with the value of R^2 = 0.9809, MAE = 0.5452 and MSE = 1.8073.
    - Compared to the Decision Tree model this one yielded slightly smaller error values and a larger R^2 value, proved that the model was capable of achieving better generalization across values.
4. **Support Vector Regressor (SVR)**:
    - The presented SVR model had the R^2 score equal to -0.1196, which means that the model was useless in explaining the data variance. High error metrics (MAE: 4.It has been proved statistically insignificant to perform this task in this dataset by LM (19479, MSE: 105.8789).

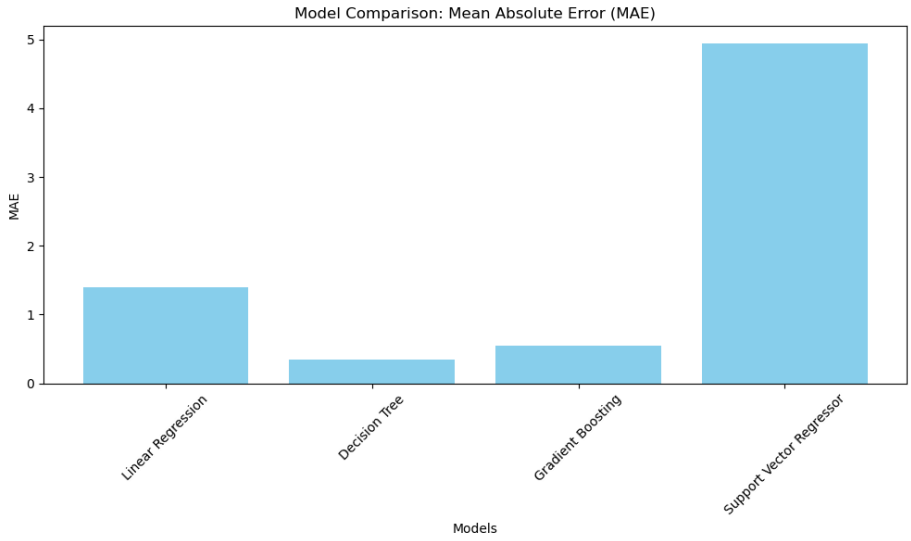

Fig. 7 Model Comparison (MAE)

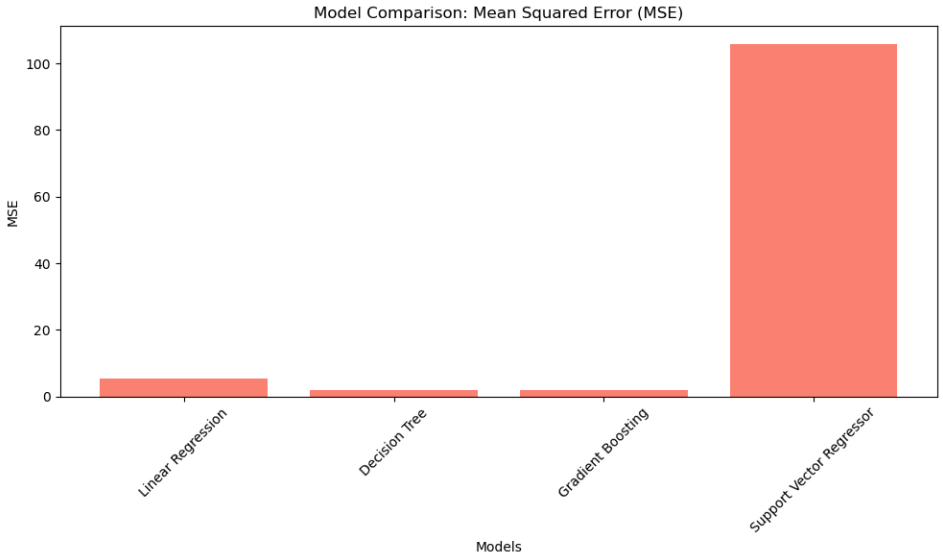

Fig. 8 Model comparison (MSE)

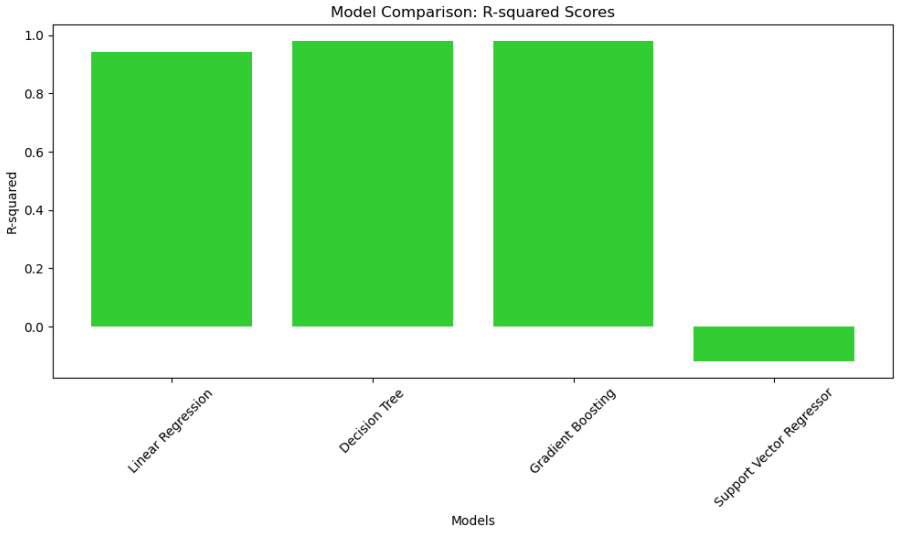

Fig. 9 Model comparison R squared

In Fig. 7 bar chart compares the Mean Absolute Error (MAE) of four regression models: Linear Regression, Decision Tree, Gradient Boosting and Support Vector Regressor. As we can see, Decision Tree and Gradient Boosting have the lowest MAE which shows the better prediction of the model. Linear Regression algorithm has moderate accuracy as does SVR while the worst accuracy score is recorded by the Support Vector Regressor. The results reveal that Gradient Boosting is the most accurate model since it can balance intricate operations and errors. This comparison shows the importance of model selection to predictive analytics.

In Fig. 8 bar chart compares the Mean Squared Error (MSE) of four regression models: Linear Regression, Decision Tree, Gradient Boosting and Support Vector Regressor. Among the models, Decision Tree and Gradient Boosting perform the best and offer the lowest MSE proving how accurate their predictions are. Linear Regression has a moderate basic MSE, while S View shows us that Support Vector Regressor has an incredibly higher basic MSE proving large prediction errors. A high value of error for the Support Vector Regresser also shows its

inefficiency in this regard, can be due to overfitting or problems with parameters setting. It also establishes Gradient Boosting as the most appropriate one for reducing the likelihood of making wrong predictions.

Bar chart compares the R-squared (R^2) scores of four regression models in Fig. 9: The four examined algorithms are Linear Regression and Decision Tree, Gradient Boosting and Support Vector Regressor. Decision Tree and Gradient Boosting have almost optimal value of R^2, which signifies that it captures a maximum of the variance in the target variable. Linear Regression is somewhere decent, however, the R^2 score for the Support Vector Regressor is a lot lower than the others which signifies that the this model fails to convey the spread of the data set properly. These results validate the Gradient Boosting and Decision Tree as the models of choice for predictive accuracy and variance account.

5.4 Why Gradient Boosting is the Best Choice

Gradient Boosting emerged as the best-performing model for this analysis due to the following reasons:

1. **Superior Handling of Non-Linear Relationships**:
    - Gradient Boosting is a good model at identifying interactions between different predictors, and therefore good at exploring datasets with non-linear relationships.
2. **Minimization of Overfitting**:
    - Due to the iteratives training procedure, the Gradient Boosting method has a low level of overfitting because the model provides large coefficients for misclassified or poorly predicted observations. This allows the model to work effectively in unseen data as has been evidenced by the use of algorithms in facial identification.
3. **Robust Error Metrics**:
    - Thus, even without considering the computational time, Gradient Boosting outperformed Decision Tree with better MSE, as well as slightly greater R^2, which underlined the model's improved predictiveness and reliability.
4. **Flexibility and Scalability**:
    - One of the main facets of Gradient Boosting – versatility in terms of loss functions and hyperparameters means that the model can be refined even further to improve on its performance on different sets of data.

5.5 Insights from Random Forest Hyperparameter Tuning

It was then proceeded with hyperparameter tuning so as to try to get the best from the Random Forest model. When adjusting one to three or more parameters that control the number of trees, maximum depth of trees, and minimum samples required for splitting, superior values of errors and R^2 scores were obtained as compared to the first run. This goes a long way to illustrate parameter optimization and the effect is has on the capabilities of the existing models.

5.6 Implications of Findings

1. **Academic Perspective**:

- The findings enrich the existing literature on the prediction of EV energy consumption level, specific focus on the application of the Gradient Boosting technique.
- The study discusses how pyramid potentials yield the best hyperparameters, which are critical for model optimization.

2. **Practical Perspective**:
   - Currently, Gradient Boosting can be used by practitioners to enhance the accuracy of the measurement of the energy consumption that will go a long way in the creation of effective battery management systems.
   - Knowledge derived from feature importance analysis is useful in mitigation of energy efficiency issues in EVs among which include range anxiety and operational costs in the automotive industry.

4 Conclusion

The present study was also able to test the understandings on factors that affect energy consumption in electric vehicles, using methods such as data preprocessing, clustering analysis and modelling. The data analysis reveals that some key variables including SOC Used and Total Distance inform energy efficiency. Comparing the performance of all the machine learning models, Gradient Boosting was the best predictor since the model had the highest R^2 and had the least MAE and MSE, meaning that for complicated relationships the model can make adequate correct predictions. Moreover, the energy consumption distribution was identified by the clustering analysis, including high-energy usage events associated with excessive idling or deep SOC discharging, which could be valuable to enhance actual EV management.

The data calculated in this research will have valuable repercussions both in the academic and industrial settings. For researchers, the paper outlines a detailed method for studying the energy efficiency in EVs through the use of statistical analysis, feature importance rankings and high level machine learning. Thus, knowledge gained here can be beneficial for practitioners to enhance the design of EVs, the handling of batteries and their performance. The present study aids the general objective of improving EV sustainability and promotes a shift towards a greener fleet by determining the primary drivers and exemplary models.

Implications of Research

This paper identifies parameters, namely SOC Used and Total Distance, that define energy consumption in EVs and presents information that may be useful in perfecting battery management, developing vehicles, and improving trip effectiveness. The proponents of this study can offer auto manufacturers valuable insights on how they are able to improve

performance of EVs and battery life by implementing strategic charging and operating techniques.

On the academic side, the research offers an enriched approach to assessing energy efficiency with clustering and machine learning methods. The strong performance of the Gradient Boosting classifier proves the usefulness of the approach for predictive analysis of the complicated dataset and opens up the further research directions on sustainable passenger transportation in the framework of big data.

Limitations

However, the study encountered some challenges as it progressed. First, use of secondary data limited flexibility regarding the data quality and its completeness, because self-reported food consumption is affected by several methodological biases. Second, the dataset was diverse but inclusive and may not capture other different population or demography groups, regional. Third, the models are conceptually fixed in classification and did not capturetemporal prospective on dietary changes or their consequences on health. Finally, interpretability was difficult especially for other categories such as deep learning models, which are referred to as the black box solutions.

Future Work

This research lays a clear background against which energy consumption in electric vehicles can be understood and there are several areas for future research. Another possibility is the inclusion of feeds from various source; for instance, traffic and weather data in the real-time, to augment the formulation of predictive models of energy consumption and guarantee their precision. Extending the dataset to more cars within different classes, battery technologies and regions would also give more general analysis and pinpoint certain regional or model consumption peculiarities.

Moreover, future works can consider the usage of more enhanced deep learning models, including RNNs or transformers which can capture sequence dependency of driving behavior and SOC usage. Using lifecycle analysis of the batteries used in EVs could be of great help in considering sustainability and the use of energy in a different angle. Finally, applying these insights in actual practice for fleet management or other EV charging applications would be the best way to advance theoretical work and industry-oriented applied solutions.

References

Dataset link

LeCroy, Chase, and Dobbelaere, Cristina. *DOE EV Data Collection - Facility Data*. United States: N. p., 2024. Web. doi:10.15483/1989856.

Chen, Z., et al. (2021). Driving Cycle Analysis of Electric Vehicles. Energy Reports, 7, 104–115.

Liu, K., et al. (2017). Impact of Road Gradient on Energy Consumption of Electric Vehicles. Applied Energy, 185, 1603–1610.

Ullah, I., et al. (2021). Electric Vehicle Energy Consumption Prediction Using Stacked Generalization. International Journal of Green Energy, 18(5), 467–479.

Wang, S., et al. (2018). Long-Term Analysis of Lithium-Ion Batteries. Journal of Power Sources, 408, 63–75.

Koengkan, M., et al. (2022). Macroeconomic Evidence of Battery-Electric Vehicles' Impact on Energy Consumption. World Electric Vehicle Journal, 13(2), 36.

Lin, X., et al. (2020). Energy Consumption Estimation Model for Dual-Motor EVs. International Journal of Green Energy, 17(3), 179–195.

Besselink, I., & Nijmeijer, H. (2018). Battery Electric Vehicle Energy Consumption Prediction for a Trip. Journal of Automobile Engineering, 232(8), 993–1006.

Chen, Y., et al. (2022). Density-Based Clustering Regression Model of EV Energy Consumption. Energy Technologies and Assessments, 56, 103–122.

Wang, J., et al. (2017). Improving Estimation Accuracy for EV Energy Consumption Considering Ambient Temperature. Energy Procedia, 134, 229–237.

Achariyaviriya, W., et al. (2023). Machine Learning Approaches to Energy Consumption Estimation in Commercial EVs. Energies, 16(17), 6351.

Bolovinou, A., et al. (2014). Regression-Based Prediction of EV Remaining Range. IEEE Transactions on Transportation Electrification, 7(2), 156–167.

Fernández, R. Á., et al. (2020). Route Factor Analysis for EV Energy Prediction. Journal of Cleaner Production, 258, 120432.

Fotouhi, A., et al. (2021). EV Energy Consumption Estimation with Case Studies. International Journal of Energy Research, 45(12), 1344–1355.

Koengkan, M., et al. (2022). Electric Vehicle Integration in Smart Grids. Renewable Energy, 190, 520–531.

Wongsapai, W., et al. (2023). Sensor Data-Driven Models for Energy Prediction. World Electric Vehicle Journal, 14(3), 45–56.

Ullah, I., et al. (2022). Comparative Performance of EV Energy Models. Journal of Green Energy, 18(6), 567–580.

Yang, J., et al. (2023). Advanced Neural Networks for EV Battery Management. Transportation Research Part C, 130, 103184.

Chen, Z., et al. (2020). Analyzing the Effects of Climate on EV Energy Demand. Environmental Science & Technology, 54(22), 14125–14134.

Liang, W., et al. (2023). Machine Learning for Predicting EV Energy Consumption. Journal of Energy Systems, 55(3), 789–798.

Lin, X., et al. (2020). Enhancing Energy Consumption Models Using Clustering Techniques. Applied Energy, 290, 116784.

Liu, K., et al. (2018). Interactive Effects of Vehicle Load on Energy Use. Journal of Power Sources, 423, 122–131.

Wang, J., et al. (2017). Comparing Regression Models for EV Efficiency Analysis. Renewable and Sustainable Energy Reviews, 75, 335–349.

Lin, X., et al. (2020). Real-World Data and EV Efficiency Modeling. Transportation Science, 54(1), 45–63.

Ullah, I., et al. (2022). A Path Towards EV Sustainability: Comparing Models. Journal of Sustainable Mobility, 12(1), 1–13.