

Article

M5Boost: A Machine Learning Approach for Driving Range Estimation in Electric Vehicles Considering Battery-Related Factors

Ibrahim Atakan Kubilay ¹, Kadriye Filiz Balbal ², Kokten Ulas Birant ¹ and Derya Birant ^{1,*}

¹ Department of Computer Engineering, Dokuz Eylul University, Izmir 35390, Turkey; atakan@cs.deu.edu.tr (I.A.K.), ulas@cs.deu.edu.tr (K.U.B.)

² Department of Computer Science, Dokuz Eylul University, Izmir 35390, Turkey; kadriyefiliz.balbal@deu.edu.tr

* Correspondence: derya@cs.deu.edu.tr

Abstract

Range estimation for electric vehicles (EVs) is critical for intelligent transportation systems since it directly affects charging planning, route optimization, driver confidence, energy management, battery utilization, and driver decision-making processes. However, current studies still suffer from issues such as limited accuracy, insufficient interpretability, high computational complexity, dependence on simulation environments, or insufficient generalization capability under dynamic driving conditions. To address these limitations, this paper proposes an M5Boost framework that successfully integrates an additive residual learning methodology with the model tree structure. Unlike conventional boosting approaches, M5Boost combines iterative residual-driven learning, multivariate leaf regression models, tailored tree pruning, and specific smoothing mechanisms to improve prediction accuracy, robustness, and generalization capability for EV range estimation. A benchmark dataset was further systematically extended with newly collected real-world battery-related driving records. Experimental validation showed that the developed model significantly outperformed state-of-the-art models reported in the literature on the same dataset.

Keywords: electric vehicles; battery; machine learning; range estimation; regression; boosting; transportation systems

1. Introduction

Electric vehicles (EVs) have become an important component of modern transportation ecosystems due to their low greenhouse gas emissions and high energy efficiency. Despite these advantages, limited battery capacity and charging infrastructure pose substantial challenges to long-distance travel. In this context, accurate EV range prediction has emerged as a critical research problem for improving driving reliability and user confidence.

Range estimation is the analytical process of predicting the maximum distance an electric vehicle can travel under specific operational and environmental conditions before its usable energy capacity is exhausted [1,2]. It is currently an active topic in intelligent transportation systems due to the growing shift toward sustainable EVs and the increasing demand for robust distance-prediction systems [3]. Providing reliable projections of

Academic Editor: Xianglin Li

Received: 8 June 2026

Revised: 9 July 2026

Accepted: 15 July 2026

Published: 16 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

driving distance is required to mitigate the critical issue of ‘range anxiety’—the psychological distress—experienced by users, thereby encouraging drivers to switch from conventional gasoline vehicles to electric mobility [4]. Accurate range estimation directly affects driver confidence, charging planning, route optimization, energy management, battery utilization, and driver decision-making processes.

Range estimation and State-of-Charge (SoC) estimation are two closely related yet fundamentally different tasks in electric vehicle research. While SoC represents the instantaneous ratio of the available battery energy to total battery capacity, typically expressed as a percentage between 0% and 100%, range estimation projects this available energy to spatial travel distance (km) based on vehicle dynamics and driving conditions. SoC estimation is primarily an internal electrochemical monitoring task that quantifies the remaining capacity inside the battery pack using local cell sensors (i.e., voltage, current, temperature). In contrast, range estimation is further compounded by highly stochastic external variables, including driving style, vehicle characteristics (e.g., tire type), ambient climate, route type, and auxiliary power consumption.

Range prediction is a highly challenging task because energy consumption in EVs is influenced by numerous dynamic and uncertain internal and external factors [5]. For instance, environmental conditions such as ambient temperature, wind resistance, and road gradient can significantly affect battery efficiency and energy usage. Similarly, operational parameters, including vehicle speed, acceleration, driving style (i.e., aggressive, normal, or eco-friendly), auxiliary heating/cooling use, and traffic situations, may drastically impact the effective driving range of EVs. Moreover, vehicle-specific parameters such as vehicle mass, battery characteristics, battery aging level, powertrain efficiency, and tire properties can also directly influence EV energy consumption behavior and, therefore, driving range performance. The interactions among all these factors are often highly complex and difficult to model using traditional linear approaches. The presence of nonlinear dependencies, uncertain operational conditions, heterogeneous driving patterns, and a lack of high-quality driving datasets further complicates the EV range estimation problem. Despite initial studies in this field, achieving highly accurate and stable range prediction under real-life conditions remains an open research problem.

The current literature reveals several limitations and drawbacks in existing range estimation frameworks [1,3]. Some previous data-driven studies relied on simulation-based datasets that might not fully represent real driving variability [1]. Some studies used simplified linear models that may fail to capture the stochastic and complex nature of driving operations [3]. Conversely, while deep learning approaches could offer high precision, they are often regarded as “black-box” models, lacking the interpretability required for critical transportation systems. In addition, they usually require computationally expensive architectures and extensive parameter processes [3]. Additionally, some previous studies still suffer from issues such as reliance on expensive optimization strategies, a lack of high-quality driving datasets, and insufficient environmental variability under unseen driving conditions [1]. These limitations indicate that further research is still needed to develop more robust and practical EV range-estimation frameworks.

This study aims to overcome the aforementioned challenges through the proposed M5Boost framework, which simultaneously addresses predictive accuracy, efficiency, and model transparency. By leveraging a real-world dataset, this research focuses on enhancing the reliability of range estimation for Volkswagen EVs.

The main contributions of this study can be summarized as follows:

- (i) It proposes an M5Boost framework for the EV range estimation that establishes an additive regression mechanism with model trees in order to iteratively reduce prediction errors while simultaneously constructing localized regression models.

- (ii) Unlike conventional boosting approaches that utilize constant-valued leaf nodes, the proposed approach employs multivariate regression models at leaf nodes to generate more flexible and region-specific predictions.
- (iii) The proposed approach incorporates specialized model tree pruning and smoothing operations within the boosting process to improve prediction continuity, robustness, and generalization capability.
- (iv) It also contributes to the literature by systematically extending a benchmark EV dataset with newly collected real-world battery-related driving records, resulting in a comprehensive dataset spanning more than 11 years.
- (v) Experimental validations revealed that the developed model significantly outperformed the state-of-the-art models reported in the literature on the same dataset.

The structural flow of this research is outlined as follows. Section 2 synthesizes current knowledge by reviewing existing literature on EV range prediction. Section 3 describes the theoretical methodology of M5Boost, along with its formal mathematical foundations. Section 4 serves as the experimental core, detailing the dataset, feature engineering operations, implementation details, the empirical setup, a comprehensive analysis of the results, and statistical evaluation findings. Section 5 presents a comparative performance assessment against competing benchmarks to validate the effectiveness of the approach. Finally, Section 6 concludes the study and discusses potential avenues for future advancements in the field.

2. Related Work

As the transition to electric mobility accelerates, researchers have focused on developing intelligent models capable of predicting driving range under diverse operational and environmental conditions. They have attempted to address “range anxiety” by building diverse EV range forecasting frameworks aimed at providing driver confidence. The studies still struggle to provide a balance among high-precision forecasting, computational efficiency, and interpretability across volatile real-world driving scenarios.

In the literature, EV range estimation approaches are generally categorized into 3 particular groups: model-based, rule-based, and data-driven methods. (i) Model-based approaches [6,7] rely on mathematical and physical formulations derived from vehicle dynamics, battery behavior, and energy consumption equations. These approaches generally require detailed physical parameters and simplified system assumptions, which may limit their applicability under highly dynamic real-world conditions and complex driver behaviors. (ii) Rule-based approaches [8,9], such as Fuzzy Logic, utilize predefined heuristics, expert-system-driven techniques, and manually designed decision mechanisms to handle driving inputs. However, these systems are frequently limited by their inability to adapt to data patterns not explicitly defined by the designer. Furthermore, their performance highly depends on the quality of manually designed rules. (iii) Data-driven approaches [10], on the other hand, utilize machine learning algorithms to learn relationships directly from historical, large-scale, high-dimensional driving data. Moreover, hybrid approaches, i.e., integrating physical constraints with data-driven insights, have also been investigated for EV range prediction. However, their integration may increase computational complexity without a proportional gain in practical reliability.

Table 1 presents a comprehensive summary of recent studies related to EV range estimation, including their vehicle brands, employed machine learning methods, geographic information, prediction objectives, strengths, and reported limitations. This table highlights the evolving landscape of the field, moving from basic regression models to complex ensemble and deep learning architectures.

Table 1. Summary of recent studies on range estimation in electric vehicles.

Ref.	Year	Methods	Country	Vehicle	Remaining Range?	Driving Range?	Strengths	Limitation
[11]	2026	ANN, RF, GBM, KNN	India	MG, Tata	Yes	No	Considers disturbances like wind and road grade	Offline calculation
[12]	2026	CatBoost, RF, LightGBM, XGBoost, SVM, ANN, GRU, LSTM	China	Generic	Yes	No	Bayesian, grid, or random search optimization	Dependent on optimization
[13]	2026	LR, ANN, GBM, SVM, KNN, LightGBM	India	Tata	Yes	No	Impact of passenger count and auxiliary loads	Small data; simulated factors
[14]	2026	ANN, SVM, Mechanistic	Simulation	Simulation	No	Yes	Multi-domain modeling	Simulation dataset, no direct physical experiments
[15]	2026	RF, ANN, XGBoost, kNN	China	Cars & Buses	Yes	No	Incorporates SOH	Lacks detailed road grade and traffic info
[16]	2026	LightGBM, GRU	USA, CR, Chile data	Nissan, Hyundai	No	Yes	Online prognostics	Limited to specific vehicle models
[17]	2025	XGBoost, DEKF	Simulation	Simulation	Yes	No	Co-estimation of SOC/SOH	Simulation dataset, relies on Kalman Filter for internal states
[18]	2025	KAN, RF, DT, KNN, XGBoost, CatBoost, GBM, LightGBM, ET, LSTM, GRU, RNN, Transformer	China	Five EVs of the same model	Yes	No	Feature selection and parameter optimization	High computational complexity for KAN
[19]	2025	ET, CatBoost, LightGBM, XGBoost, Voting, RF, GBM, DT, AdaBoost, LR, Ridge, Lasso, ElasticNet, SVM, KNN	Turkey	Musoshi	Yes	No	Analysis of factors, campus data	Small route; ignores traffic density
[20]	2025	HMM, FMM, XGBoost	China	Generic	Yes	No	Captures personalized driving behavior	High initial prediction deviations
[21]	2025	CNN-Transformer, LightGBM, GRU, ANN, LSTM	China	Six EVs	Yes	No	Multi-source data-driven	Does not analyze traffic flow patterns
[22]	2025	Deep Q-Learning, LSTM, XGBoost, LightGBM	Web data	Multi-brand (103 EVs)	No	Yes	Adaptive model	Lacks grid-level and fleet-level coordination
[23]	2025	GAN, ANN	USA	Generic	No	Yes	Federated learning, privacy-preserving	Struggles with data distribution skew
[24]	2025	BiLSTM-KAN, LSTM, TCN, GRU, CNN, Transformer	China	Tesla, BYD, Toyota, GAC	Yes	No	Introduces battery capacity index	Short data time span; no HEV validation
[25]	2024	DNN, RF, GBM, SVM, XGBoost, LR, Ridge, Lasso, DT, KNN	Holland	Multi-brand	No	Yes	Optimization	Simulated environmental factors

[26]	2024	ANN, RF, Ada-Boost, GBDT, LSTM	China	Generic	No	Yes	Estimation for electric buses	Might not accurately predict extreme values
[27]	2023	XGBoost, RF, GBM, Lasso, Ridge	USA	BYD, Freightliner	No	Yes	Models for heavy-duty EVs	Fleet-specific duty cycle knowledge gap
[28]	2023	SVM	USA	Nissan	Yes	No	Tailored for rural applications	Lacks deep learning integration
[29]	2023	LR, DT, RF, KNN, AdaBoost	Multiple	Nissan, Rover	Yes	No	Regression ensembles	Performance highly dependent on the dataset
[30]	2022	ANN, LSTM, RF	Canada	Kia	Yes	No	Performance in sub-freezing winter conditions	Lacks AC energy consumption analysis
[31]	2022	LR, Lasso, Elastic-Net, SVM	Web data	Tesla, Nissan, Fiat, etc.	No	Yes	Model specifications like curb weight and capacity	Higher error than power models
[32]	2022	HMM	China	Generic	Yes	No	Uses online traffic API	High relative error
Proposed		M5Boost	Germany	Volkswagen	No	Yes	Comprehensive study	User interface-driven accessibility

Machine learning approaches have been investigated for EV range estimation problems due to their ability to model relationships between driving conditions and energy consumption characteristics. Traditional ML techniques such as K-Nearest Neighbors (KNN) [11], Support Vector Machine (SVM) [12,13], and Artificial Neural Networks (ANN) [13–15] have been employed in previous studies. Despite their use, these methods often suffer from high sensitivity to noise or fall short of yielding satisfactory performance. Ensemble learning approaches, including Random Forest (RF) [11,15], CatBoost [12], LightGBM [16], XGBoost [11,17], and Extra Trees [18,19] have also been utilized for EV range prediction tasks. In addition, probabilistic sequential learning approaches such as Hidden Markov Models (HMM) [20] have been implemented to capture stochastic driving behaviors. More recently, deep learning-based approaches, including Convolutional Neural Networks (CNN) [21], Deep Q-Networks [22], Generative Adversarial Networks (GAN) [23], Long Short-Term Memory networks (LSTM) [18], Bidirectional LSTM (BiLSTM) [24], and Deep Neural Networks (DNN) [25], have been proposed to process EV driving data. However, these DL models often rely on computationally expensive architectures or black-box learning structures with limited interpretability.

The source of data remains a critical factor in model validation. While some studies relied on simulation environments [14,17] to generate synthetic driving cycles and battery behaviors, others [26–32] utilized real-world operational EV datasets collected under practical conditions. Since simulation-based studies provide controlled experimental environments, they may not fully reflect the variability and uncertainties encountered in real-life EV usage.

A diverse range of EV brands has been analyzed in the literature, including vehicles from manufacturers such as Tata [11,13], Nissan [16,28,29,31], Hyundai [16], MG [11], Motoshi [19], Tesla [24,31], BYD [24], Rover [29], MG [11], GAC [24], Kia [30], Fiat [31], and Toyota [24]. In addition to passenger vehicles, several studies [15,26,27] have extended their scope to electric buses. Furthermore, range estimation challenges have also been explored for specialized heavy-duty electric vehicles, such as those within the Freightliner fleet [27]. In the present study, the experimental analysis specifically focuses on Volkswagen e-Golf EVs using a real-world battery-related driving dataset collected from different users.

Geographically, the investigations span different regions worldwide with varying climatic and infrastructural characteristics. Significant research has been conducted in China [12,15,18,20,21,24,26,32], followed by the United States [16,23,27,28], India [11,13], Turkey [19], Holland [25], and Canada [30]. Some studies additionally utilized multinational or web-based datasets obtained from multiple countries [16,22,29,31]. All these findings indicate that EV range estimation research has become a globally active research area with increasing industrial and academic interest. The empirical analysis in this study is based on a comprehensive dataset sourced from a German web-based platform that reflects real-world driving conditions in the region.

The literature further indicates that EV range prediction studies are typically bifurcated into two major categories: Driving-Range Estimation (DRE) and Remaining-Range Estimation (RRE). Although both aim to estimate a distance for an EV, they differ in their prediction objectives and input data. DRE refers to predicting the distance that an EV can travel under specified operating conditions, typically considering vehicle characteristics, driving style, traffic conditions, environmental factors, and battery-related parameters such as battery State-of-Health (SoH). For example, DRE can be used before or at the beginning of a trip to estimate how far a vehicle is expected to travel under planned driving conditions. In contrast, RRE focuses on predicting how far the EV can still travel based on its current battery state during an ongoing trip. In other words, it primarily relies on the current battery State-of-Charge (SoC) together with instantaneous vehicle measurements to continuously update the remaining driving distance. For example, RRE computes that an EV operating with an active 45% SoC under aggressive highway driving behavior can travel exactly 115 km further before full depletion. Accordingly, DRE mainly supports trip planning and route selection, whereas RRE is intended for real-time and continuous range monitoring during vehicle operation. Several studies concentrated on DRE problems [16,25,26,33], whereas other studies focused on RRE tasks [12,15,17,28,34]. The existing studies primarily focus on one of these estimation types rather than investigating both of them.

As performance evaluation metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are universally employed benchmarks due to their suitability for regression-based estimation problems. MAE-based evaluations were frequently reported in studies such as [26–29], whereas RMSE-based analyses were emphasized in studies including [30,31]. These metrics provide quantitative insight into the deviation between actual and predicted EV range values under different driving conditions. However, some previous studies [32] report only a single evaluation metric, which may limit the comprehensiveness of model performance assessment.

In addition to ML model development, several studies have examined how various operational and environmental variables impact the precision of range estimation. It has been noted that vehicle mass serves as a prominent factor affecting EV range behavior [19]. Furthermore, road slope and elevation variations were shown to affect energy consumption and driving distance significantly [19,35]. Ambient temperature levels were similarly recognized as critical factors due to their effects on battery efficiency and auxiliary energy demands [19,36,37]. Driving cycles remain essential for model calibration [14,31,32]. Moreover, factors such as vehicle speed [19,38], acceleration patterns [19,39], and specific driving modes (e.g., Aggressive, Normal, or Gentle) [40] were likewise reported as influential variables affecting the accuracy of EV range estimation.

Alternative methodologies adopted a two-stage estimation strategy in which the battery State of Charge (SoC) is first predicted and subsequently utilized for EV range estimation [28,41–43]. Although this approach also provides battery-state information, prediction errors occurring during the SoC estimation stage may propagate into the final

range estimation results, thereby increasing the overall uncertainty of the prediction framework. In addition, this two-stage approach adds computational complexity.

As shown in Table 1, previous data-driven studies still suffer from issues such as limited accuracy, insufficient interpretability, dependence on simulation environments, reliance on expensive optimization strategies, lack of environmental variability, high computational complexity, and reduced generalization capability under unseen driving conditions. These limitations indicate that further research is needed to develop more robust and practical EV range-estimation frameworks.

Unlike previous studies, this research investigates the M5Boost framework that integrates the robust error-correction capabilities of additive regression with M5 Model Trees for the range estimation task. In contrast to existing studies that rely solely on conventional boosting algorithms, M5Boost combines residual-based iterative learning with localized multivariate regression modeling. M5Boost aims to provide an interpretable and computationally efficient prediction structure while maintaining high prediction accuracy. By integrating specialized tree-pruning and smoothing stages into a boosting loop, the approach provides a highly precise and continuous prediction surface that bridges the gap between traditional linear models and complex “black-box” deep learning algorithms.

3. Materials and Methods

3.1. Proposed Framework

This study proposes a comprehensive ML-based framework for precise driving-range estimation in electric vehicles (EVs). The framework was designed to transform raw vehicle usage records into an intelligent model capable of accurately estimating the distance that an EV can travel under various driving conditions. The developed framework is intended to support transportation systems, energy management, trip/charging planning, and decision-making processes by providing reliable range predictions. Given the complex non-linear dependencies inherent in EV energy depletion, it offers utility in alleviating range anxiety under dynamic real-world operational conditions.

The core element of the proposed framework is an M5Boost-based machine learning methodology that combines additive residual learning with a model-tree-based prediction structure. M5Boost iteratively fits base model trees to the residual errors of prior iterations to effectively capture both localized multi-variate trends and global non-linear variations across heterogeneous driving datasets. Critically, it leverages the interpretability advantages inherent in model trees. Incorporating dedicated pruning and smoothing operations, the proposed approach aims to improve prediction accuracy, robustness, and generalization capability for practical EV range estimation scenarios.

Figure 1 illustrates the entire procedural architecture of the proposed framework. It delineates the pipeline from raw data harvesting to the final presentation layer. Specifically, the proposed framework systematically integrates data acquisition, preprocessing, feature selection, model training, hyperparameter optimization, performance evaluation, and decision-making stages.

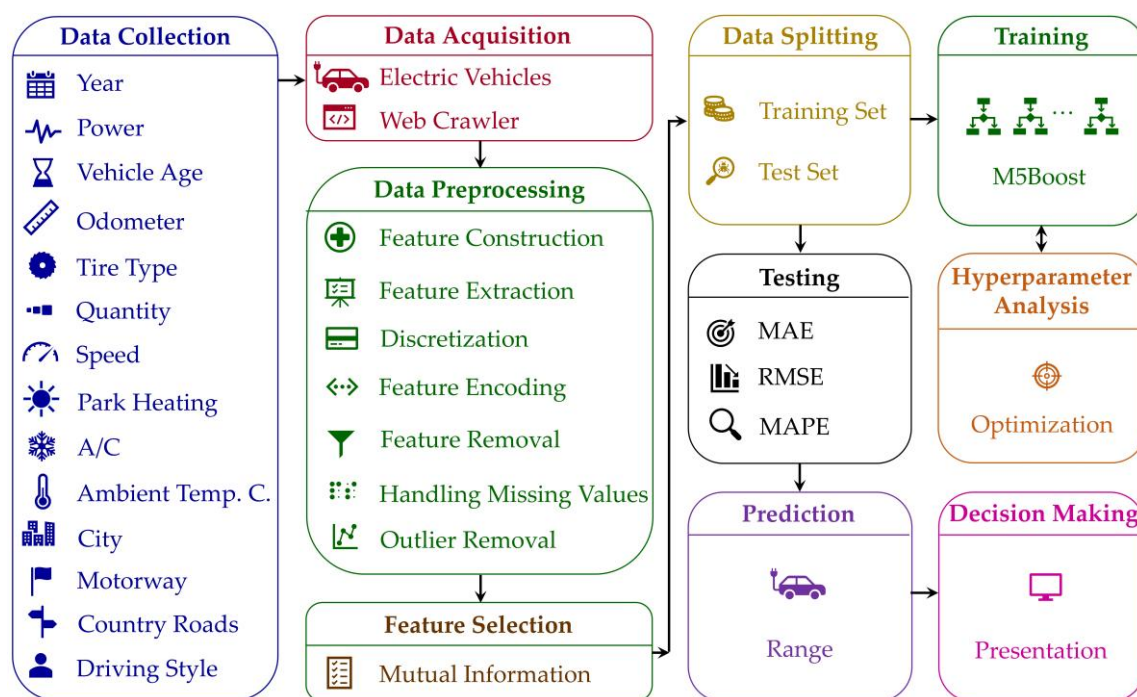


Figure 1. Overall architecture of the proposed M5Boost-based framework for electric vehicle driving range estimation.

The framework begins with the “Data Collection” stage, where crucial factors affecting EV driving range are obtained from drivers. These variables include vehicle-related specifications such as production year, power, and vehicle age; energy-related metrics such as consumed energy quantity; operational parameters including odometer reading and average driving speed; environmental factors such as ambient temperature category and road usage types including city, motorway, and country roads; vehicle equipment activation variables such as air conditioning (A/C) and park heating; and driver behavior characteristics represented by driving style.

The “Data Acquisition” stage uses a specialized web crawler to automatically retrieve, extract, aggregate, and store vehicle records in a structured dataset for subsequent analysis. To ensure high data quality, a “Data Preprocessing” stage is executed, consisting of multiple sequential operations, including Feature Construction, Feature Extraction, Discretization, Feature Encoding, Feature Removal, Handling Missing Values, and Noise/Outlier Removal. Feature construction generates new variables, such as vehicle age, while feature extraction derives additional information from existing attributes, such as driving month information from the date value. Discretization is applied to improve data representation, whereas feature encoding converts categorical variables into machine-readable numerical formats. Redundant attributes are eliminated during feature removal, missing records are handled appropriately, and abnormal observations are identified and removed using outlier detection techniques.

Once the data is cleaned, a “Feature Selection” stage is employed to identify the most informative variables for EV range estimation. In this study, the Mutual Information method was used to mathematically assess the dependency between the input features and the target variable. The aim of this stage is to discard irrelevant information, reduce computational cost, and improve model accuracy. The resulting dataset then undergoes “Data Splitting” to be divided into two distinct subsets: the training set and the test set. The training set is used to construct predictive models, while the test set is reserved for performance evaluation.

During the “Training” phase, the core M5Boost algorithm is fitted to the training partition and operates in direct coordination with the “Hyperparameter Analysis” stage to dynamically determine suitable model parameters. The optimization process explores various parameter combinations to improve predictive performance while maintaining generalization capability. Subsequently, the “Testing” phase utilizes statistical metrics, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) to evaluate the effectiveness of the trained model on the test set. These metrics quantify the deviation between actual and predicted driving ranges and provide an objective evaluation of model accuracy. Finally, the framework proceeds to the “Prediction” stage to generate range predictions for unseen EV driving conditions. The predicted range values are then used in the “Decision-Making” stage, where results are presented to drivers to support operational planning, route selection, charging management, and other range-related decisions. Through this end-to-end workflow, the proposed framework provides a practical and effective solution for EV range estimation applications.

The advantages of M5Boost can be summarized as follows:

- **Enhanced Non-linear Mapping:** By combining additive regression with model trees, M5Boost can approximate highly complex non-linear functions more accurately compared to a single tree.
- **Reduced Bias:** The iterative nature of M5Boost systematically reduces the bias of the model by focusing on previously unexplained variance in each step; therefore, it successively minimizes global error.
- **Interpretability:** Since each learner in M5Boost is a model tree, the decision boundaries remain human-interpretable compared to black-box boosting methods. Furthermore, leaf coefficients provide insights into feature importance.
- **Handling High Dimensionality:** M5Boost effectively performs feature selection during the tree-building process, focusing on the most relevant predictors for residual reduction.
- **Smooth Predictions:** Through the integration of child and parent predictions, M5Boost produces smoother output transitions across the feature space than standard regression tree ensembles. In this way, it reduces abrupt changes in estimates between neighboring regions.
- **Advanced Leaf Nodes:** Unlike traditional Gradient Boosting that uses constant values in leaves, M5Boost employs multivariate regression models at leaf nodes, allowing the method to generate more flexible and region-specific predictions.
- **Robustness to Overfitting:** The shrinkage parameter in the additive stage provides a dual-layer defense against noise.
- **Efficient Handling of Continuous Data:** M5Boost is natively designed for regression tasks involving continuous numerical features, eliminating the need for extensive data discretization.
- **Efficiency:** M5Boost is computationally efficient compared to many deep learning-based methods while still maintaining high predictive performance.
- **Stability:** The pruning mechanism reduces unnecessary model complexity and improves model stability.

3.2. M5Boost-Based Methodology

This study proposes a boosting-based approach, referred to as M5Boost, that is based on additive regression and model trees. Specifically, it integrates the M5 model tree structure with an additive regression meta-learning strategy in order to improve prediction accuracy for estimation problems. This approach iteratively constructs multiple M5 Model

Trees, where each newly generated model aims to reduce the prediction errors of the previous ensemble.

Let the training dataset be represented as:

$$D = \{(x_i, y_i)\}_{i=1}^n \quad (1)$$

where n is the number of training instances, $x_i \in \mathbb{R}^d$ represents the d -dimensional input feature vector of the i -th instance, and $y_i \in \mathbb{R}$ refers to the corresponding target variable. Each instance is defined in the feature space:

$$x_i = x_{i1}, x_{i2}, \dots, x_{id} \quad (2)$$

where x_{ij} is the j -th feature value in the i -th instance.

The M5Boost process begins with an initialization step, $F_0(x)$, which is defined as the mean value of the target attribute across the training set:

$$F_0(x) = \frac{1}{n} \sum_{i=1}^n y_i \quad (3)$$

This initial global average serves as the baseline for the subsequent additive updates, ensuring that the first tree focuses on the variance around the mean.

In each t -th iteration, M5Boost calculates the residual for each instance, which represents the unexplained variance of the current ensemble. The residual $r_i^{(t)}$ for instance i at iteration t is formulated as:

$$r_i^{(t)} = y_i - F_{t-1}(x_i) \quad (4)$$

where $F_{t-1}(x_i)$ denotes the ensemble prediction obtained from the previous iteration. Following the residual calculation, a new training subset is defined as:

$$D^{(t)} = \{(x_i, r_i^{(t)})\}_{i=1}^n \quad (5)$$

This $D^{(t)}$ dataset is then used to build a new M5 Model Tree, h_t . Unlike conventional standalone regression trees, M5Boost incrementally refines the prediction model by adding correction functions generated by successive M5 Model Trees. The ensemble is updated at each iteration by adding the newly trained tree to the existing model. The update rule for the t -th iteration of M5Boost is expressed as:

$$F_t(x) = F_{t-1}(x) + v h_t(x) \quad (6)$$

where $v \in (0, 1]$ denotes the shrinkage parameter controlling the contribution of each M5 Model Tree to the final prediction model. The shrinkage mechanism reduces the impact of individual learners and improves the generalization capability of M5Boost by preventing excessively large corrective updates during training.

After T boosting iterations, the final M5Boost prediction model is defined as the cumulative sum:

$$F_T(x) = F_0(x) + v \sum_{t=1}^T h_t(x) \quad (7)$$

where $h_t(x)$ denotes the t -th M5 Model Tree and v represents the shrinkage parameter (learning rate) used to control the contribution of each tree. This final M5Boost model provides a comprehensive mapping of the input features to the target by systematically reducing the global error through these local approximations.

M5Boost minimizes the Mean Squared Error (MSE) loss function defined as:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n (y_i - F_T(x_i))^2 \quad (8)$$

At each iteration, the newly generated M5 Model Tree attempts to minimize the remaining residual error of the current ensemble model. Consequently, M5Boost gradually improves prediction performance by sequentially correcting previously misestimated observations.

The M5Boost model can simultaneously capture nonlinear partitioning structures within data. This modeling capability makes M5Boost particularly suitable for complex EV range estimation problems involving heterogeneous driving conditions, nonlinear energy consumption behavior, and environmental variations.

3.3. Formal Description of M5Boost

M5Boost builds multiple model trees within the additive regression framework to iteratively minimize residual prediction errors. In the ensemble framework, each tree recursively partitions the feature space into multiple subregions and constructs local regression models at leaf nodes. This learning structure enables M5Boost to simultaneously capture nonlinear feature interactions and localized relationships within estimation problems.

Let the dataset associated with a tree node be represented as:

$$D_q = \{(x_i, y_i)\}_{i=1}^m \quad (9)$$

where m denotes the number of samples belonging to node q .

Tree Construction:

During the tree construction stage of M5Boost, candidate splitting attributes are evaluated using the Standard Deviation Reduction (SDR) criterion. The standard deviation of node q is calculated as:

$$\sigma(D_q) = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \bar{y}_q)^2} \quad (10)$$

where σ denotes the standard deviation and $\bar{y}_q$ denotes the mean target value of the samples located at node q .

Splitting Criterion:

Suppose that a candidate splits partitions node D_q into K child subsets denoted by $D_{q1}, D_{q2}, \dots, D_{qK}$. The expected standard deviation after splitting is computed as:

$$\sigma_{split}(D_q) = \sum_{k=1}^K \frac{|D_{qk}|}{|D_q|} \sigma(D_{qk}) \quad (11)$$

Accordingly, the SDR value used by M5Boost is expressed as:

$$\text{SDR} = \sigma(D_q) - \sigma_{split}(D_q) \quad (12)$$

In M5Boost, the splitting attribute and threshold that yield the maximum SDR are selected for recursive partitioning. M5Boost aims to maximize this expression to ensure the residuals in the child nodes are more homogeneous. The recursive partitioning process continues until predefined stopping criteria are satisfied, such as insufficient variance reduction or minimum sample constraints.

Leaf Nodes:

Unlike conventional regression trees that assign constant prediction values to leaf nodes, the M5Boost method constructs multivariate linear regression models at leaf nodes. For a leaf node l , the local regression model is defined as:

$$g_l(x) = \beta_{l0} + \sum_{j=1}^d \beta_{lj} x_j \quad (13)$$

where β_{10} denotes the intercept coefficient, β_{ij} represents the regression coefficient associated with the j -th feature, and the variable d represents the dimensionality of the input feature space.

M5Boost ensures model robustness through two critical post-processing steps: pruning and smoothing.

Pruning:

To improve generalization capability and reduce overfitting, the proposed approach employs pruning operations during tree construction. In the pruning stage, subtree structures are recursively evaluated and replaced by regression models whenever the estimated prediction error of the subtree exceeds the error generated by the corresponding model. The estimated pruning error (E) used in M5Boost is defined as:

$$E = \frac{n_q + p}{n_q - p} \times \frac{1}{n_q} \sum_{i=1}^{n_q} |r_i^{(t)} - \hat{f}_t(x_i)| \quad (14)$$

where n_q represents the number of training instances associated with node q , p denotes the number of parameters (coefficients) in the corresponding regression model, $r_i^{(t)}$ represents the residual value of the i -th sample at boosting iteration t , and $\hat{f}_t(x_i)$ denotes the prediction generated by the model for input sample x_i . This pruning mechanism simplifies the tree structure while maintaining predictive performance and reducing unnecessary model complexity.

Smoothing:

M5Boost also employs smoothing operations between parent and child nodes in order to reduce abrupt prediction discontinuities between neighboring leaf nodes. The smoothing process combines the predictions generated by a child node and its parent node, thereby improving prediction stability and robustness throughout the tree structure. The smoothing equation used in M5Boost is expressed as:

$$\hat{y}_{\text{smooth}} = \frac{n_{\text{sub}} \hat{y}_{\text{child}} + w \hat{y}_{\text{parent}}}{n_{\text{sub}} + w} \quad (15)$$

where $\hat{y}_{\text{smooth}}$ denotes the final smoothed prediction value, $\hat{y}_{\text{child}}$ represents the prediction generated by the child node, $\hat{y}_{\text{parent}}$ refers to the prediction generated by the parent node, n_{sub} is the number of training samples associated with the child node, and w denotes the smoothing constant controlling the contribution of the parent prediction.

Prediction:

During the prediction stage, an input sample is first routed through the hierarchical decision structure of the M5 Model Tree according to the learned splitting conditions. After reaching the corresponding leaf node, the prediction value is obtained from the associated local regression equation.

By integrating recursive partitioning, local regression modeling, tree-pruning operations, smoothing mechanisms, and additive residual correction, M5Boost provides an effective and interpretable solution for estimation problems characterized by nonlinear and complex relationships.

Algorithm 1 presents pseudocode illustrating the iterative execution flow of M5Boost and detailing how the final predictive model is systematically constructed. First, the algorithm is initialized by computing the mean target value of the training dataset to establish a baseline $F_0(x)$. At each boosting iteration, residual errors are computed by subtracting the current ensemble predictions from the actual target values and the subset $D^{(t)}$ is defined. Subsequently, a new M5 Model Tree is trained using these residual values in order to fit the remaining prediction errors that could not be captured by the previous ensemble structure. During the tree construction stage, candidate splitting attributes are recursively evaluated using the standard deviation reduction (SDR) criterion, and the attribute

providing the maximum SDR value is selected to partition the data into more homogeneous subsets. Accordingly, the current node D_q is recursively partitioned into left (D_{qL}) and right (D_{qR}) child subsets, and the node collection is updated. This recursive partitioning process continues until the predefined stopping criteria are satisfied, resulting in a hierarchical tree topology $\mathcal{T}_t$. After the tree topology has been established, local multivariate linear regression models are generated for all leaf nodes $q \in \mathcal{L}$. The algorithm maps the tree topology $\mathcal{T}_t$ and its corresponding leaf equations into a base tree structure, h_t . To improve predictive robustness and generalization capability, pruning and smoothing are subsequently applied to the constructed model tree as post-processing operations. After all these phases, the newly generated M5 Model Tree is incorporated into the ensemble model through the additive regression updating mechanism. By applying the shrinkage parameter v to each tree, M5Boost gradually refines the ensemble's accuracy while maintaining generalization. This iterative learning process continues until the predefined number of boosting iterations (T) is reached. As a result, M5Boost constructs a strong ensemble regression model composed of multiple complementary model trees.

Algorithm 1: M5Boost

```

1.   $D = \{(x_i, y_i)\}_{i=1}^n, T, v$ 
2.   $F_0(x) = 1/n \sum_{i=1}^n y_i$ 
3.   $t \leftarrow 1$ 
4.  while  $t \leq T$  do
5.       $r_i^{(t)} \leftarrow y_i - F_{t-1}(x_i) \quad \forall i \in \{1, \dots, n\}$ 
6.       $D^{(t)} = \{(x_i, r_i^{(t)})\}_{i=1}^n$ 
7.      while stopping criteria not satisfied do
8.           $SDR_j = \sigma(D_q) - \sigma_{split}(D_q)$ 
9.           $j^* \leftarrow \arg \max_j SDR_j$ 
10.          $D_{qL}, D_{qR} \leftarrow Split(D_q, j^*)$ 
11.          $D_q \leftarrow \{D_{qL}, D_{qR}\}$ 
12.     end while
13.      $g_q(x) \leftarrow \beta_{q0} + \sum_{k=1}^d \beta_{qk} x_k \quad \forall q \in \mathcal{L}$ 
14.      $h_t \leftarrow (\mathcal{T}_t, \{g_q(x)\}_{q \in \mathcal{L}})$ 
15.      $h_t \leftarrow Prune(h_t)$ 
16.      $h_t \leftarrow Smooth(h_t)$ 
17.      $F_t(x) = F_{t-1}(x) + v h_t(x)$ 
18.      $t \leftarrow t + 1$ 
19. end while
20. return  $F_T(x)$ 

```

4. Experimental Studies

4.1. Dataset Description

This study used a publicly available dataset obtained from the SpritMonitor (<https://www.spritmonitor.de/>) (accessed on 25 April 2026) platform and stored in a GitHub version 2.55.0 (<https://github.com/armiro/crawlers/tree/master/SpritMonitor-Crawler>) (accessed on 25 April 2026) repository [44]. SpritMonitor is a collaborative crowdsourced web platform where vehicle owners voluntarily share detailed information regarding their charging events and driving activities. The platform contains extensive

real-world operational data for electric vehicles, including trip distance, average speed, driving conditions, and additional vehicle-related information. Unlike standardized laboratory data, this platform stores dynamic field data from thousands of electric vehicle users, encompassing diverse driving styles, environmental conditions, and inherent variability of EV usage. Due to its extensive scale and continuously updated structure, it serves as a high-potential data source for developing and evaluating advanced range estimation models in the EV domain.

The original dataset [44] was collected using web crawlers from the website and included records over the 2014–2019 period. In this study, the dataset was further systematically extended with newly collected records from the same platform. As a result, the final dataset spans the period from 18 December 2014 to 7 April 2026, corresponding to a total duration of more than 11 years.

The experiments were conducted using data belonging to the Volkswagen e-Golf, which was selected due to its large number of available records and balanced distribution across various users. The final dataset consists of records collected from 10 distinct EVs owned by different drivers. In total, the dataset contains 8501 trip records that reflect diverse real-world battery-aware driving scenarios. Table 2 summarizes the specifications of the vehicles incorporated in the dataset, including production year, power rating, gearbox type, number of records, total traveled distance, and data collection period.

Table 2. Characteristics of the vehicles used in the proposed range estimation framework.

Vehicle ID	Driver ID	Year	Power (kW)	Gearbox	Number of Records	Total Trip Distance	Start Date	End Date
01	786327	2016	85	automatic	2433	150182	23.12.2016	05.04.2026
02	1503266	2019	100	automatic	312	36240	18.12.2023	07.04.2026
03	1281734	2019	100	automatic	288	42303	02.09.2021	25.03.2026
04	1547073	2020	100	automatic	142	19788	09.05.2024	01.04.2026
05	1168655	2020	100	continuously variable	268	35918	01.08.2020	25.07.2025
06	1131375	2020	100	continuously variable	411	51545	30.11.2020	24.11.2023
07	979887	2018	100	automatic	402	36767	17.11.2018	18.02.2022
08	925855	2018	100	continuously variable	157	30863	25.02.2018	03.12.2019
09	861231	2015	85	automatic	1394	99134	18.08.2017	12.10.2019
10	679341	2014	85	continuously variable	2694	85218	18.12.2014	02.06.2019

Table 3 describes the features integrated into the study. A total of 14 input features and 1 target variable were used in the proposed ML-based range estimation framework. The input space is categorized into five primary domains: (i) Vehicle Specifications: It captures intrinsic hardware attributes such as production year, maximum power output, and the relative age of the EV. (ii) Usage Info: It incorporates ordinal-encoded odometer readings and nominal tire types to account for mechanical wear and rolling resistance. (iii) Energy Metrics: It contains continuous variables, including trip-specific energy consumption (kWh) and average speed (km/h), which serve as primary indicators of energy intensity. (iv) Environmental Factors: It accounts for external influences through seasonal temperature categories and binary indicators for auxiliary energy loads (A/C and Park Heating). (v) Driving Profiles: It represents the spatial context of the trip (City, Motorway, Country Roads) and the subjective driving style, ranging from normal to fast.

Table 3. Description of the features utilized in the proposed range prediction model.

Category	Feature	Description	Unit	Data Type
Vehicle Specifications	Year	The production year of the EV.	year	Integer
	Power	Maximum power output under normal cond.	kW	Integer
	Vehicle Age	The age of the EV.	years	Integer
Usage Info	Odometer	Categorized the total distance traveled since initial use, i.e., low, medium, and high.	0: $x \leq 40,000$ 1: $x \leq 90,000$ 2: $x > 90,000$	Categorical (Ordinal)
	Tire Type	The type of tires equipped on the vehicle.	0: Winter 1: All-year 2: Summer	Categorical (Nominal)
Energy Metrics	Quantity	Energy consumed by the EV during the trip.	kWh	Float
	Average Speed	Average speed at which the vehicle is driven.	km/h	Integer
Environmental Factors	Ambient Temperature Category	Seasonal temperature category derived from the trip month.	0: Cold 1: Mild 2: Hot	Categorical (Ordinal)
	A/C	Use of air conditioning.	0: Off 1: On	Binary
	Park Heating	Use of park heating.	0: Off 1: On	Binary
Driving Profiles	City	EV drives in the city or not.	0: No 1: Yes	Binary
	Motorway	EV drives on the motorway or not.	0: No 1: Yes	Binary
	Country Roads	EV drives on the country roads or not.	0: No 1: Yes	Binary
	Driving Style	The manner in which the vehicle is driven.	0: Normal 1: Moderate 2: Fast	Categorical (Ordinal)
Target Variable	Trip Distance	The range the vehicle traveled on a single charge.	km	Integer

Table 4 presents representative sample records within the dataset used in this study. These sample instances demonstrate the diversity of the dataset in terms of vehicle production year, tire type, driving style, ambient temperature conditions, and average driving speed. The presented records also highlight the variability in the traveled EV trip range under different operational conditions, which is essential for developing a robust and generalized ML-based range estimation model.

Table 4. Sample records from the EV usage data used for building the proposed model.

Year	Power	Vehicle Age	Odometer	Tire Type	Quantity	Avg. Speed	Ambient Temp. Category	A/C	Park Heating	City	Motor Way	Country Roads	Driving Style	Trip Distance
2016	85	11	2	All-year	15.23	62	Mild	1	0	1	1	1	Normal	112
2015	85	5	2	Summer	15.09	71	Mild	0	0	0	1	1	Moderate	104
2015	85	4	1	Winter	4.09	92	Mild	0	0	0	1	0	Fast	87
2014	85	2	0	Winter	3.54	33	Mild	0	1	1	0	0	Normal	30
2019	100	6	1	Winter	27.26	39	Cold	0	1	1	0	1	Normal	164
2018	100	4	0	Summer	23.27	45	Hot	1	0	1	0	1	Normal	155
2015	85	4	0	Winter	13	69	Cold	0	1	0	0	1	Moderate	88
2018	100	5	1	Winter	7.9	61	Cold	0	1	0	1	0	Fast	45
2020	100	6	0	All-year	28.45	40	Mild	1	0	1	1	1	Normal	183
2020	100	5	1	Summer	22.05	57	Hot	1	0	1	1	1	Normal	126

4.2. Feature Engineering

Feature engineering plays a critical role in ML-based range estimation frameworks since the driving range of EVs is impacted by numerous operational, vehicle-related, and environmental factors. In this study, several feature engineering operations were performed to improve data quality and therefore enhance prediction performance. The overall process consisted of feature construction, feature extraction, discretization, feature encoding, feature removal, handling missing values, and noise/outlier removal.

Step 1—Feature Construction

In the first stage, a new feature named Vehicle Age was constructed by subtracting the vehicle production year from the charging year. The age of an EV can significantly affect its driving range due to battery degradation over time. For instance, lithium-ion batteries gradually lose their energy storage capacity as the number of charge–discharge cycles increases. Older vehicles generally exhibit reduced battery efficiency and lower usable battery capacity compared to newer vehicles. As a result, the same amount of energy yields a shorter driving distance in older EVs.

Step 2—Feature Extraction

In the second stage, the month information was extracted from the date value and then used as an indicator of ambient temperature. Ambient temperature has a direct impact on EV driving range. In cold weather conditions, the kinetics of electrochemical reactions within the battery cells inherently slow down, leading to reduced battery capacity and increased energy consumption. Moreover, cabin heating systems consume additional electrical energy, which further decreases the available driving range. In contrast, hot weather conditions may also negatively affect the net vehicular range due to increased air-conditioning usage and active thermal regulation requirements.

Step 3—Discretization

While discretization may reduce the granularity of numerical information and potentially lead to information loss, directly feeding highly fluctuating real-world variables into tree-based architectures can amplify noise effects and reduce the stability of decision boundaries due to data variability. To evaluate whether discretization is beneficial for the considered EV range estimation problem, a comparative analysis was conducted using both the raw continuous representations and the discretized versions of the ambient temperature and odometer features. The empirical evaluation revealed that the discretized representation achieved superior predictive performance (MAE = 3.8411, RMSE = 7.8485) compared with the continuous representation (MAE = 3.9145, RMSE = 8.0680). This observation confirmed that the discretization process effectively captured meaningful insights while filtering out non-informative micro-variations, thereby enhancing the predictive accuracy of the global model.

Since the adopted real-world dataset does not provide recorded ambient temperature measurements or geographic information (e.g., city or region) for individual trips, the month value was used as a proxy for seasonal weather conditions. Since feeding continuous month/temperature approximations into the model actually degraded the prediction performance as described above, the months December, January, and February were categorized as Cold, the months from June to August were assigned to Hot, and the remaining months were labeled as Mild. Furthermore, the numerical Odometer feature was converted into three-tier categorical usage levels based on the cumulative distance traveled by the vehicle. Vehicles with odometer values not exceeding 40,000 km were categorized as Low usage level and denoted by 0, vehicles with mileage from 40,000 km to 90,000 km were assigned to Medium usage level and denoted by 1, and finally, vehicles exhibiting odometer readings greater than 90,000 km were classified as High usage level and denoted

by 2. The total traveled distance of a vehicle may influence EV range due to long-term battery degradation and mechanical aging effects. Vehicles with higher odometer values are generally subjected to a greater number of charging cycles, which may consequently lead to reduced battery performance and operational efficiency. Instead of using raw odometer values directly, discretization was applied to simplify the interpretation of vehicle wear levels, reduce the influence of extreme numerical values, and improve the generalization of the M5Boost model within the ML framework.

Step 4—Feature Encoding

During this stage, several categorical attributes were encoded into numerical representations to ensure compatibility with ML algorithms. The Driving Style feature was mapped to ordinal values: Normal = 0, Moderate = 1, and Fast = 2. Driving behavior is one of the influential factors affecting EV range. Fast and aggressive driving generally leads to higher acceleration rates, increased power demand, and greater energy consumption. Consequently, the available driving range decreases significantly under high-speed driving conditions. Similarly, the Tire Type attribute was encoded as follows: Winter tires were represented by 0, All-year tires by 1, and Summer tires by 2. Tire characteristics directly influence rolling resistance and thereby dictate vehicle energy efficiency. Winter tires generally exhibit higher rolling resistance due to the inherent nature of their softer rubber compounds and specialized tread structures, which may increase energy consumption and reduce the achievable driving range. Summer tires typically provide reduced rolling resistance and enhanced efficiency under suitable temperature conditions, while all-year tires offer a balanced intermediate performance.

Step 5—Feature Removal

In this preprocessing stage, several features were removed from the data since they did not provide statistically meaningful discriminative information for machine learning. Since the dataset only contains Volkswagen e-Golf vehicles, the attributes manufacturer, model, and version consisted of a single unique value throughout the entire dataset and were therefore discarded. Similarly, the fuel-type attribute was eliminated because all vehicles in the dataset were solely EVs. Furthermore, the consumption feature was excluded since it is computationally derived from energy quantity and trip distance values. Including this variable could introduce redundancy and potentially lead to data leakage during M5Boost model training.

Step 6—Handling Missing Values

In this stage, a small number of records with missing values were removed from the dataset since they could negatively affect the M5Boost model training and evaluation processes. They were particularly observed in the Average Speed attribute and the target attribute, Trip Distance.

Step 7—Noise and Outlier Removal

In the final preprocessing stage, noise and outlier samples were eliminated using the Isolation Forest algorithm. Isolation Forest is an effective anomaly detection method that can successfully identify abnormal observations based on random partitioning principles without requiring assumptions about data distribution. Outlier removal is particularly important in EV range estimation problems since sensor errors, incorrect user entries, or unrealistic trip records may negatively affect the model learning and reduce prediction accuracy.

Table 5 presents the descriptive statistics for the numerical features in the cleaned dataset after completing the feature engineering operations. The table summarizes the minimum, maximum, mean, and standard deviation values of the attributes used in the proposed EV range estimation framework. According to the statistics, the Quantity fea-

ture ranges from 0.01 kWh to 34.97 kWh, with an average value of 8.71 kWh, indicating considerable fluctuations in energy consumption patterns among different trips. Similarly, the Average Speed attribute reflects the diversity of driving environments and driver behaviors contained in the dataset. The Vehicle Age feature spans from 1 to 11 years, demonstrating that the dataset includes both relatively new and aged EVs, which is essential for analyzing battery aging effects on driving range. Furthermore, the Trip Distance target variable, with a mean value of 58.99 km, highlights the diverse nature of real-world EV driving conditions and trip characteristics.

Table 5. The descriptive statistics of numerical features in the cleaned dataset after feature engineering.

Feature	Min	Max	Mean	Std
Quantity	0.01	34.97	8.71	7.06
Average Speed	6.00	136.00	46.70	14.39
Vehicle Age	1.00	11.00	4.76	2.32
Trip Distance (Range)	1.00	280.00	58.99	46.52

Figure 2 presents the distribution of driving range values for the entire dataset, with a probability density histogram superimposed on a Kernel Density Estimation (KDE) curve. As can be observed, the dataset is naturally skewed toward shorter driving ranges (below 120 km), with progressively fewer samples for long-distance trips (up to 280 km). This distribution accurately reflects the empirical nature of real-world electric vehicle utilization, where daily regional routines and urban commutes dominate the dataset, while long-range inter-city trips represent a smaller fraction of operations.

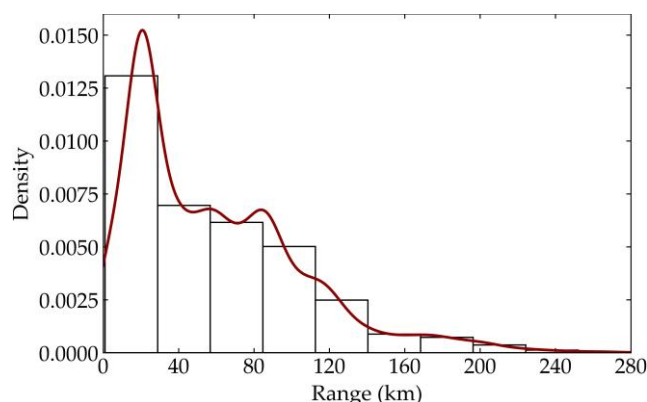


Figure 2. Distribution of driving range values for the entire dataset. The red line shows the KDE distribution of the histogram.

4.3. Feature Selection

Feature selection is a critical phase in the machine learning pipeline, as it directly influences the predictive performance and interpretability of models. This process systematically examines the underlying statistical relationships between the input feature space and the target variable. By identifying and selecting the most informative attributes, it effectively achieves significant computational complexity reduction, model generalization capability improvement, and the elimination of redundant or irrelevant information. In this study, Mutual Information (MI) was employed as the feature selection technique. Formally, MI quantifies the amount of information obtained about a variable through observing the other. Higher MI values indicate stronger dependency and greater predictive relevance between the feature and the target variable. It is calculated as:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (16)$$

where X denotes an input feature and Y refers to the continuous target variable, $MI(X, Y)$ is mutual information between them. Specifically, the symbol $p(x, y)$ corresponds to the joint probability of these values (X and Y). Here, the notations $p(x)$ and $p(y)$ are the individual marginal probabilities of X and Y , respectively. According to the foundational theorem: If X and Y maintain absolute independence then $p(x, y) = p(x)p(y)$, which simplifies the logarithmic component to $\log(1) = 0$ and $MI = 0$. Conversely, whenever X provides information about Y , the MI score strictly exceeds zero.

Figure 3 shows the calculated MI scores for all candidate features in descending order. As revealed by the empirical data, Quantity exhibited the highest informational gain, indicating its role as the most influential predictor of EV range estimation. This is followed by temporal and dynamic factors, namely Year and Average Speed. Furthermore, vehicle age, motorway usage, and country road driving also demonstrated relatively high MI values, suggesting their significant contributions to range estimation. Overall, the results confirm that range estimation is influenced by a combination of vehicle characteristics, energy-related variables, driving conditions, and operational factors.

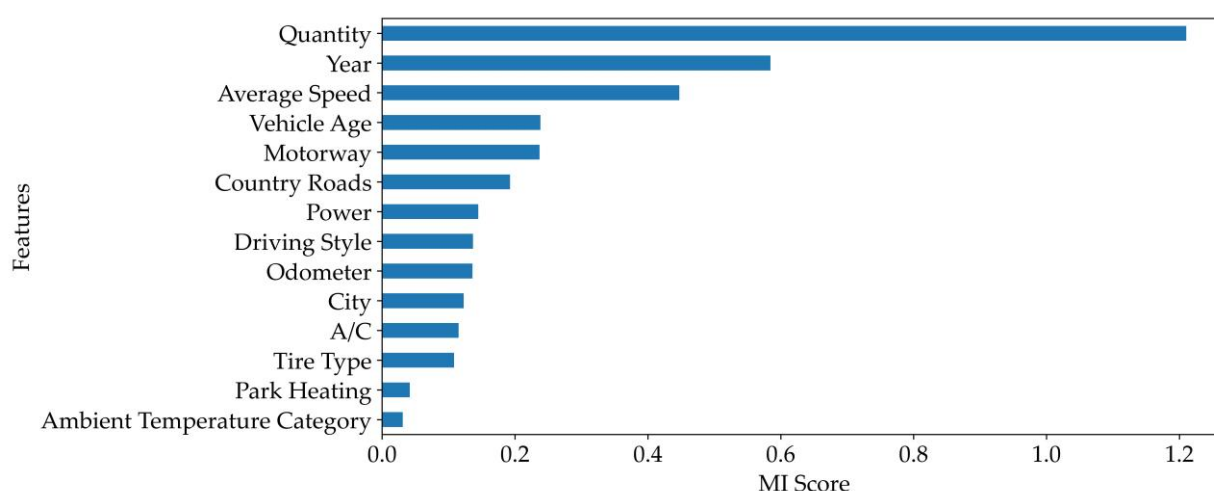


Figure 3. Feature importance scores computed between candidate input factors and the target driving range variable.

To determine the optimal feature subset, an iterative analysis was conducted using the MI ranking results. Starting from the lowest-scoring attributes, features were systematically removed one by one, and the model was retrained after each elimination step. Table 6 presents the performance metrics obtained during this sequential elimination procedure. The findings demonstrated that the best predictive performance was achieved when the top eight features were used. These results indicate that a compact subset of highly informative features can provide better generalization capability while reducing model complexity.

Table 6. Performance analysis of the model across varying numbers of selected features based on sequential elimination.

Number of Features	MAE	RMSE	MAPE
4	4.6894	8.6560	35.4712
5	4.4096	8.4207	36.3835
6	3.9713	8.0547	35.5126
7	3.9713	8.0547	35.5126

8	3.8411	7.8485	34.0127
9	4.3474	7.9381	38.0126
10	4.1899	7.9308	38.0747
11	4.3270	8.0901	36.3205
12	4.2865	8.4195	34.5963
13	4.0184	8.3122	34.5576

4.4. Implementation Details

M5Boost was implemented in the Java programming language by leveraging the extensive algorithmic capabilities of the Weka version 3.8.6 (University of Waikato, Hamilton, New Zealand) machine learning library [45] within the Eclipse IDE version 4.26.0 (The Eclipse Foundation, Brussels, Belgium) development environment. The implementation process included the integration of the M5 Model Tree algorithm [46] with the additive regression meta-learning strategy in order to construct a boosting-based EV range estimation framework.

To comprehensively validate the effectiveness of M5Boost, performance comparisons were carried out against 11 prominent ML regression approaches. The compared methods include K-Nearest Neighbors (KNN), KStar, Locally Weighted Learning (LWL), Linear Regression, Neural Network, Support Vector Machine (SVM), Random Forest, ElasticNet, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost).

Prior to the training phase, feature selection was applied to identify the most informative attributes affecting EV range estimation performance. The details of the feature selection procedure are presented in Section 4.3. For evaluation, the experimental workflow followed a hold-out validation strategy. Specifically, 90% of the dataset was allocated for model training, while the remaining 10% was reserved strictly for final blind performance testing. This experimental configuration enabled the proposed M5Boost framework and the competing ML models to be assessed under identical data distribution conditions. Furthermore, as described in Section 4.6, hyperparameter tuning processes were performed for all ML models to ensure fair and reliable performance comparisons. Specifically, the entire training pipeline was repeatedly executed from scratch using different hyperparameter configurations through independent runs, and the predictive performance of each configuration was assessed using previously unseen test samples. In each run of the grid-search methodology, the model was trained exclusively on the training partition, and the test partition was never used during model fitting. Therefore, no information from the test data was incorporated into the training process. Moreover, a specialized sensitivity analysis was integrated into the study to evaluate the impact of human-centric variables on the model's performance, which is detailed in Section 4.7.

For evaluation, the experimental workflow followed a hold-out validation strategy. Specifically, 90% of the dataset was allocated for model training, while the remaining 10% was reserved for final performance testing. This experimental configuration enabled M5Boost and the competing ML models to be assessed under identical data distribution conditions.

To evaluate the predictive accuracy of the proposed M5Boost framework for EV range estimation, three widely used regression assessment measures, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), were used. In EV range estimation problems, these measures are specifically tailored to analyze the deviation between the actual and predicted driving ranges obtained from the machine learning model.

Mean Absolute Error (MAE): It provides a direct measure of how many kilometers, on average, the M5Boost model's predictions deviate from the true range. For the range forecasting task in this study, it is formulated as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |R_{\text{actual},i} - R_{\text{estimated},i}| \quad (17)$$

where n is the number of trips in the test set, $R_{\text{actual},i}$ denotes the true range recorded for the i -th trip, and $R_{\text{estimated},i}$ is the corresponding distance projected by the model. A lower MAE value indicates a more consistent and reliable prediction for the EV user.

Root Mean Squared Error (RMSE): Since the RMSE metric penalizes larger discrepancies in range estimation, it provides additional insight related to the stability and robustness of models under challenging driving conditions. It evaluates the square root of the average squared difference between the actual EV range and the predicted EV range. The RMSE formulation employed for range prediction is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (R_{\text{actual},i} - R_{\text{estimated},i})^2} \quad (18)$$

Lower RMSE values indicate that the proposed M5Boost model produces more reliable and consistent EV range predictions across different operational scenarios.

Mean Absolute Percentage Error (MAPE): It is adopted as a core statistical metric by computing the normalized percentage deviation between the projected distances and the actual trip lengths. The MAPE is mathematically formulated as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{R_{\text{actual},i} - R_{\text{estimated},i}}{R_{\text{actual},i}} \right| \times 100 \quad (19)$$

4.5. Results

Table 7 presents the comparative performance of M5Boost against 11 popular ML models on MAE, RMSE, and MAPE. According to the experimental outcomes, the proposed M5Boost model outperformed all other models, producing the lowest error rates with a MAE of 3.8411. These results indicate that M5Boost achieved more accurate range estimation performance compared to its alternative approaches. Among the competing ensemble learning algorithms, XGBoost and LightGBM also demonstrated strong performance; however, their prediction errors remained higher than those of the proposed M5Boost model. In contrast, traditional regression methods such as Linear Regression exhibited significantly larger error rates, indicating limited capability for modeling complex EV range estimation patterns. The superior performance of M5Boost can be attributed to the combination of additive residual learning, localized regression modeling, pruning operations, and smoothing mechanisms provided for the tree-based structure.

Table 7. Performance comparison of the predictive models for electric vehicle range estimation.

Method	MAE	RMSE	MAPE
K-Nearest Neighbors	10.9389	18.8137	30.1278
KStar	11.6011	20.3963	32.8884
Locally Weighted Learning	11.4887	20.8652	31.0043
Linear Regression	15.0344	23.4866	47.4332
ElasticNet	7.5138	11.8895	47.5482
Neural Network	12.0943	19.1505	47.0267
Support Vector Machine	14.8012	23.6292	44.3051
Random Forest	11.5279	19.1814	37.5111

LightGBM	4.1276	8.6247	34.5897
XGBoost	4.2748	8.1986	40.7584
CatBoost	5.5828	8.4609	40.8642
M5Boost (proposed)	3.8411	7.8485	34.0127

The MAPE results provide insights into the relative prediction accuracy of the developed models across trips with different driving ranges. While relatively high MAPE values may mainly be attributed to short-trip samples with small actual range values, lower MAPE values indicate that the predicted driving ranges are closer to the actual ranges in percentage terms.

To statistically validate the predictive precision of the proposed M5Boost approach, two different non-parametric tests (Friedman Aligned Ranks Test and Quade Test) were performed at a significance level of 0.05 in a one-vs-all configuration. The null hypothesis (H_0) states that the performances of two or more methods are statistically equivalent. The Friedman Aligned Ranks Test obtained a p -value of 0.04823, which is lower than the significance threshold of 0.05, leading to the rejection of the null-hypothesis (H_0). Similarly, the Quade Test resulted in a highly significant p -value of 0.00242, further confirming the rejection of H_0 . Therefore, the null-hypothesis was rejected in both statistical tests. These findings demonstrate that the performance differences between M5Boost and the competing ML methods are statistically significant and not due to random chance drift. Consequently, the experimental results confirm the robustness of M5Boost for EV range estimation problems.

Figure 4 presents the scatter plot of actual versus estimated values generated by M5Boost for the EV range. Ideally, accurate predictions are expected to be distributed close to the dashed diagonal line ($x = y$), indicating a high degree of agreement between actual and projected range values. As illustrated in the figure, the estimation points are generally concentrated near the reference line, demonstrating the high predictive accuracy of M5Boost across both short-range trips and longer-distance trips. The overall distribution confirms that M5Boost successfully captures the underlying patterns of EV driving range characteristics under varying operational conditions.

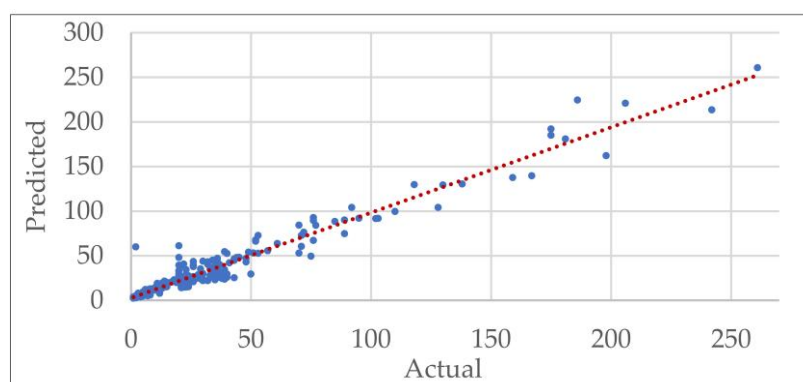


Figure 4. Correlation analysis between actual and predicted range values. The red dashed line shows the reference line, the blue dots show actual and predicted data.

Figure 5 illustrates the distribution of residuals, representing the differences between actual and predicted driving range values. The histogram shows that the residuals are primarily centered around zero, indicating that M5Boost produced predictions with relatively small estimation errors. Furthermore, the residual distribution exhibits an approximately symmetric structure, following a nearly normal distribution as indicated by the overlaid density curve. The lack of significant skewness in the distribution confirms that the model does not consistently overestimate or underestimate the range, demonstrating

unbiased predictive behavior across different EV operating conditions and trip characteristics. The relatively narrow spread of residual values is highly acceptable for real-world EV range estimation.

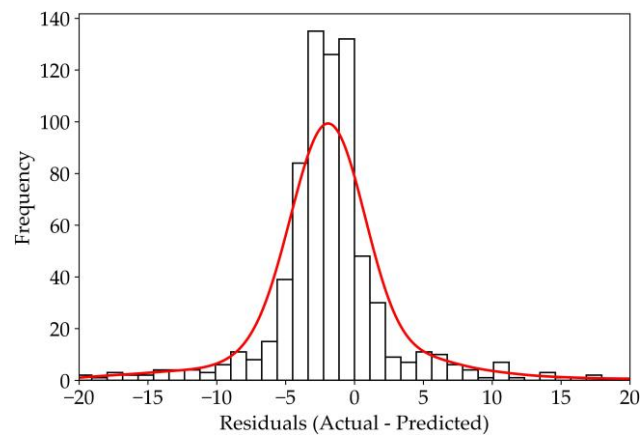
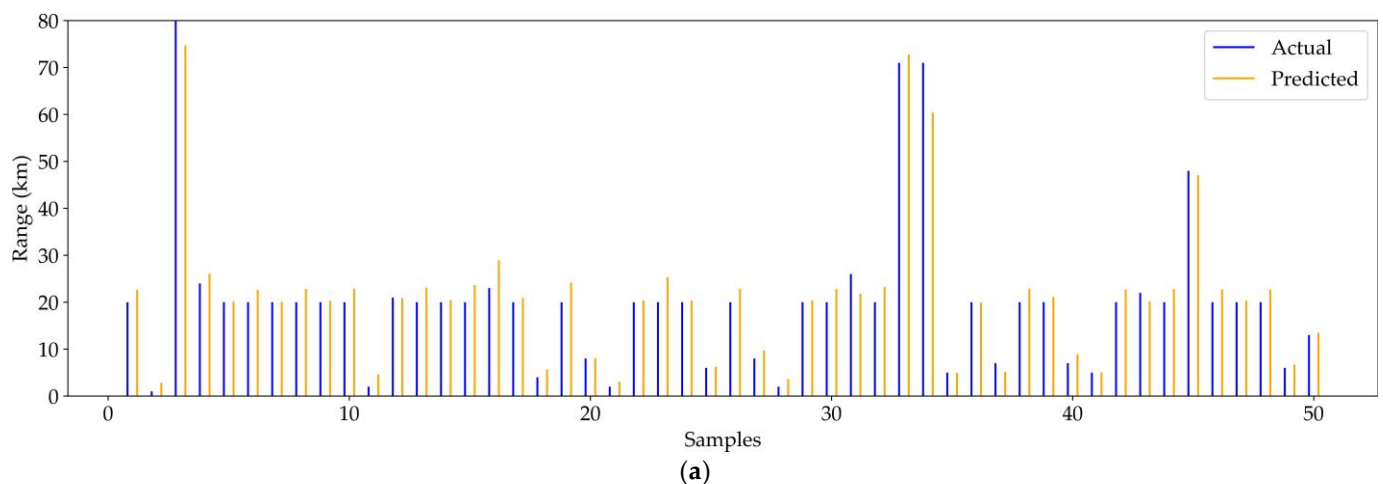


Figure 5. Distribution of prediction residuals. The red line shows the KDE distribution of the histogram.

To illustrate the prediction behavior of the model across different scales, Figure 6 provides a two-part stem plot visualization comparing the actual and corresponding estimated range values generated by M5Boost. Specifically, Figure 6a presents a clear, zoomed-in look at a subset of 50 trips to prevent visual clutter, while Figure 6b displays the entirety of the test samples across the full distribution of trip lengths to provide a comprehensive overview. As observed, the blue lines (actual) and orange lines (predicted) show remarkable synchronization across test samples. The predicted range values closely follow the actual range values, demonstrating the capability of M5Boost to accurately estimate EV driving range under different operating conditions. This sample-based comparison confirms that the proposed framework is robust enough to handle the volatility inherent in real-world driving data.



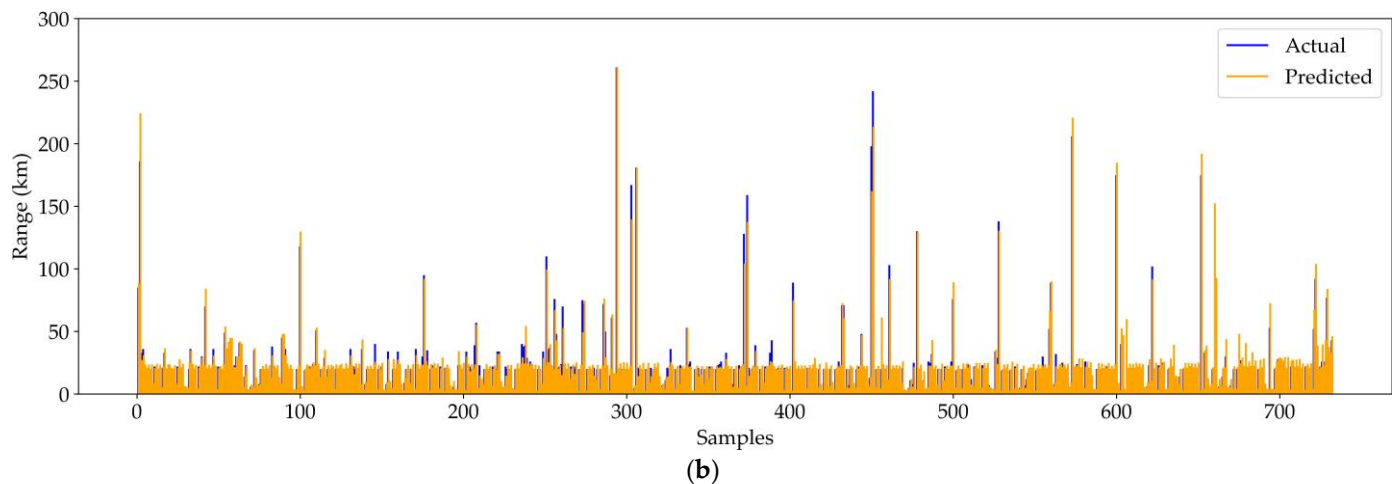


Figure 6. Comparison of actual and predicted range values via stem plot representation: (a) a representative subset of 50 test samples; (b) the comprehensive distribution across all test samples.

4.6. Hyperparameter Analysis

Choosing appropriate hyperparameters is one of the most important stages for building a robust and high-performing predictive machine learning model. It directly regulates the learning behavior, structural complexity, convergence characteristics, and overfitting tendency of the model and prevents generalization errors. Since range estimation involves highly heterogeneous and nonlinear driving patterns, selecting suitable hyperparameter configurations becomes particularly important for achieving satisfactory prediction performance under varying EV operational conditions. Therefore, a grid search was conducted for all evaluated ML approaches in this study, ensuring a fair and systematic comparison under optimized experimental settings.

For each ML method used in the experimental analysis, Table 8 presents the alternative hyperparameter configurations and the corresponding optimal values that yielded the peak estimation accuracy on the dataset. For instance-based learning methods, neighborhood-related parameters such as the number of nearest neighbors were analyzed to find a balance between localized patterns and noise smoothing. For linear regression-based approaches, regularization-related parameters were investigated to control model complexity and prevent overfitting. In the neural network configuration, learning rate and hidden layer architecture were optimized to improve nonlinear learning capability. For ensemble learning approaches, maximum tree depth, iteration count, and learning rate parameters were searched due to their significant impact on prediction capability.

Table 8. Hyperparameter configurations for each ML algorithm.

Regressor Category	Method	Hyperparameter	Search Space	Optimal Value
Instance-Based	K-Nearest Neighbors	n_neighbors	{3, 5, 7}	7
	KStar	global blend	{10, 20, 30, 40}	10
	Locally Weighted Learning	n_neighbors	{3, 5, 7}	7
Linear Models	Linear Regression	ridge	$\{1.0 \times 10^{-8}, 1.0 \times 10^{-4}, 0.01, 0.1\}$	0.01
		alpha	{0.01, 0.1}	0.1
	ElasticNet	l1_ratio	{0.2, 0.5, 0.8}	0.2
Neural Network	Multi-Layer Perceptron	learning_rate	{0.01, 0.1}	0.1
		hidden_layers	{{3, 3}, {5, 5}, {7, 7}}	{7, 7}
Support Vector Machines	Support Vector Regression	complexity	{1, 10, 20}	10

Bagging Ensemble	Random Forest	maxDepth	{3, 5, 7}	7
		numIterations	{10, 100, 200, 300}	100
Boosting Ensemble	LightGBM	learning_rate	{0.01, 0.1}	0.1
		max_depth	{3, 5, 7}	7
		n_estimators	{10, 100, 200, 300}	100
	XGBoost	learning_rate	{0.01, 0.1}	0.1
		max_depth	{3, 5, 7}	7
		n_estimators	{10, 100, 200, 300}	100
	CatBoost	learning_rate	{0.01, 0.1}	0.1
		max_depth	{3, 5, 7}	7
		n_estimators	{10, 100, 200, 300}	300
	M5Boost	shrinkage	{0.01, 0.1}	0.1
		minNumInstances	{5, 15, 25, 35, 45}	35
		numIterations	{10, 20, 30, 40, 50}	50

The hyperparameter tuning for M5Boost was carefully executed to harmonize its additive learning structure with the underlying multivariate regression models. The optimization space for M5Boost explicitly targeted the shrinkage factor, the minimum number of instances per leaf node (minNumInstances), and the total number of boosting iterations (numIterations). As shown in Table 8, M5Boost achieved the highest accuracy with a shrinkage parameter of 0.1, a minimum node capacity of 35, and 50 boosting iterations. This hyperparameter configuration enabled M5Boost to construct sufficiently generalized model trees while iteratively minimizing residual estimation errors without excessive overfitting.

4.7. Sensitivity Analysis

Sensitivity analysis constitutes a fundamental procedure in evaluating the structural stability and behavior of a machine learning model under different parameter settings. In this study, an analysis was conducted to investigate the operational impact of three critical hyperparameters governing M5Boost: the number of iterations (T), the shrinkage factor (v), and the minimum number of instances per leaf node (minNumInstances). The objective of this analysis is to identify appropriate parameter values that achieve an effective trade-off between prediction accuracy, model complexity, and generalization capability.

Figure 7 illustrates the convergence behavior of the model against a progressively increasing number of boosting iterations. A sharp reduction in MAE was visible in the initial phases, demonstrating the effectiveness of the additive residual learning mechanism. Upon reaching 50 iterations, the error margin attained its minimum value and stabilized, indicating that the algorithm had already captured the majority of the learnable information contained in the dataset. Therefore, additional iterations did not provide any further predictive gains and only increased computational cost. Based on these findings, the optimal number of iterations was selected as 50.

In M5Boost, the shrinkage parameter acts as a regularization component. Smaller shrinkage values lead to slower learning and require more iterations, whereas larger values accelerate the learning process. Figure 8 presents the relationship between the shrinkage factor and the resulting MAE scores. The empirical curve demonstrated that the error decreased consistently as the shrinkage value increased, indicating the importance of the parameter in controlling the learning dynamics of M5Boost. Furthermore, it could be observed that the MAE stabilized around a shrinkage value of 0.1, indicating that larger values did not significantly alter the final prediction performance. Consequently, a shrinkage value of 0.1 was selected for the proposed M5Boost framework.

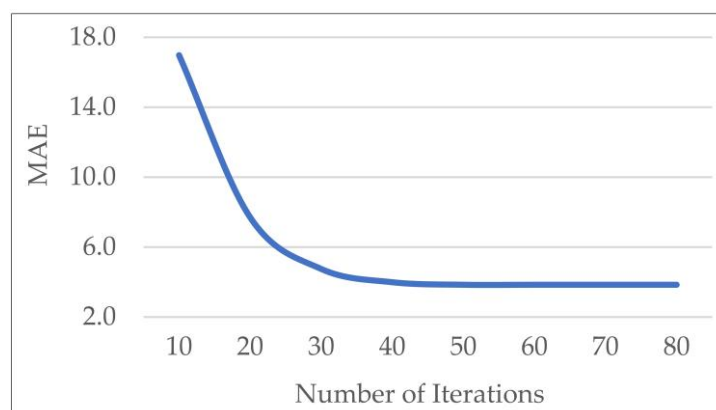


Figure 7. Effect of the number of iterations on the M5Boost performance.

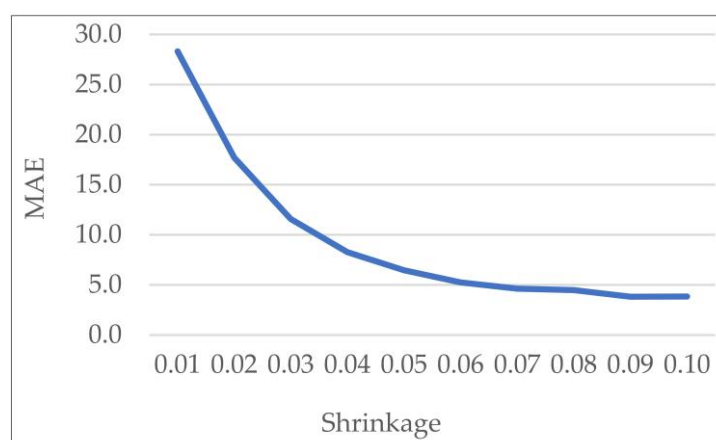


Figure 8. Effect of shrinkage on the M5Boost performance.

Table 9 presents the error values obtained across varying thresholds for the minimum number of instances per leaf node (`minNumInstances`), ranging from 5 to 45 in increments of 10. This parameter regulates the granularity of the recursive partitioning process; smaller values allow highly intricate tree topologies that risk overfitting, while excessively large values induce underfitting by restricting local linear model fitting. The experimental results demonstrated that the best performance was successfully achieved at a value of 35, suggesting the best balance between model complexity and predictive accuracy. Therefore, this value was selected in the final M5Boost configuration to address the EV driving range prediction problem.

Table 9. Effect of the minimum number of instances on the M5Boost performance.

MinNumInstances	5	15	25	35	45
MAE	4.2771	4.4842	4.5321	3.8411	3.8768
RMSE	8.3333	8.2663	8.1572	7.8485	8.0660
MAPE	37.6477	40.0014	40.4472	34.0127	32.4782

Table 10 shows the prediction performance of M5Boost under alternative train-test partitioning strategies, including 75:25, 80:20, 85:15, and 90:10. As the proportion of training samples increases, the model is able to learn more representative patterns from the data, resulting in improved prediction accuracy. The 90:10 split provided sufficient training samples for effectively learning the underlying nonlinear complex relationships while maintaining an independent test set for reliable performance evaluation. Furthermore, train-test partitions with high training ratios are widely adopted in machine learning ap-

plications involving small datasets, as they provide a sufficient number of training samples to effectively learn the volatile patterns while yielding a statistically robust test boundary.

Table 10. Prediction performance of M5Boost under alternative train-test partitioning strategies.

Data Split	90–10%	85–15%	80–20%	75–25%
MAE	3.8411	3.9304	4.3883	5.0828
RMSE	7.8485	8.9284	8.6971	8.8267
MAPE	34.0127	26.0322	29.4089	31.5562

A leave-one-vehicle-out evaluation was also conducted, in which one vehicle was entirely excluded from the training process and used exclusively for testing, representing a substantially more challenging evaluation scenario than the adopted hold-out strategy. The experimental observations showed that prediction performance decreased under this protocol, indicating that vehicle-specific driving behavior, battery characteristics, and operational conditions introduce inter-vehicle variability into the range estimation problem. Nevertheless, since the objective of this study is to construct a generalized prediction model using a fleet of real-world driving records, the adopted hold-out evaluation protocol provides an appropriate and representative experimental setting for the targeted application.

4.8. Tree Structure of the Proposed Machine Learning Model

To demonstrate the structural transparency and internal decision logic, Figure 9 presents a tree topology corresponding to a base learner within the M5Boost model. Through two main components (decision nodes and leaf nodes), the proposed method reveals its routing pathways explicitly. The decision nodes define the structural boundaries by partitioning the feature space according to threshold-based conditions, whereas the leaf nodes contain local regression models used to generate the final prediction values. During prediction, an input instance strictly traverses the tree based on the established split criteria until it reaches a terminal leaf node. The corresponding local regression model is then used to estimate the driving range. For instance, the recursive partitioning begins at the root node, which performs the initial data split based on the condition $Quantity \leq 4.795$. If a given driving sample satisfies this condition (=yes), the algorithm routes the instance to the left branch; otherwise, it follows the right branch. Subsequent decision boundaries further partition the feature space using variables such as year, vehicle age, motorway, and average speed. Instead of outputting static constant values, the leaf nodes represent localized, multivariate linear regression models, denoted by LM1, LM2, ..., LM22. These localized equations compute precise numerical estimations tailored specifically to the data partition that falls into that particular leaf. For example, the mathematical formulation of the local regression model associated with LM1 is given in Equation (20).

$$\begin{aligned}
 \text{trip_distance} = & 3.253 * \text{quantity} + 0.0429 * \text{year} + 0.0053 * \text{avg_speed} - 0.0072 * \text{vehicle_age} \\
 & - 0.0946 * \text{motor_way} + 0.188 * \text{country_roads} - 0.0118 * \text{driving_style} \\
 & - 113.5135
 \end{aligned} \quad (20)$$

The tree structure significantly contributes to the Explainable Artificial Intelligence (XAI) property of M5Boost, providing operational transparency. Unlike many black-box machine learning approaches, the decision boundaries generated by this approach remain human-readable and can be directly interpreted by users. Drivers can easily audit, trace, and validate how a specific EV driving range estimation is generated. Furthermore, the hierarchical structure of the tree also provides valuable information about feature importance, since features appearing near the root generally have a greater influence on the prediction process than those at deeper levels. According to Figure 9, the Quantity feature

is the most prominent variable because it is selected as the root splitting attribute. This observation is consistent with the Mutual Information analysis presented earlier.

In addition to global feature positioning, the localized regression models at the leaf level also provide deeper insights into feature importance through their mathematical weights. As shown in Equation (19), the largest regression coefficient belongs to the Quantity feature, indicating its dominant contribution to range estimation. These representative examples demonstrate that both the decision boundaries and the leaf-level regression equations remain highly interpretable and transparent compared to traditional black-box boosting methods. By providing insights into the prediction mechanism, the proposed M5Boost approach bridges the gap between high-precision accuracy and human-understandable model interpretability.

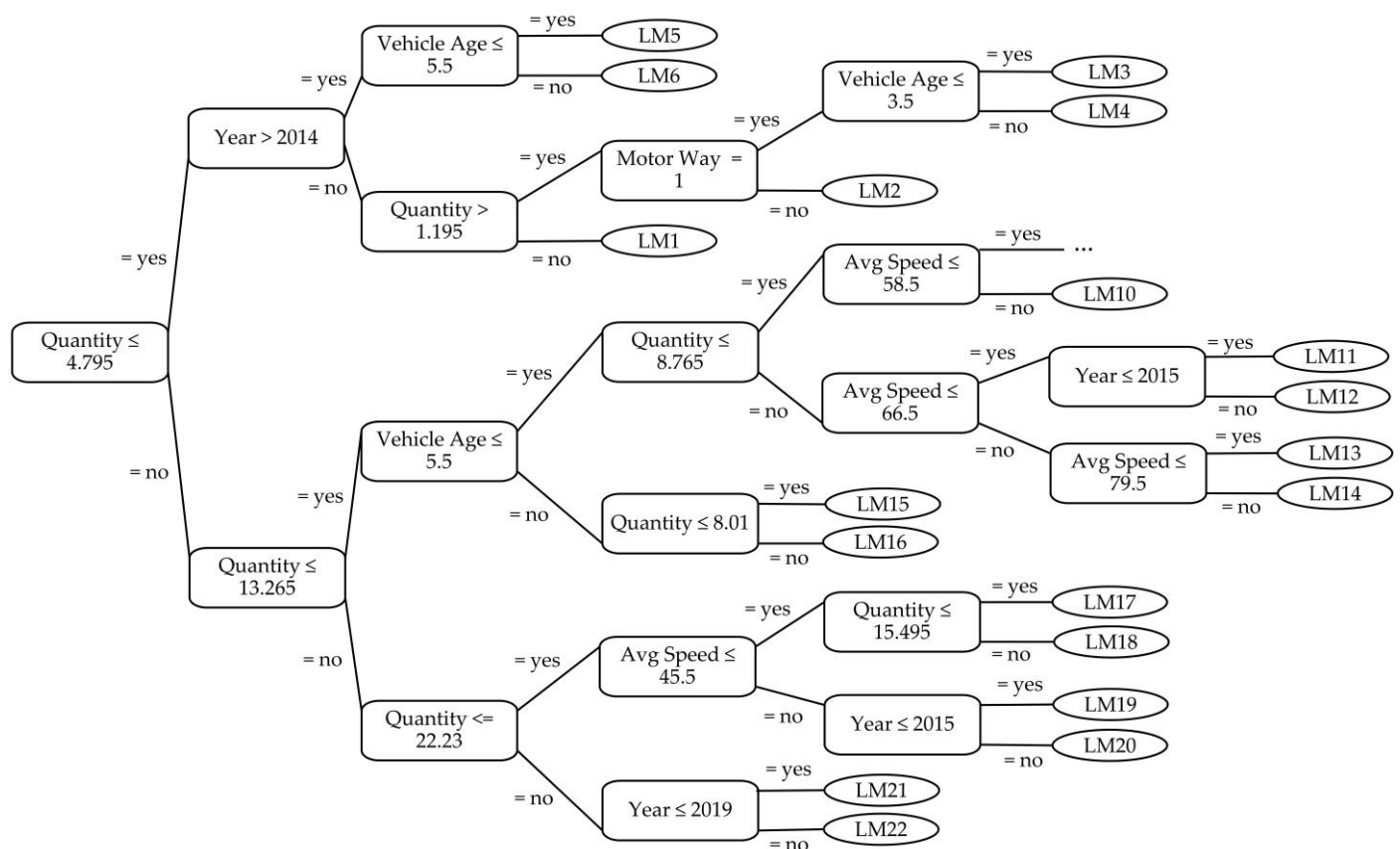


Figure 9. The structure of a model tree generated by M5Boost.

5. Comparison with Existing Methods

To objectively evaluate the practical significance of the proposed approach, a comprehensive benchmarking was performed against previously reported methods [47–51] on the same dataset [44]. Table 11 presents a comparative analysis between M5Boost and various state-of-the-art models spanning from conventional regression techniques to sophisticated deep learning frameworks. According to the reported results, the proposed M5Boost framework achieved the most superior result across the evaluation metric, yielding an MAE value of 3.9888. Compared to the most recent study presented in [47], M5Boost significantly outperformed the ETRARO_DRP model (MAE: 6.6380) and other ensemble variants like LightGBM, CatBoost, and XGBoost. Similarly, it demonstrated superior precision over the results in [48], reducing the RMSE value from 11.8738 to 8.0282. M5Boost also surpassed deep learning-based approaches such as DNN [47] reported in previous studies. In comparison with the deep learning-focused study [50], which produced a MAE of 5.1150 with Deep MLP, M5Boost provided a further error reduction

(MAE: 3.9888). These findings demonstrated that the integration of the iterative error-correction mechanism of additive regression with M5 model trees enabled the proposed framework to provide more accurate and stable EV range estimation performance than existing state-of-the-art approaches. The superior performance of M5Boost can be attributed to its capability to iteratively reduce residual estimation errors while simultaneously modeling region-specific regression behaviors through the tree structure.

To provide a broader evaluation perspective, M5Boost was compared with a model-based approach reported in the literature [51]. The model-based baseline yielded an MAE of 8.9137 and an RMSE of 14.4949, whereas M5Boost significantly reduced these errors to 3.9888 and 8.0282, respectively. Furthermore, comparisons with recent hybrid approaches [49] showed that M5Boost consistently achieved lower prediction errors. Although hybrid methods combine multiple paradigms, they often suffer from severe drawbacks in terms of model complexity, higher computational requirements, and reduced explainability, which limit their deployment in real-world vehicle electronic control units. In contrast, M5Boost offers a favorable balance between predictive performance, efficiency, and interpretability through its model-tree-based structure, making it highly suitable for practical EV range estimation applications.

Table 11. Performance comparison of M5Boost with benchmark methods on the same dataset.

Reference	Year		MAE	RMSE
[47]	2025	ETRARO_DRP	6.6380	14.8740
		DNN	6.7100	13.4620
		RF	7.0340	14.9900
		LightGBM	7.2140	15.3090
		XGBoost	7.3410	15.9700
		GB	7.6360	15.4030
		CatBoost	7.8100	15.0520
		Bagging	8.8010	20.3040
		DT	9.6880	22.4380
[48]	2024	Linear Regression	NA	17.7274
		Random Forest	NA	13.6498
		Deep MLP	NA	11.8738
[49]	2024	ELM	12.649	NA
		XGBoost	12.710	NA
		MLR	16.490	NA
		MLP	11.131	NA
		Deep MLP	11.738	NA
		RF	10.698	NA
		AdaBoost	15.155	NA
		RF and XGBoost	10.459	NA
		MLR and XGBoost	12.656	NA
[50]	2019	RF and MLP	9.5100	NA
		Linear Regression	11.689	NA
		MLP	7.2200	NA
		Deep MLP	5.1150	NA
		Random Forest	7.2680	NA
[51]	2013	AdaBoost	12.6800	NA
		Model-based approach	8.9137	14.4949
Proposed		Average	9.7981	15.8114
		M5Boost	3.9888	8.0282

6. Conclusions and Future Work

This study presents a machine learning approach, M5Boost, aimed at improving the accuracy of EV range estimation. Accurate prediction of EV driving range is essential for reducing uncertainty during driving and enabling more efficient trip planning, charging schedules, and battery utilization. However, EV range prediction remains a highly challenging problem due to the complex and nonlinear interactions among environmental conditions, driving behaviors, traffic situations, battery characteristics, and vehicle operational parameters. Current studies still suffer from issues related to limited interpretability, dependence on simulation-based scenarios, high computational complexity, and insufficient robustness under diverse real-world driving conditions.

M5Boost establishes multiple model trees with additive regression as a boosting strategy in order to iteratively reduce residual estimation errors while constructing localized regression structures. In addition, specialized tree pruning and smoothing operations were incorporated into the learning process to improve robustness and generalization capability. Unlike conventional boosting mechanisms that rely on constant-valued terminal nodes, the proposed framework utilizes multivariate linear regression models at its leaf nodes. This synergistic design improves robustness and mathematical interpretability.

To validate the efficacy of M5Boost, a benchmark dataset was systematically extended with newly collected real-world driving records for battery-related use cases, resulting in a comprehensive dataset spanning more than 11 years of EV operational data. Extensive experimental studies were conducted using eleven prominent ML methods, including KNN, KStar, LWL, LR, ANN, SVM, RF, ElasticNet, LightGBM, XGBoost, and CatBoost. Experimental results demonstrated that M5Boost achieved the best overall predictive performance among all evaluated algorithms. Statistical tests additionally confirmed that the observed performance improvements were statistically significant. Moreover, comparative analyses with previously published state-of-the-art studies conducted on the same dataset showed that M5Boost significantly outperformed existing approaches across both MAE and RMSE metrics. All these empirical results confirm that M5Boost effectively captures the highly nonlinear and complex dynamics of the EV range estimation problem without compromising computational efficiency. Therefore, the proposed approach has a strong potential for practical intelligent transportation and EV energy management systems operating under diverse environmental and driving conditions.

The generalizability of data-driven estimation models across diverse electric vehicle brands, models, and powertrain configurations represents a critical operational consideration. In this study, the empirical evaluation was conducted utilizing real-world driving data collected from Volkswagen e-Golf vehicles. Although the reported performance values are directly associated with this vehicle family, M5Boost is designed as a generic machine learning framework rather than a vehicle-dependent model. The input features, such as speed, vehicle age, ambient temperature, tire type, driving style, and road environment characteristics, represent common factors that influence the driving range of other battery electric vehicles as well. Consequently, the structural logic of the M5Boost approach is inherently brand-agnostic. Since it is fundamentally rooted in universal vehicle physics and principles, the framework can be systematically retrained and adapted to different EV models without requiring structural modifications, when representative training data are available.

As future work, a user-friendly graphical interface is planned to be developed in order to improve the accessibility and usability of the proposed M5Boost framework for real-world EV users and intelligent transportation web/mobile applications. This digital tool will aim to visualize dynamic range predictions under varying driving profiles, allowing users to interactively assess information and seamlessly manage their charging schedules during active trip planning.

Author Contributions: Conceptualization, I.A.K. and K.F.B.; methodology, I.A.K. and K.F.B.; software, D.B.; validation, I.A.K. and K.F.B.; formal analysis, K.U.B.; investigation, I.A.K. and K.F.B.; resources, K.U.B.; data curation, K.U.B.; writing—original draft preparation, I.A.K., K.F.B. and D.B.; writing—review and editing, K.U.B.; visualization, K.U.B. and D.B.; supervision, D.B.; project administration, D.B.; funding acquisition, I.A.K. and K.F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset utilized in this study is publicly available [44] on the GitHub repository (<https://github.com/armiro/crawlers/tree/master/SpritMonitor-Crawler>, accessed on 25 April 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AdaBoost	Adaptive boosting
ANN	Artificial neural networks
BiLSTM	Bidirectional long short-term memory
CatBoost	Categorical boosting
CNN	Convolutional neural network
DEKF	Dual extended Kalman filter
DNN	Deep neural networks
DT	Decision trees
EKF	Extended Kalman filter
ET	Extra Trees
ETRARO_DRP	Extra tree regressor with artificial rabbit optimization for driving range prediction
EV	Electric vehicles
FMM	Fuzzy Markov model
GAN	Generative adversarial networks
GB	Gradient boosting
GBDT	Gradient boosted decision trees
GRU	Gated recurrent unit
HMM	Hidden Markov model
KAN	Kolmogorov–Arnold Network
KF	Kalman filter
KNN	K-nearest neighbors
LFP	Lithium iron phosphate
LR	Linear regression
LSTM	Long short-term memory
MAE	Mean absolute error
ML	Machine learning
MLP	Multi-layer perceptron
NCM	Nickel cobalt manganese
NN	Neural network
RF	Random forest
RMSE	Root mean squared error
RNN	Recurrent neural networks
SoC	State of charge
SVM	Support vector machine
TCN	Temporal convolutional network
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting

References

1. Amin, A.; Amin, M.S.; Park, H.; Lee, D. Electric Vehicle Range Prediction Models: A Systematic Review of Machine Learning, Mathematical, and Simulation Approaches. *World Electr. Veh. J.* **2025**, *16*, 607. <https://doi.org/10.3390/wevj16110607>.
2. Caferler, B.; Ünal, A.; Keser, S.B.; Yazıcı, A. A Review for Remaining Driving Range Prediction of Electric Vehicles Using Machine Learning Algorithms. *J. Data Sci. Intell. Syst.* **2025**, *in press*. <https://doi.org/10.47852/bonviewJDSIS52025131>.
3. Mei, P.; Karimi, H.R.; Huang, C.; Chen, F.; Yang, S. Remaining driving range prediction for electric vehicles: Key challenges and outlook. *IET Control Theory Appl.* **2023**, *17*, 1875–1893. <https://doi.org/10.1049/cth2.12486>.
4. Dong, J.; Wu, X.; Liu, C.; Lin, Z.; Hu, L. The impact of reliable range estimation on battery electric vehicle feasibility. *Int. J. Sustain. Transp.* **2020**, *14*, 833–842. <https://doi.org/10.1080/15568318.2019.1639085>.
5. Gurusamy, A.; Ashok, B.; Mason, B. Prediction of Electric Vehicle Driving Range and Performance Characteristics: A Review on Analytical Modeling Strategies with its Influential Factors and Improvisation Techniques. *IEEE Access* **2023**, *11*, 131521–131548. <https://doi.org/10.1109/ACCESS.2023.3334620>.
6. Dollinger, M.; Fischerauer, G. Model-Based Range Prediction for Electric Cars and Trucks under Real-World Conditions. *Energies* **2021**, *14*, 5804. <https://doi.org/10.3390/en14185804>.
7. Cannavacciuolo, G.; Maino, C.; Misul, D.A.; Spessa, E. A Model for the Estimation of the Residual Driving Range of Battery Electric Vehicles Including Battery Ageing, Thermal Effects and Auxiliaries. *Appl. Sci.* **2021**, *11*, 9316. <https://doi.org/10.3390/app11199316>.
8. Çeven, S.; Albayrak, A.; Bayır, R. Real-time range estimation in electric vehicles using fuzzy logic classifier. *Comput. Electr. Eng.* **2020**, *83*, 106577. <https://doi.org/10.1016/j.compeleceng.2020.106577>.
9. Chellaswamy, C.; Geetha, T.S.; Kannan, G.; Vanathi, A. Fuzzy Classification Based Driving Distance Estimation for Electric Vehicles. *Trends Sci.* **2021**, *18*, 32. <https://doi.org/10.48048/tis.2021.32>.
10. Thorgeirsson, A.T.; Vaillant, M.; Scheubner, S.; Gauterin, F. Evaluating system architectures for driving range estimation and charge planning for electric vehicles. *Softw. Pract. Exp.* **2021**, *51*, 72–90. <https://doi.org/10.1002/spe.2914>.
11. Unni, K.; Thale, S.S. Regression algorithm based residual range prediction and validation on EV travel data. *Aust. J. Electr. Electron. Eng.* **2026**, *in press*. <https://doi.org/10.1080/1448837X.2025.2457256>.
12. Ye, J.; Lou, B.; Hu, L.; Liu, R.; Tang, J. Prediction of electric vehicles remaining driving range based on BO-CatBoost-SHAP model. *Transp. A Transp. Sci.* **2026**, *in press*. <https://doi.org/10.1080/23249935.2025.2519568>.
13. Unni, K.; Thale, S. Machine-learning-based range estimation for electric vehicles with an Indian road perspective. *Int. J. Ambient Energy* **2026**, *47*, 2625939. <https://doi.org/10.1080/01430750.2026.2625939>.
14. Khodijah, S.; Rizki, C.A.; Hasanuddin, M. Integrated Multi-Domain Modeling Framework for Energy Efficiency and Range Prediction in Modern Electric Vehicle Systems. *Int. J. Appl. Sci. Technol. Appl.* **2026**, *1*, 1–9. <https://doi.org/10.62712/ijapset.v1i1.1>.
15. Zhou, L.; Wang, Z.; Mao, Z.; Wang, Q.; Zhang, D.; Zhang, Z.; Chen, Z. Data-driven remaining driving range estimation and analysis framework for electric vehicles under real-world conditions. *Appl. Energy* **2026**, *402*, 126881. <https://doi.org/10.1016/j.apenergy.2025.126881>.
16. García Bustos, J.E.; Brito Schiele, B.; Masserano, B.; Salas-Espiñeira, R.; Troncoso-Kurtovic, D.; Acuña-Ureta, D.E.; Jaramillo-Montoya, F.; Orchard, M.E.; Silva, J.F.; Perez, A. Enabling online maximum driving range prognostics in electric vehicles via uncertain event likelihood functions. *Eng. Appl. Artif. Intell.* **2026**, *173*, 114449. <https://doi.org/10.1016/j.engappai.2026.114449>.
17. Han, Z.; Chen, Z.; Yang, Z.; Zhang, Z.; Zhang, B. Remaining mileage estimation for electric vehicles based on dual extended Kalman filter and eXtreme gradient boosting. *Green Energy Intell. Transp.* **2025**, *4*, 100259. <https://doi.org/10.1016/j.geits.2025.100259>.
18. Li, D.; Sun, Q. Prediction of remaining driving range for electric vehicles based on IVY feature selection and parameter optimization of KAN. *J. Clean. Prod.* **2025**, *527*, 146680. <https://doi.org/10.1016/j.jclepro.2025.146680>.
19. Polat, A.A.; Bozkurt Keser, S.; Sarıççek, İ.; Yazıcı, A. Analysis of Factors Affecting Electric Vehicle Range Estimation: A Case Study of the Eskisehir Osmangazi University Campus. *Sustainability* **2025**, *17*, 3488. <https://doi.org/10.3390/su17083488>.
20. Jiang, Y.; Guo, J.; Xie, D.; Hou, Z.; Deng, J. Remaining Driving Range Prediction of Electric Vehicles Based on Personalized Driving Behavior in Complex Traffic Scenarios. *IEEE Open J. Veh. Technol.* **2025**, *6*, 1333–1347. <https://doi.org/10.1109/OJVT.2025.3565498>.
21. Lou, B.; Tang, J.; Hu, L.; Ye, J. Multi-source data-driven short-term remaining driving range prediction for electric vehicles: A hybrid CNN-transformer framework. *Energy* **2025**, *334*, 137564. <https://doi.org/10.1016/j.energy.2025.137564>.
22. Khekare, U.; Vedaraj, I.S. Adaptive deep Q-networks for accurate electric vehicle range estimation. *Front. Big Data* **2025**, *8*, 1697478. <https://doi.org/10.3389/fdata.2025.1697478>.

23. Dora, P.; Harini, N. Enhancing Electric Vehicle Range Prediction: A Federated Learning Framework Addressing Non-IID Data Challenges. *IEEE Internet Comput.* **2025**, *29*, 44–54. <https://doi.org/10.1109/MIC.2025.3622738>.
24. Ma, X.; Li, J.; Cui, H.; Cheng, L.; Ji, Y.; Wang, J. Electric vehicle range prediction considering real-time driving factors and battery capacity index. *Transp. Res. Part D Transp. Environ.* **2025**, *144*, 104795. <https://doi.org/10.1016/j.trd.2025.104795>.
25. Hasib, S.A.; Gulzar, M.M.; Shakoor, A.; Habib, S.; Murtaza, A.F. Optimizing electric vehicle driving range prediction using deep learning: A deep neural network (DNN) approach. *Results Eng.* **2024**, *24*, 103630. <https://doi.org/10.1016/j.rineng.2024.103630>.
26. Ke, H.; Bi, J.; Wang, Y.; Zhang, Y. Driving range estimation for electric bus based on atomic orbital search and back propagation neural network. *IET Intell. Transp. Syst.* **2024**, *18*, 2884–2895. <https://doi.org/10.1049/itr2.12592>.
27. Qiu, Y.; Dobbelaere, C.; Song, S. Energy Cost Analysis and Operational Range Prediction Based on Medium- and Heavy-Duty Electric Vehicle Real-World Deployments across the United States. *World Electr. Veh. J.* **2023**, *14*, 330. <https://doi.org/10.3390/wevj14120330>.
28. Eissa, M.A.; Chen, P. Machine Learning-based Electric Vehicle Battery State of Charge Prediction and Driving Range Estimation for Rural Applications. *IFAC-PapersOnLine* **2023**, *56*, 355–360. <https://doi.org/10.1016/j.ifacol.2023.12.050>.
29. Albuquerque, D.A.S.G.; Ferreira, A.J.; Coutinho, D.P. An Approach to Estimate Electric Vehicle Remaining Driving Distance. *i-ETC ISEL Acad. J. Electron. Telecommun. Comput.* **2023**, *9*, 1–11.
30. Ayevide, F.K.; Kelouwani, S.; Amamou, A.; Kandidayeni, M.; Chaoui, H. Estimation of a battery electric vehicle output power and remaining driving range under subfreezing conditions. *J. Energy Storage* **2022**, *55*, 105554. <https://doi.org/10.1016/j.est.2022.105554>.
31. Ahmed, M.; Mao, Z.; Zheng, Y.; Chen, T.; Chen, Z. Electric Vehicle Range Estimation Using Regression Techniques. *World Electr. Veh. J.* **2022**, *13*, 105. <https://doi.org/10.3390/wevj13060105>.
32. Sun, T.; Xu, Y.; Feng, L.; Xu, B.; Chen, D.; Zhang, F.; Han, X.; Zhao, L.; Zheng, Y. A vehicle-cloud collaboration strategy for remaining driving range estimation based on online traffic route information and future operation condition prediction. *Energy* **2022**, *248*, 123608. <https://doi.org/10.1016/j.energy.2022.123608>.
33. Armenta-Déu, C.; Cortés, H. A Hybrid Method to Calculate the Real Driving Range of Electric Vehicles on Intercity Routes. *Vehicles* **2023**, *5*, 482–497. <https://doi.org/10.3390/vehicles5020027>.
34. Zhao, L.; Yao, W.; Wang, Y.; Hu, J. Machine Learning-Based Method for Remaining Range Prediction of Electric Vehicles. *IEEE Access* **2020**, *8*, 212423–212441. <https://doi.org/10.1109/ACCESS.2020.3039815>.
35. Micu, D.A.; Băţăuș, M.V.; Renţea, C.A.; Ancuţa, A.A.; Mancaş, R. The Influence of Road Gradient Resistance on the Driving Range of Electric Vehicles. *Vehicles* **2026**, *8*, 44. <https://doi.org/10.3390/vehicles8030044>.
36. Szumska, E.M.; Jurecki, R.S. Parameters Influencing on Electric Vehicle Range. *Energies* **2021**, *14*, 4821. <https://doi.org/10.3390/en14164821>.
37. Tripathy, Y.; McGordon, A.; Barai, A. Improving Accessible Capacity Tracking at Low Ambient Temperatures for Range Estimation of Battery Electric Vehicles. *Energies* **2020**, *13*, 2021. <https://doi.org/10.3390/en13082021>.
38. Wang, F.; Dai, T.; Zheng, X.; Pan, X.; Lei, L. Analysis of Factors Affecting the Accuracy Evaluation of Range Estimation of Pure Electric Vehicles Based on CLTC-P. *J. Phys. Conf. Ser.* **2020**, *1607*, 012012. <https://doi.org/10.1088/1742-6596/1607/1/012012>.
39. Harth, P.; Nagy, A. Estimation of the Range of a Light-Duty Commercial Vehicle. *Eng. Proc.* **2025**, *113*, 3. <https://doi.org/10.3390/engproc2025113003>.
40. Armenta-Déu, C.; Cattin, E. Real Driving Range in Electric Vehicles: Influence on Fuel Consumption and Carbon Emissions. *World Electr. Veh. J.* **2021**, *12*, 166. <https://doi.org/10.3390/wevj12040166>.
41. Prasad, S.L.; Gudipalli, A. An effective range estimation and state-of-charge to mitigate range anxiety in electric vehicles. *Heliyon* **2025**, *11*, e41494. <https://doi.org/10.1016/j.heliyon.2024.e41494>.
42. Wei, H.; He, C.; Li, J.; Zhao, L. Online estimation of driving range for battery electric vehicles based on SOC-segmented actual driving cycle. *J. Energy Storage* **2022**, *49*, 104091. <https://doi.org/10.1016/j.est.2022.104091>.
43. Sangeetha, E.P.; Subashini, N.; Santhosh, T.K.; Augusti Lindiya, S.; Uma, D. Validation of EKF based SoC estimation using vehicle dynamic modelling for range prediction. *Electr. Power Syst. Res.* **2024**, *226*, 109905. <https://doi.org/10.1016/j.epsr.2023.109905>.
44. Arandia, I.; Cejudo, I.; Irigoyen, E.; Urbieta, I.; Arregui, H.; Loyo, E. Analyzing the Influence of Driver, Route and Vehicle-Related Factors in Electric Vehicle Energy Consumption, Based on Real Life Data. In Proceedings of the 3rd International Conference on Robotics Systems and Vehicle Technology (RSVT 2022), Singapore, 22–24 July 2022; pp. 88–94. <https://doi.org/10.1145/3560453.3560467>.

45. Frank, E.; Hall, M.A.; Witten, I.H. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann: San Francisco, CA, USA, 2016.
46. Quinlan, R.J. Learning with Continuous Classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Singapore, 16–18 November 1992; pp. 343–348.
47. Keser, S.B. Artificial Rabbits Optimization for Refining Extra Trees Regression in Accurate Electric Vehicle Range Prediction. *IET Intell. Transp. Syst.* **2025**, *19*, e70085. <https://doi.org/10.1049/itr2.70085>.
48. Mishra, D.P.; Kumar, P.; Rai, P.; Kumar, A.; Salkuti, S.R. Exploratory data analysis for electric vehicle driving range prediction: Insights and evaluation. *Int. J. Appl. Power Eng.* **2024**, *13*, 474–482. <https://doi.org/10.11591/ijape.v13.i2.pp474-482>.
49. Hosseini, S.; Yassine, A.; Akilan, T. Ensemble-based Robust Model for Accurate Driving Range Estimation of EVs Leveraging Big Data. In Proceedings of the IEEE 8th Energy Conference, Doha, Qatar, 4–7 March 2024; pp. 1–6. <https://doi.org/10.1109/ENERGYCON58629.2024.10488792>.
50. Amirkhani, A.; Haghani, A.; Mosavi, M.R. Electric Vehicles Driving Range and Energy Consumption Investigation: A Comparative Study of Machine Learning Techniques. In Proceedings of the 5th Conference on Signal Processing and Intelligent Systems (ICSPIS), Shahrood, Iran, 18–19 December 2019; pp. 1–6. <https://doi.org/10.1109/ICSPIS48872.2019.9066042>.
51. Grewal, K.S.; Darnell, P.M. Model-based EV range prediction for Electric Hybrid Vehicles. In Proceedings of the IET Hybrid and Electric Vehicles Conference, London, UK, 6–7 November 2013; pp. 1–6. <https://doi.org/10.1049/cp.2013.1895>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.